Technical Report

UCAM-CL-TR-939
ISSN 1476-2986

Number 939

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

Effects of timing on users' perceived control when interacting with intelligent systems

Christine Guo Yu

August 2019

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted May 2018 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Gonville & Caius Gollege.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This research relates to the usability of mixed-initiative interaction systems, in which actions can be initiated either through a choice by the user or through intelligent decisions taken by the system. The key issue addressed here is how to preserve the user's perceived control ("sense of agency") when the control of the interaction is being transferred between the system and the user in a back-and-forth manner.

Previous research in social psychology and cognitive neuroscience suggests timing is a factor that can influence perceived control in such back-and-forth interactions. This dissertation explores the hypothesis that in mixed-initiative interaction, a predictable interaction rhythm can preserve the user's sense of control and enhance their experience during a task (e.g. higher confidence in task performance, stronger temporal alignment, lower perceived levels of stress and effort), whereas irregular interaction timing can have the opposite effect. Three controlled experiments compare alternative rhythmic strategies when users interact with simple visual stimuli, simple auditory stimuli, and a more realistic assisted text labelling task. The results of all three experiments support the hypothesis that a predictable interaction rhythm is beneficial in a range of interaction modalities and applications.

This research contributes to the field of human-computer interaction (HCI) in four ways. Firstly, it builds novel connections between existing theories in cognitive neuroscience, social psychology and HCI, highlighting how rhythmic temporal structures can be beneficial to the user's experience: particularly, their sense of control. Secondly, it establishes timing as a crucial design resource for mixed-initiative interaction, and provides empirical evidence of how the user's perceived control and other task experiences (such as reported levels of confidence, stress and effort) can be influenced by the manipulation of timing. Thirdly, it provides quantitative measures for the user's entrainment behaviours that are applicable to a wide range of interaction timescales. Lastly, it contextualises the design of timing in a realistic application scenario and offers insights to the design of general end-user automation and decision support tools.

Acknowledgements

First of all, I wish to express my heartfelt thanks to my PhD supervisor, Professor Alan Blackwell, who has given me patient guidance, inspirational advice and unhesitating support in both my research and personal development. Most importantly, his unwavering confidence in me has motivated and encouraged me to overcome many hardships and adversities, and to pursue my research and life goals. His admirable openness, insights and critical thinking have set an example which I intend to follow in my future career.

I would also like to thank my co-supervisor, Professor Ian Cross, whose extensive cross-disciplinary knowledge has broadened my intellectual horizons. Without his advice, the theoretical and experimental fabric of this research would have been much thinner.

I was able to concentrate my full energy on this research thanks to the generous financial support of the Cambridge Commonwealth, European and International Trust and the China Scholarship Council. I also feel honoured to have received a research grant from the Department of Computer Science and Technology, and studentships from the Cambridge Philosophical Society and Gonville & Caius College.

I am very lucky to have worked with excellent colleagues who created a cordial and supportive atmosphere throughout my time at Cambridge. My colleagues Zhen Bai, Alistair Stead, Advait Sarkar, Isak Herman, Mariana Mărășoiu, Luke Church, Sam Aaron, Alexander Simpson, Moe Hadhrawi and Xiaomeng Wang in the Rainbow Group, as well as David Greatrex, Li-Ching Wang and Jiaxi Liu in the Centre for Music and Science have kindly helped and encouraged me whenever I needed it.

Living outside my home county for the first time, I have been embraced by a bigger family. With Anna Greatrex, Daniela Peris, Tim Ellis, Ye Zhu, Maria Tsapali, Oliver Chick, James Bell, Michael Kosicki, Wiktor Olszowy, Beñat Mencia Uranga and my other friends at Caius and in the University, I never felt lonely.

My parents and the rest of my family have given me their blessings and support as I fight for my dreams. I hope that I can make them proud.

Finally, I am grateful to Alex Shattock, whose wit and humour, faith and companionship have become an irreplaceable source of strength that has supported me through the years at Cambridge and will continue for many adventures ahead.

Contents

CHAPTER 1

Introduction

1.1 Research background

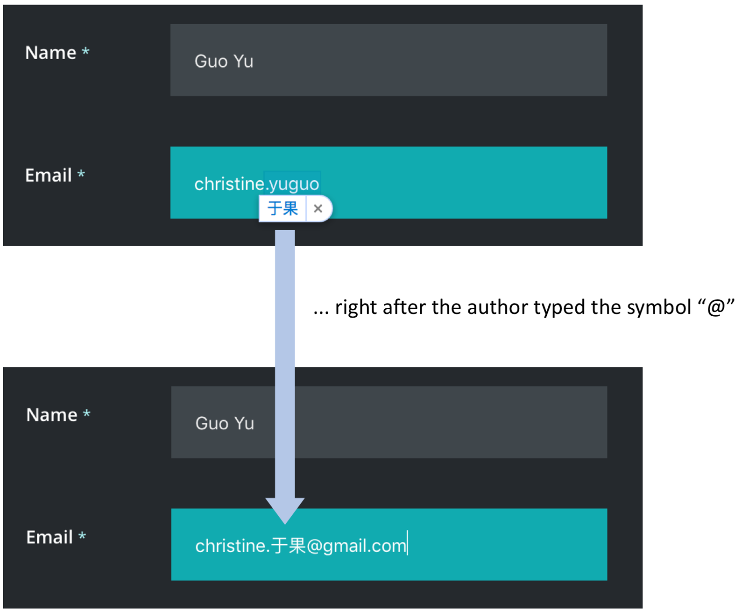
We all like to be in control. We often experience a sense of control when we rotate a steering wheel and see the car turn, or press a button and feel the room heat up, or in general, when we take a conscious action, and through that action, influence the immediate environment around us in the way that we want (Moore & Obhi, 2012). This subjective experience is called a *sense of agency*.

Our sense of agency is a malleable experience. Cognitive neuroscience studies find that a sense of agency arises when we ascribe authorship to our action and its effects or consequences in the external world. In other words, we should be able to say, "*I* caused that to happen" (Aarts, Custers, & Wegner, 2005; Coyle, Moore, Kristensson, Fletcher, & Blackwell, 2012). This inferential authorship ascription process is, however, prone to errors. When a person fails to observe the causal link between their action and its effects, they may not ascribe authorship to the effects, and therefore may not experience a sense of agency (Blakemore, Wolpert, & Frith, 2002; Farrer, Bouchereau, Jeannerod, & Franck, 2008). Alternatively, a person may falsely establish a causal link between their action and some external effects, and infer authorship of both based on a "belief-like mental state", thus experiencing an illusory sense of agency (Aarts et al., 2005). Nevertheless, being able to experience agency is crucial to our mental wellbeing, and if this ability is impaired, we may suffer from confusion or even schizophrenic symptoms (Mikesell, 2010).

Users' experience of agency is therefore a significant consideration in human-computer interaction design (Coyle et al., 2012). According to the *eight golden rules of interface design* proposed by Shneiderman, a good interface should support the user's " internal locus of control" (Shneiderman, 2010). One of the most influential design solutions that aims to improve the user's control experience is through *direct manipulation*, an interaction style popularised in the early 1980s. In contrast to abstract command lines, direct manipulation allows the user to manipulate graphical objects in a similar way to manipulating physical objects in the real world (Shneiderman, 1981, 1982), and the user can receive continuous and vivid feedback of their incremental actions (Kwon, Javed, Elmqvist, & Yi, 2011). A broader application of direct manipulation is what-you-see-is-what-you-get (WYSIWYG) editors, such as Microsoft Word/Excel and Adobe PageMill.

Modern interfaces increasingly incorporate automation and artificial intelligence (AI) components. While users welcome the convenience and assistance brought by AI, their sense of agency is faced with new challenges (Coyle et al., 2012). Users have already experienced the anxiety of losing control over simple inference-based functions such as auto-correction and Amazon recommendations (Madison, 2012; Blackwell, 2015) (and Figure 1.1 is merely one example of such frustration), let alone interacting with an intelligent system that appears to have its own mental model of the world (Kulesza, Burnett, Wong, & Stumpf, 2015; Blackwell, 2015).

Researchers tried to bring the merits of direct manipulation and of automated interface agents together when *mixed-initiative interaction* emerged as a new interaction paradigm (Horvitz, 1999a). In mixed-initiative interaction, the user and the automated system will take turns as in a dialogue, and their contributions are interwoven. Programming-by-example (PbE) applications generally exhibit mixed-initiative characteristics (Lieberman, 2000), where the user demonstrates desirable behaviours to the system (e.g. putting old E-mails from Alex into the folder named "Alex"), then the system learns to emulate and automate similar behaviours (e.g. putting new E-mails from Alex into his folder, and putting Eevee's E-mails into the "Eevee" folder), then the user manually confirms or corrects the system's behaviour, and so on and so forth. Inevitably, the controlling role in an interaction of this kind will be passed between the user and the system in a relatively frequent back-and-forth manner. However, very



Name * Guo Yu

Email * christine.yuguo

... right after the author typed the symbol "@"

Name * Guo Yu

Email * christine.于果@gmail.com

Figure 1.1: An example of the irritation caused by auto-correction: The author's computer apparently remembers her Chinese name, so it kindly and keenly replaces any "yuguo" with two Chinese characters *immediately*, even when she *chose* to type in English. The author has not worked out how to cancel this function and has decided to use copy-and-paste to input email addresses.

few studies have specifically studied or empirically measured the user's sense of agency during the transfer of initiative. This dissertation is, therefore, concerned with design factors that influence users' agency experience in mixed-initiative systems.

1.2 Research motivation and questions

The notion that the *timing* of mixed-initiative interaction is a crucial design consideration was raised by Horvitz (1999b) two decades ago:

> Mixed-initiative systems must consider a set of key decisions in their efforts to support joint activity and grounding. These include *when* (emphasis in original) to engage users with a service, *how* (emphasis in original) to best contribute to solving a problem, *when* (emphasis in original) to pass control of problem solving back to users for refinement or guidance, and

when to query a user for additional information in pursuit of minimizing uncertainty about a task.

However, most studies that considered the timing of mixed-initiative event invocation and presentation did not specifically investigate its effects on the user's sense of agency or give practical design guidelines (e.g., Horvitz & Barry, 1995; Wolber & Myers, 2001).

Given that the user tends to treat a computer system as a social actor (Nass, Steuer, & Tauber, 1994), and mixed-initiative interaction itself is like a dialogue (Bauer, Dengler, & Paul, 2001; Sarkar, 2017), this dissertation is particularly interested in the ways that the *timing* of such interactions might emulate interaction between two humans.

Inappropriate timing during human interactions can be unpleasant, and we use expressions like "you jumped down my throat" or "don't hijack the conversation" when someone takes over the initiative *too* fast or abruptly. For a long time, the design of timing in HCI was influenced by real time systems engineering and classical operations research: because "time is commodity" (Becker, 1965), users/customers should be happier with faster services, hence user interfaces should strive to respond as fast as possible (Weinberg, 2000; Rose & Straub, 2001). More recent studies, however, found that sometimes users were actually less satisfied when the system responded *too* fast. One example is that participants preferred waiting for an online flight-booking or dating website to search for up to 15 seconds (with sham searching progress presented on the screen) to receiving answers instantaneously, despite the search results being identical (Buell & Norton, 2011). It may be because a longer waiting time creates a "labour illusion" (i.e. as if more effort is being put into the task), thereby inducing feelings of reciprocity and valuation. The transparency of progress can also reduce perceived uncertainty (Buell & Norton, 2011) or support a stronger sense of control. Though this example only shows a classical command-respond interaction on a non-AI interface, it leads to the first research question of this dissertation:

Research Question 1: What timing characteristics are appropriate for mixed-initiative interaction?

Cognitive neuroscience research has found that our subjective experience of time is closely associated with the internal experience of agency (Ebert & Wegner, 2010; Moore & Obhi, 2012), and it can easily be warped: a person will perceive their voluntary action as being taken later than it actually was, and perceive the consequence of that action as happening sooner than the actual time. In other words, the action and the consequence were perceived as "being attracted" together temporally (Aarts et al., 2005; Engbert, Wohlschläger, Thomas, & Haggard, 2007). By measuring the distortion of a person's time perception, their sense of agency can be measured implicitly. This phenomenon is called "intentional binding", and will be further introduced in **Section 2.1.3** of this dissertation.

Our timing perception is also very susceptible to external influence. For example, manipulating a webpage's background colour can alter the user's perceived quickness of a download task (Gorn, Chattopadhyay, Sengupta, & Tripathi, 2004), adding animated images in a filler interface can reduce perceived waiting time (Lee, Chen, & Ilie, 2012), and adjusting the speed linearity of a progress bar can make it appear faster (Harrison, Amento, Kuznetsov, & Bell, 2007). Based on those findings, the second research question of this dissertation is:

Research Question 2: Can the timing of events become a design resource, which can be manipulated in a way that affects the user's agency experience?

Follow-up Question: If yes, then how can timing be manipulated to achieve this effect?

Mixed-initiative interaction that involves interactive machine learning components is recognised as a dialogue-like joint problem solving activity (Horvitz, 1999b; Sarkar, 2017), but a recent critique has pointed out that the behaviour of intelligent systems often falls short of the "basic courtesies of personal service" (Blackwell, 2015). Among the abundant social psychology and communication theories we can draw on, the theory of *rhythmic entrainment* is particularly relevant to the research questions raised above, and can offer insights into how we can achieve appropriate temporal co-ordination in the context of mixed-initiative interaction. The cognitive neuroscience basis, applications and implications of rhythmic entrainment will be reviewed in detail in **Sections 2.2.3**, **2.3.3** and **2.3.4** of this dissertation respectively.

Social interactions are essentially oscillatory processes that convey meaning (Clayton, Sager, & Will, 2005), and two or more such processes may entrain with each other as they interact with and adapt to each other's rhythm, and finally reach a relatively stable temporal synchrony (Clayton, 2012). Such temporal co-ordination can improve the predictability of an interaction, allowing interactants to exert *anticipatory control* (Keller, Knoblich, & Repp, 2007; Pecenka & Keller, 2011; Nowicki, Prinz, Grosjean, Repp, & Keller, 2013). Consequently, entrainment can be beneficial to social interactions: it can enhance interactants' pro-sociality and empathy (Spiro & Himberg, 2012; Spiro, Schofield, & Himberg, 2013), build rapport and mutual affiliation (Miles, Nind, & Macrae, 2009; Cross, 2013), and facilitate joint problem solving processes (Hawkins, Cross, & Ogden, 2013). The merits of entrainments in social interactions are motivating enough to ask the third research question of this dissertation:

Research Question 3: Can the rhythmic entrainment of a mixed-initiative interaction positively affect the user's experience, such as their sense of agency, perceived stress level, confidence and task performance?

Follow-up Question: If yes, then what are the design guidelines?

1.3 Dissertation overview

The research questions raised above are explored in each of the following chapters. In order to approach the questions, **Chapter 2** reviews the literature on three themes that this research draws on. The theme of **Section 2.1** is *agency.* I first introduce the definition of agency and its production mechanisms, then summarise five general approaches the HCI community has adopted to study the concept of agency, and analyse three aspects of new challenges to the user's agency experience brought by modern mixed-initiative interaction paradigms that incorporate artificial intelligence components. I then review existing philosophical models that explain how people attribute mental properties and agency to a computer system, and how people infer causality through observation. Then I review both the explicit measures (e.g. self reporting on a numeric scale) and the implicit measures (e.g. estimating event time on a Libet clock) for agency, analysing the pros and cons of each measure, as well as the differences and correlation between them. I also give a comprehensive review of four sets of factors that can affect the user's experience of agency.

The second theme, the user's *expectation* in HCI, is reviewed in **Section 2.2**. After defining "expectation", I introduce relevant cognitive-behavioural models in social psychology that could inform how the user's expectation can influence their perceptions of the computer system in mixed-initiative interaction. I also review the factors that can influence users' expectation and the underlying neural mechanisms that allow our brain to form temporal expectations.

In **Section 2.3**, I reviewed the third theme of literature, the *rhythmic entrainment* in interaction. I first define the term "rhythm" and introduce two important roles played by rhythm. Then I introduce the rhythmic entrainment theory well-established in musicology and social psychology literature, and review the effects of entrainment on interpersonal interaction. I also summarise existing studies on rhythm and entrainment in the HCI literature, which have mainly treated rhythm as a passive attribute and studied entrainment from a verbal/gestural perspective. I then highlight the potential research value of temporal entrainment in mixed-initiative interaction.

In **Chapter 3**, I set out the framework established for my PhD research. To answer my research questions, I propose four sets of research hypotheses accordingly, which were tested empirically in three experiments reported in later chapters. My considerations in adopting an empirical approach in this research are also explained in this chapter. I then lay the theoretical basis upon which the hypotheses are formulated, drawing on the three themes of the literature on agency, temporal expectation, and rhythmic entrainment reviewed in **Chapter 2**. Accordingly, for each hypothesis, I select dependent variables and their measurements that have been adopted by previous empirical works.

In **Chapters 4**, **5** and **6**, I report three experiments that I completed during the course of my PhD study.

Experiment 1 served as a first step in exploring the research questions empirically. As reported in **Chapter 4**, the main purpose of this experiment was twofold: 1) to find valid ways of manipulating the timing of initiative taking in a highly controlled manner, 2) to see whether or not the timing manipulation has caused significant effects on the user's sense of control, perceived stress level, entrainment behaviours and task performance as predicted in the hypotheses in **Chapter 3**. As will be reported in **Section 4.1**, Experiment 1 was a within-subject design that had the timing pattern of initiative taking as the only independent variable. There were four treatments, each of which corresponded to a kind of initiative-taking setting: 1) the system took the initiative at irregular intervals, 2) the system took the initiative at rhythmic intervals,

3) the user took the initiative first, then the system took the initiative aligning with the user's pace, 4) the user took the initiative in their own pace. The tasks were adapted from simple visual stimulus-response paradigms conventionally used in ergonomics and cognitive psychology studies, and were carried out following strictly designed protocols. By analysing the results in **Section 4.2**, I confirmed that the manipulation of timing in this experiment had been valid and had caused significant effects on participants' sense of control, reported level of stress and effort, tendency to entrainment, and task performance, as predicted in the four sets of hypotheses. In addition, I propose four design implications in **Section 4.3.1**. For example, the user may be happy to devote more physical effort in exchange for a higher sense of control and less mental stress, and the user tends to maintain their own rhythm against external temporal structures and may have adopted it as a way of preserving the sense of control. I discuss the limitations of the design and findings of Experiment 1 in **Section 4.3.2**. Based on the findings, I provide answers to the research questions in **Section 4.4**: the timing of mixed-initiative interaction can be manipulated as a design resource and can influence the user's experience and performance. It will usually be appropriate to allow the user to take the initiative in their own pace or let the system take the initiative rhythmically, because the user will have a relatively higher sense of control, better task performance and more confidence. Conversely, letting the system take the initiative irregularly can result in an impaired sense of control and task performance as well as a higher amount of perceived effort.

Having tested and supported my hypotheses in the context of interacting with visual stimuli, I report Experiment 2 in **Chapter 5**, in which participants needed to interact with auditory stimuli while observing a Libet clock. This experiment aimed to obtain further evidence to support the hypotheses while investigating how the timing of system-initiated events can influence the user's perception of time. Experiment 2 was also a within-subject design, and shared the same independent variable and its manipulation as Experiment 1. The details of task design and measurements are reported in **Section 5.1**. According to the results presented in **Section 5.2**, I found that the effects of timing on the user's sense of control, confidence in task performance and their reported level of stress and effort remained congruent whether the interaction happened in the visual or the auditory modality, thereby consolidating my hypotheses. I also noted that participants were able to recognise it when the system was emulating their pace, and appreciated it as being helpful and adaptive. In addition, I looked into how grouping effect may have contributed to participants' temporal expectation, and suggested that grouping irregular individual events in a regular pattern could be a way to mitigate the loss of agency caused by the temporal unpredictability of an interaction. Drawing on the findings above, I propose another three design implications and one observation-based prediction in **Section 5.3.1** and discuss the limitations that need to be considered when generalising the findings of Experiment 2 in **Section 5.3.2**.

Experiment 3 is reported in **Chapter 6**. This experiment investigated whether or not the hypotheses in **Chapter 3** could still hold in a relatively more realistic HCI context, with the aim of providing the HCI and machine learning community with a concrete showcase and practical design insights. I start with introducing the crucial role of labelling when training artificial intelligence algorithms in **Section 6.1.1**, and define the software systems that aim to improve the efficiency of labelling tasks and/or the quality of the labels given by human users as "assisted labelling" tools, and those with AI components as "AI-assisted labelling" tools. In **Section 6.1.2**, I identify that while AI-assisted labelling tasks often have mixed-initiative characteristics, the timing of labelling is an important yet underinvestigated design resource. In **Section 6.1.3**, I propose four hypotheses derived from those tested in Experiments 1 and 2 to fit the context of an AI-assisted labelling system. The design of Experiment 3 is introduced in **Section 6.2**. The experiment adopted the Wizard-of-Oz paradigm, in which participants were interacting with a simulation of an AI-assisted labelling interface in an imaginary task scenario. The independent variable and its manipulation were the same as Experiments 1 and 2, and the calculation of dependent variables was accommodated to the design of tasks. My main finding from the results in **Section 6.3** was that the effect of timing on the user's sense of control and their perceived stress and effort during the tasks observed in Experiments 1 and 2 also appeared in the context of AI-assisted labelling, thereby fulfilling the main purpose of contextualising theoretical findings in a more realistic application. I also found that when the system took the initiative and pushed messages to label at random times, participants would speed up to cope with the irregularity and feel more stressed and rushed, whereas when participants took the initiative, they would feel less stressed or rushed, despite the fact that they did not make the label decisions any more slowly. In addition, I noted that when participants had full control of the timing of labelling, they may wrongly reject more correct recommendations made by the system. I discuss the above observation and more design implications in **Section 6.4.1**. Finally, I analyse the limitations of the design and the findings of Experiment 3 and suggested potential directions for future studies in **Section 6.4.2**.

I conclude this dissertation with **Chapter 7**. I first summarise the findings of all three experiments as the answers to each of the research questions I proposed earlier in this chapter. I then expound the four contributions this dissertation has made to the field of human-computer interaction. I also discuss the limitations of the research methods and the results, and suggest potential directions for future research.

1.4 Research contributions

The major contributions of this dissertation are as follows:

1. It provides a cross-disciplinary review of the literature in the fields of human-computer interaction, cognitive neuroscience and social psychology, and establishes connections between the existing theories in the three fields to inform the design of mixed-initiative interaction that can preserve the user's perceived control (in **Chapters 2** and **3**);

2. It demonstrates the importance of timing during mixed-initiative interaction, proposes that the timing of an interaction, on both the visual and auditory modalities, can be manipulated as a design resource, and empirically tests the effect of timing on the user's perceived control (Experiments 1 and 2, in **Chapters 4** and **5**);

3. It provides quantitative measures for the user's entrainment behaviours during the handover of initiative on a relatively broad timescale, ranging from 250 milliseconds (Experiment 1 in **Chapter 4**) to 20 seconds (Experiment 3 in **Chapter 6**);

4. It showcases how rhythmic entrainment principles can be applied to the design of mixed-initiative systems such as AI-assisted labelling tools (Experiment 3 in **Chapter 6**), offering insights that can inform the design of the temporal aspects of mixed-initiative systems that incorporate inference-based components (in **Chapter 7**).

This dissertation has drawn on the content of the following publications of the author during the course of this PhD study:

1. **Yu, G.**, Blackwell, A. F., & Cross I. (2016). Understanding timing in mixed-initiative interaction. *In the Doctoral Consortium of the 27th the Psychology of Programming Interest Group (PPIG'16).* URL: http://www.ppig.org/sites/ppig.org/files/2016-PPIG-27th-Yu.pdf

2. **Yu, G.**, & Blackwell, A. F. (2017). "Are you following?" Agency and timing in mixed-initiative interaction. *Poster presentation on the 4th Oxbridge Women in Computer Science Conference (OWCSC'17).* URL: http://www.cl.cam.ac.uk/~gy238/Poster-OWCSC-2017-GuoYu.pdf

3. **Yu, G.**, & Blackwell, A. F. (2017). Effects of timing on users' agency during mixed-initiative interaction. *In Proceedings of the 31st British Human Computer Interaction Conference (BHCI'17)* (pp. 1–12). Electronic Workshop in Computing (eWiC). URL: https://ewic.bcs.org/upload/pdf/ewic_hci17_fp_paper3.pdf. DOI: 10.14236/ewic/HCI2017.35. DOI: 10.14236/ewic/HCI2017.35

CHAPTER 2

Literature review

2.1 Agency in mixed-initiative interaction

2.1.1 Definition of agency

Agency has many different definitions. In earlier philosophical and psychological studies (Bratman, 1999; McCann, 1998), the concept of agency was defined as a person's conscious state, in which they observe an outcome or an impact in the external world caused by their own action (Nowak & Biocca, 2003), and through that action and observation, a person can experience the *sense of agency* (SoA).

Cognitive neuroscientists take one of two opposing stances in explaining how a sense of agency arises (Moore & Haggard, 2008). The first stance argues that a person will compare the predictions made by their motor control system with the actual sensory/proprioceptive consequences. When there is a match, they will experience a sense of agency. The *comparator model* is one of the well established models taking this stance (Blakemore et al., 2002; Moore & Haggard, 2008), which has essentially two parts. To achieve optimal motor control, a person needs to be able to predict both the upcoming states of the motor system captured by the *forward dynamic model*, and the potential sensory consequences of movements based on the *forward sensory model* (Wolpert & Ghahramani, 2000). Such a predictive account of action is considered as *intrinsic* to the agent, and it contributes to the awareness of the action ownership *before* the action. The alternative stance maintains that the sense of agency arises not from predictive motor control processes, but from retrospective and "postdictive" inferential processes instead. In other words, a person will use sensory information and evidence to "make sense" of their actions and their subsequent outcomes (Moore & Haggard, 2008; Wegner & Wheatley, 1999; Hon, Poh, & Soon, 2013). When they build a causal link between them, they will experience a sense of agency. One major theory taking this stance is the *apparent mental causation model.* This model maintains that a person can infer a causal link if the action occurred prior to the outcome, if the outcome was consistent with their expectation and if the action was the only plausible cause of the outcome (Wegner & Wheatley, 1999). That means that the awareness of the action ownership is *not* a person's intrinsic knowledge, but is rather a restored link between their conscious action and the consequences built *after* the action. Nevertheless, recent research on the *cue integration model* suggested that the two stances are not necessarily mutually exclusive, because there may be a processing mechanism that brings together *both* internal cues (e.g. motor predictions, direct and indirect sensory feedbacks, action-relevant thoughts, etc.) and external cues (e.g. priming, environmental factors, social cues, action consequences, etc.). In other words, both internal motor signals and external information contribute to the formation and experience of agency, and when one source is not available, the other plays a greater role (Wegner & Sparrow, 2004; Moore, Wegner, & Haggard, 2009).

In the field of human-computer interaction, research explores the concept of agency from five different angles (Coyle et al., 2012):

1. The first is *media agency*, which centres around the *media equation theory* (Nass et al., 1994; Reeves & Nass, 1996). It suggests that people tend to treat information media (such as a computer system) as a social actor and respond to it in a similar way to how they respond to other human beings.

2. *Intelligent agents/interface* is the second angle (Franklin & Graesser, 1996; Faratin, Sierra, & Jennings, 1998; Klingspor, Demiris, & Kaiser, 1997), as computer systems are becoming more capable of observing the user's behaviours and providing the user with better assistance, or even serving as a delegate that can make decisions and execute actions with a certain level of autonomy on the user's behalf. Such ability is considered as "intelligence". It indicates that the user find those systems are exhibiting autonomy, resulting in a blurred boundary

between the agency of the user and that of the systems.

3. The third angle is *design agency* (Fogg, 1998, 2002). Design agency is where the user perceives a system or interface as a product that embodies a message or intention from its designer, hence attributes the agency to the designer.

4. The fourth angle angle is *agency in the laboratory* (Collins & Kusch, 1999; Suchman, 2007), which focuses on a) what role machines play in humans' knowledge production and interpretation processes (either as a neutral tool or as an entity, both are subjected to human observation), and b) how machines are attributed with characteristics through their participation.

5. The fifth and final angle, to which this dissertation is devoted, is the *sense of agency* of the user, highlighting a person's subjective experience of agency, or the sense of control, when interacting with a computer system (Coyle et al., 2012; Limerick, Moore, & Coyle, 2015).

As shown in the definition of agency given above, there are three key elements in the production of the experience of agency: the action, its consequence, and the sense of authorship of both. In mixed-initiative interaction, each element is subject to new challenges. First, intelligent interfaces increasingly complete the user's actions, and even automatically make decisions on our behalf. Therefore the sense of agency gained from "taking an action" can be reduced. Secondly, modern AI largely relies on machine learning (ML) algorithms that behave in a probabilistic manner, and their internal inferred models may "carry new consequences" (Blackwell, 2015). For instance, an ML algorithm may inferentially select features or make decisions that are not recognisable or comprehensible to human perception (Lowe, 1999; Nguyen, Yosinski, & Clune, 2015). When a consequence appears to be inconsistent with a person's prediction or prior knowledge, they may not be able to restore a causal link between their action and the consequence, hence their sense of agency may not arise (Wegner & Wheatley, 1999). Thirdly, mixed-initiative interaction is essentially a joint problem solving activity (Horvitz, 1999b), and the boundary between the user's and the system's contribution is blurred - even more, the behaviour of an interactive ML algorithm is the product of the interaction between the user and the algorithm itself (Blackwell, 2015). This will result in ambiguity during the ascription of authorship, and the user cannot be sure whether there was an alternative cause other than their own action that had led to that consequence (Berthaut, Coyle, Moore, & Limerick, 2015).

2.1.2 Agency perception and attribution

In the field of virtual reality and computer-mediated communication, the terms *avatar* and *agent* are defined based on the concept of agency (Bailenson, Blascovich, Beall, & Loomis, 2003; Mehdi, Nico, Dugdale, & Pavard, 2004; Guadagno, Blascovich, Bailenson, & Mccall, 2007; Skalski & Tamborini, 2007). An entity that behaves fully under the control of a human operator in real time is called an *avatar*, while an entity that operates on its own is called an *agent*. The key element that distinguishes the two terms is whether the computer entity is able to take an action and exert an impact in the external world on its own.

In mixed-initiative interaction, both the user and their computer counterpart are entities that can take actions and cause effects, and therefore the computer is also an agent, just like the user. On the user's side, they will experience a sense of agency when they consider themselves to have taken an action intentionally and can build a causal link between their action and the observed consequences. Meanwhile, they can observe the actions and consequences produced by the computer agent and perceive it as having agency. An underlying mechanism of this perceived agency is the attribution of *intentionality*.

Neuropsychological and behavioural studies have shown that humans are "hard-wired to respond to cues that suggest an entity has intentionality" (Nowak & Biocca, 2003), and with that perceived intentionality, humans may perceive the entity as "living" rather than "nonliving" (Warrington & Shallice, 1984; Gainotti, Silveri, Daniel, & Giustolisi, 1995). Reeves and Nass proposed that when interacting an entity that appears or behaves like alive, a person will exhibit *automatic social responsiveness* (Reeves & Nass, 1996) and pay more attention to that entity (Reeves & Nass, 1996; André, Klesen, Gebhard, Allen, & Rist, 2000; Martínez-Miranda, Bresó, & García-Gómez, 2012). In other words, the illusion of humanity/aliveness of an entity can trigger a person to treat it in a way that they would treat another human.

According to Dennett's *Three Stances* system, when a person decides to treat an entity as having mental properties (e.g. intelligence, intention) and tries to understand, explain and/or predict its behaviours, they can view the entity from three levels of abstraction, or *intentional stances*: the *physical stance*, the *design stance* and the *intentional stance* (Dennett, 1989). There are other variants of the abstraction levels proposed by other cognitive scientists, such as Pylyshyn's "Levels of Organization" (Pylyshyn, 1988), Newell's "Levels of Description" (Newell, 1982), and Marr's "Levels of Analysis" (Marr & Vision, 1982), but they essentially share the same structure, as presented in Table 2.1.

Dennett (1989)'s Intentional Stances	*Pylyshyn (1988)'s Levels of Organization*	*Newell (1982)'s Levels of Description*	*Marr and Vision (1982)'s Levels of Analysis*
Physical stance	Physical/ Biological level	Physical/ Device level	Hardware implementation/ Mechanism level
Design stance	Symbol level	Program/Symbol level	Representation and algorithm level
Intentional stance	Semantic/ Knowledge level	Knowledge level	Computational theory level

Table 2.1: Four different three-level structures that describe humans' knowledge abstraction

The first stance is the most basic and concrete one. It is concerned with objective principles in the physical world: a person can predict the future states of an object based on a specific set of physical or chemical laws, initial conditions and structural configurations. The second one is about predicting how a complex system - like a muscle group or a running vehicle - is supposed to operate assuming that it is not malfunctioning. Such predictions are derived from our knowledge about for what purpose the system is designed for. The last stance is the intentional stance. When a person knows neither about the structure nor about the design, but has only the knowledge of an agent's mental states, they will predict the agent's behaviours by deducing its intentions based on the assumptions that any agent will always take actions according to its beliefs and desires for the purpose of getting exactly what it wants. According to Dennett, our understanding and predictions of an entity's behaviour would be more accurate if we viewed it from a more concrete abstraction level. If we viewed it from a higher abstraction level that allows us to "zoom out", we could gain a greater computational power by filtering out impertinent details (Dennett, 1989). This view is also supported by Pylyshyn's argument that human cognitive processes are essentially *a species of computing*, and a higher knowledge level can explain a broader domain of behaviours (Pylyshyn, 1988).

In short, a computer system can be an agent because it is composed of “a set of actions, a set of goals and a body” (Newell, 1982). It also processes the knowledge to determine which actions to take following the behaviour law *principle of rationality*: the system chooses the actions that can achieve its goals (Newell, 1982). The user will go through a cognitive process described by the *Three Stances* to attribute intentionality, and therefore agency, to the computer system during human-computer interaction.

One relevant theory for this attribution process claims that experiencing a sense of agency of one’s own and attributing agency to oneself or another agent happen on different cognition levels. Specifically, the theory holds that the sense of agency is generated on first-order cognition based on *bottom-up* accounts of pre-reflective neuronal mechanisms, whereas the attribution of agency requires higher-order cognition that draws on reflective experience (Gallagher, 2007).

The human cognitive system can not only attribute agency based on prediction, but also *infer* causality based on observation. According to the *apparent mental causation model* (Wegner & Wheatley, 1999), a person would experience a conscious will when they can “draw the inference that their thought has caused their action” regardless of the correctness of the inference (Wegner, 2003), and they would ascribe the ownership of the action to themselves when three criteria are met:

1. Priority, which refers to the temporal sequence that a conscious thought occurs before an action within a time window that is close enough to bind up the two events.

2. Consistency, which is the congruence when the action taken agrees with the prior thought or intention.

3. Exclusiveness, which means the thought or intention is the most possible cause of the action, while alternative causes can be ruled out.

The perception and attribution of agency can be distorted. A person can experience an “agentic shift” when following external commands (Milgram & Gudehus, 1978). In some extreme cases such as in facilitated communication, the perception of another agent can completely cancel out a person's experience of conscious will and the perceived ownership of their own action (Jacobson, Mulick, & Schwartz, 1995). As mentioned in **Section 2.1.1**, when a computer system exhibits more autonomy and intelligence by completing a human user's action or even acting on their behalf, it becomes more difficult to draw a clear boundary between a user's own agency and the system's agency. Recent HCI research has therefore expanded the range of the *apparent mental causation model* from attributing agency to oneself to attributing agency to another agent, either a human or a computer system (Coyle et al., 2012; Berthaut et al., 2015).

2.1.3 Explicit and implicit measures for sense of agency

2.1.3.1 Explicit measure: subjective judgements

A sense of agency can be measured explicitly as a sense of control or authorship by asking a person to report their subjective perception directly (Mellor, 1970; Ebert & Wegner, 2010; Coyle et al., 2012). It can be as straightforward as asking the participants to rate on a numeric scale (e.g. "*I allowed that to happen*" vs. "*I intended that to happen*", "*It was not at all me*" vs. "*It was absolutely me*") (Wegner & Wheatley, 1999; Aarts et al., 2005). However, subjective reports on agency experience can be susceptible to different contexts, prior beliefs and task expectations (Gawronski, LeBel, & Peters, 2007). For instance, when a person initiates a gamble rather than letting others initiate it, they would report an exaggerated sense of agency in the gamble, even though the odds remain the same. In other activities in which participants claim to have experienced unwilled actions, like during table turning or Ouija-board spelling, they would report a reduced level of agency experience, despite the source of the observed actions being the participants themselves (Coyle et al., 2012; Wegner, 2003).

2.1.3.2 Implicit measure: intentional binding

The malleability of humans' subjective agency perception urges us to find a more robust metric. The theory of *intentional binding* in cognitive research offers an implicit metric for agency through empirical measurements (Haggard, Clark, & Kalogeras, 2002; Moore & Obhi, 2012; Hughes, Desantis, & Waszak, 2013). A person's perception of time is tightly associated with their intentions and actions, and can be distorted differently depending on whether they are consciously taking an action, or just passively experiencing an action (Libet, Gleason, Wright, & Pearl, 1983; Haggard & Eimer, 1999; Moore & Haggard, 2008; Moore, Wegner, & Haggard, 2009; Moore, Lagnado, Deal, & Haggard, 2009). In particular, a person will perceive the time interval between an intentional action and its corresponding outcome to be *shorter* than the actual interval, and the time between an unintentional action and its outcome to be *longer* than the actual length (Coyle et al., 2012; Moore & Obhi, 2012; Hughes et al., 2013), as illustrated in Figure 2.1. This distortion is a systematic error in human cognition. It is considered to be made up of two components: *action binding* and *outcome binding*, meaning that a person tends to consider a voluntary action as having been taken later than it actually was, whereas its outcome to have appeared earlier than it actually did. Hence, the perceived interval between the action and its outcome will be shorter than its actual length. In other words, the perceived time of an intentional action and the perceived time of its outcome will be attracted together. For an involuntary or unintentional action, on the contrary, the warp in a person's time perception will have a prolonging effect, where a person perceives the action as having been taken earlier than it had while the outcome occurred later than it actually did. Hence, the perceived interval will be longer than the actual one (Ebert & Wegner, 2010).

There are two methods to measure the intentional binding effect derived from this theory. The first is simply asking the user to estimate the length of the perceived interval between an action and its outcome repeatedly, then calculating the average error between the actual interval and the user's estimations (Engbert et al., 2007). This method is easy to apply, and is suitable for experiment tasks that involve visual targets. However, it is less robust in complex contexts, and with it we can only calculate the total binding effect and cannot calculate the action binding and the outcome binding separately.

Another method is the Libet Clock (Libet et al., 1983; Coyle et al., 2012) paradigm. As shown in Figure 2.2, the Libet clock has an appearance of an analogue clock, with a full cycle of 2560ms, twelve evenly distributed number labels (starting from 5 at the direction of 1 o'clock, ending with 60 at the direction of 12 o'clock) along the outer perimeter of the clock face, and a single clock hand rotating in a constant speed. The clock is displayed on a normal computer screen, and is deliberately designed to be small compared with the screen size, therefore the user does not have

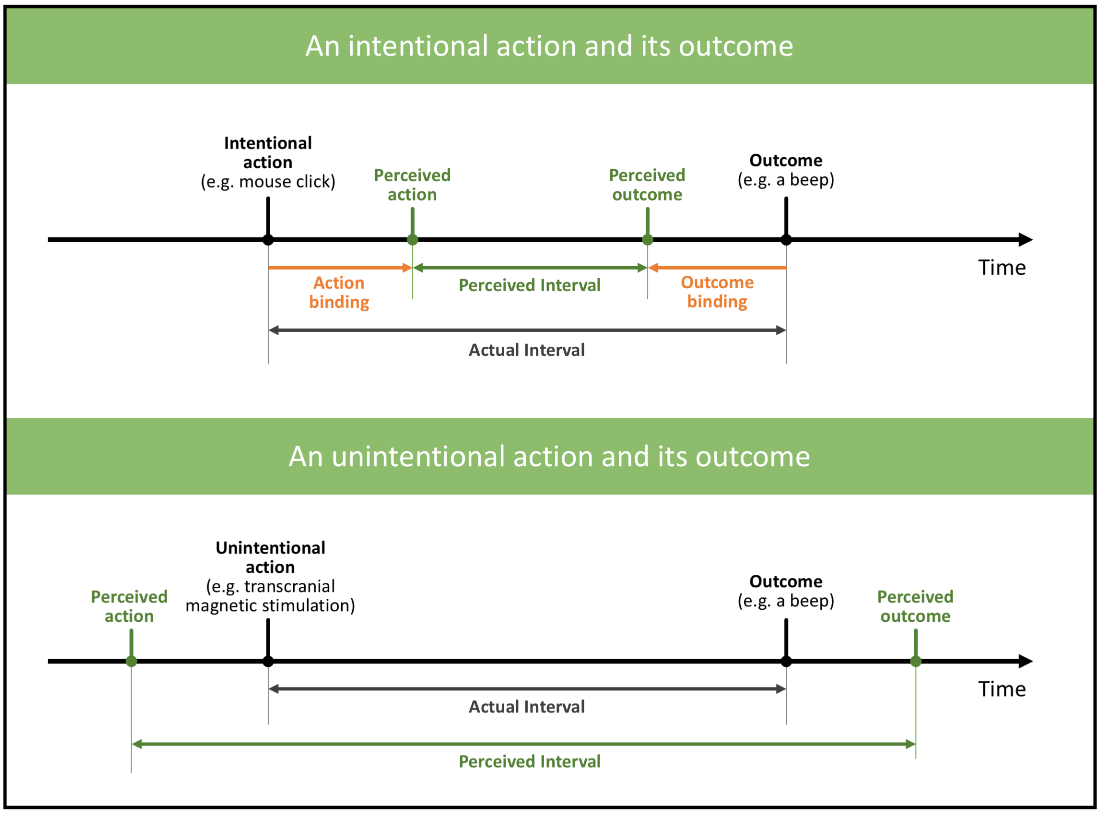


Figure 2.1: Warped time perception during an intentional action and an unintentional action. In this figure, the upper half illustrates that the perceived times of an intentional action and its outcome are attracted together, whereas the lower half shows that the distortion of temporal perception happens in an opposite manner with an unintentional action and an unintended outcome. An example of an intentional action is a voluntary mouse click. An example of an unintentional action is an involuntary movement (such as a muscle twitch) induced by a transcranial magnetic stimulation (TMS) applied on a person's motor cortex. An example of an outcome is an audible beep.

Action phase (Action: key press)		
Baseline error (BE):	A participant is required to press a key whenever they want to while observing a Libet clock. The key press does *not* generate any effect. The participant then reported their perceived clock hand position at which they pressed the key. The actual time of their key press was recorded by the system.	BE = actual time - perceived time
Active error (AE):	A participant is required to press a key whenever they want to while observing a Libet clock. The key press generates a beep sound. The participant then reported their perceived clock hand position at which they pressed the key. The actual time of their key press was recorded by the system.	AE = actual time - perceived time
		Action binding = action(AE) - action(BE) (measure and calculate repetitively)
Outcome phase (Outcome: beep sound)		
Baseline error (BE):	A participant is required to take *no* action whilst observing a Libet clock. The system randomly generates a beep sound. The participant then reported their perceived clock hand position at which they heard the beep. The actual time of the beep was recorded by the system.	BE = actual time - perceived time
Active error (AE):	A participant is required to press a key whenever they want to whilst observing a Libet clock. The key press generates a beep sound. The participant then reported their perceived clock hand position at which they heard the beep. The actual time of the beep was recorded by the system.	AE = actual time - perceived time
		Outcome binding = outcome(AE) - outcome(BE) (measure and calculate repetitively)
		Total binding = Action binding + Outcome binding

Table 2.2: The method for measuring and calculating intentional binding effects

to move their head or eyes significantly when they observe the clock. The user will be asked to report where the clock hand was pointing at when an action and its outcome occurred respectively. Then we can calculate the respective binding effect according to Table 2.2 adapted from the work of Coyle et al. (2012). This measure is more robust and accurate, and the results can reflect the action binding and the outcome binding separately. However, it is not always suitable for empirical application, because it requires a considerable number of repetitive measurements, which means that running the experiment can be time consuming. It also requires the user to devote visual attention to the clock during the experiment, so the task design cannot include important visual information.

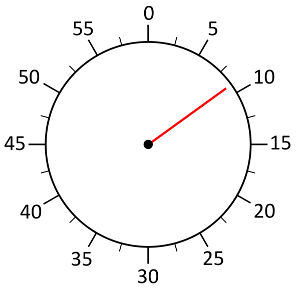


Figure 2.2: The appearance of a Libet clock

2.1.3.3 Are they measuring the same thing?

Existing studies adopt either the explicit measures or the implicit measures (or sometimes both) when investigating the user's sense of agency (SoA). It is found that the results of the two kinds of measures are often, but *not always*, congruent or positively correlated (Moore, Wegner, & Haggard, 2009; Ebert & Wegner, 2010). One theory is that the sense of agency is a concept with heterogeneous aspects, and explicit and implicit measures provide different accounts of SoA (Synofzik, Vosgerau, & Newen, 2008):

1. Implicit measures, such as the intentional binding effect on one's time perception, represent the *feeling of agency* (FoA) aspect of SoA (Synofzik et al., 2008). The FoA is a passive *reflexive feeling* that takes place at a lower level, which is primary and perceptual, but not conceptual (Ebert & Wegner, 2010). Hence

the comparator model introduced in **Section 2.1.1** that relies on sensorimotor information (e.g. feed-forward cues, proprioception, sensory feedback) can explain the FoA well (Synofzik et al., 2008). The FoA only categorises an action as either "self-caused" or "not self-caused", while the concept of "self" here is merely implicitly represented through FoA (Synofzik et al., 2008). In other words, the action is not actively and conceptually attributed by the user to themself. Moreover, it is found that binding effects that arise from FoA are prominent only when the delay between an action and an outcome lasts "a fraction of a second" (Ebert & Wegner, 2010) (e.g. not longer than 200-250ms (Stetson, Cui, Montague, & Eagleman, 2006; Choi & Scholl, 2006)).

2. Explicit measures, such as a subjective report on how much control one feels, represent the *judgement of agency* (JoA) aspect of SoA, which is an active *reflective attribution* process takes place at a higher level (Ebert & Wegner, 2010; Gallagher, 2007). Unlike the FoA, the JoA is an "*explicit conceptual, interpretative judgement*" of oneself being the agent (Synofzik et al., 2008). The JoA arises from rule-based belief formation and authorship attribution processes (e.g. goals, intentions, thoughts, social cues, contextual cues), which require a considerable amount of cognitive capacity (Smith & DeCoster, 1999; Sloman, 1996; Ebert & Wegner, 2010). In some cases simply holding a belief-like mental status can be sufficient to support one's JoA, even though the observed outcome was not caused by one's action (Aarts et al., 2005; Gawronski et al., 2007). This can explain the reason why people would judge an event as caused by themselves even when it is delayed by several seconds (Shanks, Pearson, & Dickinson, 1989; Ebert & Wegner, 2010).

Nevertheless, it is found that both the FoA and JoA contribute to the overall SoA, but how much each of them contributes will depend on the "context and task requirements" (Synofzik et al., 2008; Ebert & Wegner, 2010). This is because the FoA and JoA arise from different levels based on different authorship indicators, such as proprioceptive influences, direct bodily feedforward, visual action feedback and social cues (Wegner & Sparrow, 2004), and some authorship indicators can have a greater impact on the explicit JoA measure than on the implicit FoA measure or vice versa (Moore, Wegner, & Haggard, 2009; Ebert & Wegner, 2010). As a result, it is possible to observe either a positive correlation or a dissociation between the measures for the two aspects (Moore, Wegner, & Haggard, 2009; Ebert & Wegner, 2010).

2.1.4 Factors that influence the experience of agency in HCI

2.1.4.1 Human factors

In human-computer interaction, the user's experience of agency can be influenced by many human factors. For instance, a link has been found between ageing and an impaired perception of personal competence and agency, and a person's preference in how much control they assume (Rodin, 1986; Schieman & Campbell, 2001). Mental diseases like schizophrenia can cause a person to experience the "alien hand syndrome" (which makes them feel their hand is moving due to an external source of power rather than under their own control) (Mellor, 1970; Carruthers, 2007), or hear an inner voice that is actually their own thoughts but they mistake for others' (R. E. Hoffman, 1986). A higher working memory load can impair a person's sense of agency too, suggesting that agency judgements (as a retrospective inferential process introduced in **Section 2.1.1**) are moderated by the availability of conscious cognitive resources (Hon et al., 2013; Limerick et al., 2015; Shneiderman, 2000). A person's involvement in the system implementation, and their familiarity and expertise gained from previous experience in interacting with a computer system can also affect the level of control they would expect and their sense of being in control (Baronas & Louis, 1988; Hardian, 2006; Obendorf, 2009; Shneiderman, 2010; Iacovides, Cox, Kennedy, Cairns, & Jennett, 2015).

2.1.4.2 Taking an action: Input and operations

Although many human factors are often not controllable in interaction design, researchers and designers can still decide how the user takes an *action* by designing a variety of modalities for user input. Different input methods and devices can have a different impact on the user's experience of control. For instance, the user will experience a significantly greater sense of agency when using skin or body-based input compared with a traditional keypad (Coyle et al., 2012). There have also been research and design projects focusing on improving the traditional keypad input method, such as mapping the keypads across difference devices (like a TV remote controller and a computer keyboard) (Brusky, Frederick, & Lininger, 1999). Many well-commercialised game consoles have adopted motion and gesture control, aiming at improving the controlling experience by offering the players a closer coupling between control gestures and real gestures (Francese, Passero, & Tortora, 2012). Within the field of ubiquitous computing, designers have developed "tangible user interfaces" (TUIs) that allow the user to control the computer system by manipulating surrounding physical objects or surfaces (e.g. in office environment like *metaDESK*, *ambientROOM*, and *transBOARD* (Ishii & Ullmer, 1997), or in interactive music system like *Block Jam* (Newton-Dunn, Nakano, & Gibson, 2003)). Speech input is useful when the user needs to control the system while they are visually engaged in other tasks or are motor-impaired (Shneiderman, 2000), but due to the latency and accuracy problems in even state-of-the-art speech recognition systems, as well as the amount of working memory that speech input demands, the user may not experience a strong sense of agency (Limerick et al., 2015). Controlling the computer system through eye gaze has provided opportunities for users who are less capable of using a computer mouse or making movements due to physical conditions (Hutchinson, White, Martin, Reichert, & Frey, 1989; Murata, 2006). While young able-bodied users reported that they experienced less sense of control using eye-gaze or gesture control compared with a mouse and keyboard (Hyrskykari, Istance, & Vickers, 2012), elderly users (aged over 64) welcomed eye-gaze control, especially when they needed to point at a small target: they reported a higher "ease of input" and achieved faster pointing time using eye-gaze control compared with mouse input, and their performance was almost as good as young users' (Murata, 2006).

2.1.4.3 Observing an outcome: Output and feedbacks

We can decide how the system shows the *outcomes* caused by the user's action by offering different presentations of system output, which can also influence the user's control experience. Given that the experience of agency and agency attribution may rely on a cue integration process (as introduced in **Section 2.1.1**) (Wegner & Sparrow, 2004; Moore, Wegner, & Haggard, 2009), a task-dependent, sensorily accurate, and well-timed system feedback is crucial to the user's perception of agency. When designing *output* presentations, there are many ways to code and convey information, such as visual, auditory, tactile feedback (Akamatsu, MacKenzie, & Hasbroucq, 1995; Biocca, Inoue, Lee, Polinsky, & Tang, 2002; Hoggan, Crossan, Brewster, & Kaaresoja, 2009; Iio et al., 2011) or other conversational backchannels (Inden, Malisz, Wagner, & Wachsmuth, 2013; Jung et al., 2013). Each has its advantages and pitfalls given different task requirements. According to the *forward model* of the sense of agency, when there is a mismatch between a person's predicted sensory feedback of an action and its actual sensory consequence, the experience of agency will be impaired. An example is that when the consequences of an action are displaced spatially or temporally, it can distort participants' agency attribution (Farrer et al., 2008).

2.1.4.4 Process: Interaction flow and its nature

Apart from choosing the appropriate output modalities, providing the end user with sufficient and timely system responses in a "flow" manner can also enhance their control experience (Tanimoto, 1990; Church, Nash, & Blackwell, 2010; Nash, 2012; Berthaut et al., 2015). Previous studies have defined six levels of "liveness" of an interaction, as summarised in Table 2.3 adapted from the work of Church et al. (2010) and Tanimoto (2013).

In a complex and autonomous system, the user's control experience can also be influenced by the transparency and the availability of context information of the system status and its working process (Hardian, 2006), as well as the level of autonomy of the intelligent interface/agent (Franklin & Graesser, 1996). For instance, when a computer system can facilitate the user's action, there may exist a "sweet spot" regarding the level of system assistance where the user will experience the strongest sense of control. When the system provides too much assistance, it can impair the user's agency experience instead (Coyle et al., 2012).

During the process of human-computer interaction, the nature of the interaction itself can also significantly contribute to the user's sense of agency. The nature here refers to what *function* an interaction serves, what *goal* it leads to, and what *form* it takes. For instance, when a user is having a conversation with a computer system, if their goal is to make the conversation as engaging and natural as possible, then the interaction itself is both the goal and form, while also functioning as the path to the goal. The user may evaluate their experience of agency based on the vividness, smoothness, immersion or other parameters of the conversation. But if the goal is to let the user make a critical decision assisted by the computer, the conversation will merely be the form of this interaction, and the function it serves may facilitate the decision-making process. Hence, the user's agency may come from the judgement of

Level	Desciption	Features	Example
Level 1 *"ancillary"*	A continuously visible visual representation that supports software design	informative	Drawing a flowchart on a notepad
Level 2 *"executable"*	A continuously visible visual representation that can be manually executed, mapping the user's macro action on the representation and the program's behaviour	informative, significant	Re-compiling an edited program
Level 3 *"edit-triggered"*	A continuously visible visual representation that can immediately respond (e.g. execute and apply changes automatically) to the user's micro action such as editing	informative, significant, responsive	Code completion in programming IDEs
Level 4 *"stream-driven"*	A continuously visible visual representation that is constantly active, presenting the user with all the changes made to the program in a real-time manner	informative, significant, responsive, live	Live coding in music
Level 5 *"one-step-ahead"*	The environment predicts the next programmer action. It stays a step *ahead* of the programmer, rather than lagging behind, or just keeping up with them	tactically predictive	Interacting with a statistical model of a machine learning algorithm
Level 6 *"strategic"*	The environment infers the programmer's desires or intentions and makes strategic predictions, based on which it can predict desirable behaviours and synthesise a program with those behaviours	strategically predictive	Using an IDE that has access to a rich knowledge base and can plan actions towards the programmer's goal

Table 2.3: Six levels of *liveness* during an interaction - describing the feedback flow both in the program notation and in its execution environment

the quality of the decision and the amount and the helpfulness of the aids provided by the computer.

Moreover, the nature of the interaction can determine how the user perceives their relationship with their computer counterpart: it can serve as a neutral information media (Moon, 1999; Guadagno & Cialdini, 2002) (e.g. using instant messaging software), a competitive rival (Williams & Clippinger, 2002; Kätsyri, Hari, Ravaja, & Nummenmaa, 2013) (e.g. playing games with a computer), an adviser, coach or mentor (Desmarais, Giroux, & Larochelle, 1993; Baylor, 2000), a cooperative partner or teammate (Clarke & Smyth, 1993; Nass, Fogg, & Moon, 1996). As will be introduced in **Section 2.2.2**, the perception of the relationship with the computer system can provide the user with a basis for distributing their expectation on either themselves or the computer system. In addition, it determines where the user's agency would come from - the effectiveness and satisfaction obtained from the interaction process, the sense of achievement when getting a task completed, the power when taking control of the interaction, or other aspects of the interaction.

2.2 Expectation in human-computer interaction

During human-computer interaction, the user's interaction behaviours and subjective experience can both be greatly influenced by their expectations and beliefs about their computer counterpart. For instance, if the user is told that they are to interact with a "basic" computer system (e.g. having only a limited vocabulary), they are more likely to adapt their word choice to the computer's words, while they appear to be significantly less adaptive when they *believe* the computer system is "advanced" (e.g. having a copious vocabulary), despite the fact that the computer system exhibits the same capabilities (Pearson, Hu, Branigan, Pickering, & Nass, 2006). Previous research has also found that the user will be less likely to blame a computer system for incorrect decisions when they perceived the computer to be similar to themselves, and the user will also be more likely to share credit with the computer for successful attempts during the task (Moon & Nass, 1996; Bonito, Burgoon, & Bengtsson, 1999). Another study suggested that compared with playing games with a human opponent in present, playing with a computer opponent can induce a higher level of aggression on the user, and such aggression may be reduced by humanising the computer counterpart (Williams & Clippinger, 2002). HCI researchers explain such phenomena using theories of *expectation* from social psychology.

2.2.1 Definition of expectation

Expectation refers to the anticipated events in the external world or the anticipated behaviours of others, and those anticipated events or behaviours are "typical, modal or normative" (Bonito et al., 1999). In social psychology, studies on expectation are usually built on a more abstract concept, *expectancy.* Expectancy in interpersonal interaction refers to "an enduring pattern of anticipated behaviour" (Burgoon, 1993), which can be either verbal or nonverbal. Burgoon claimed that there are two kinds of expectancy. One is the *central tendency*, which is the regularity and predicability of a behavioural pattern in a certain culture or environment. The other is the *idealised standards of conduct*, under which one's behaviours are considered as "appropriate, desired or preferred" (Burgoon, 1995).

Expectancy is almost always associated with interpretations and evaluations, which will interfere with further inference-based expectancies (Burgoon, 1993; Fişek, Berger, & Norman, 1995; Scherer, Zentner, & Stern, 2004; Tzur & Berger, 2009). A study has shown that participants' pre-induced expectations (either positive or negative) about their interaction partner resulted in an either pleasant or unpleasant interaction. The pre-induced expectations formed such a persistent impression that they had an impact even on the participants' subsequent conclusions, irrespective of how that interaction partner actually behaved in the experiment (Burgoon & Poire, 1993). This result suggests two things. First, expectancy is a *framing device* that "defines and shapes" social interactions, and a person would choose the way they communicate with others based on how they anticipate others' communication style to be. Second, expectancy is a *perceptual filter* that decides how social information is processed (Burgoon, 1993).

The disconfirmation between a person's expectation and the actual event/behaviour is called *expectancy violation.* Expectancy violations have valences, which can be either positive or negative. Violations have positive valence when the consequences go beyond a person's expectations in a desirable way. A positive violation of an expectation may cause a surprise effect, in which relevant behaviours may be more influential than those that only conform to the expectations. Violations with negative valence occur when the consequences fail to live up to a person's expectations, and they are more likely to have a negative effect on the interaction than when the expectations were not violated.

Theories of expectation have been applied to human-computer interaction (Bonito et al., 1999). Expectations influence human-computer interaction in two ways. One happens *before* an interaction, the other *after*.

- *Before* an interaction, the user will evaluate their own ability to contribute to the current task by comparing themselves to their human partners or the computer counterpart. By anticipating how well each member including themselves would perform and how much each other would contribute, the user distributes their expectations to each member. The user can then determine how much they want to get involved and how much influence they can have. Sometimes the user may put higher expectation on their computer counterpart and anticipate it to contribute more than other human partners. This is called the "expectation advantage" (Bonito et al., 1999), though it is likely to backfire if the system fails to live up to the expectation.
- *After* the interaction, the user uses their expectations as a reference when they assess the behaviours during the interaction just now. For a given behaviour or event, the user would interpret it based on how they expected it to be and whether their expectations were violated. When the behaviour or event violated their expectations, either positively or negatively, the user would re-interpret and re-evaluate the situation, update their expectations towards it, and adjust their future behaviours and contribution accordingly (Bonito et al., 1999).

2.2.2 Factors that shape and update expectations

There are two major theories that can throw light on the mechanism by which expectations operate during an interaction, namely the *expectation states theory* (EST) and the *expectancy violation theory* (EVT). Both theories complement each other. The expectation states theory (EST) focuses on how a person uses the initial distribution of expectations in a task-oriented group when predicting the consequences of their interaction (Skvoretz, 1988; Balkwell, 1991; Fişek, Berger, & Norman, 1991; Fişek et al., 1995). As described in **Section 2.2.1**, this applies *before* an interaction. In a heterogeneous group where members have salient status characteristics (e.g. sex, age, class), a member would evaluate the amount of contributions as well as the usefulness of the contributions differentially when it comes to evaluating themselves and the others, taking into account the characteristics of each member and their ability to participate and exert influence. This “state organising process” forms a network of expectation attributions that can moderate members’ behaviours. For instance, members with a low status would be given less chances to speak, and when they spoke their comments would also be given less attention and credit.

In a task-oriented status-homogeneous group, members’ expectations are not formed based on status characteristics but behaviour patterns instead. Researchers in social psychology started by investigating the effect of consistent behaviours patterns of *initiator-reactor* or *leader-follower* between group members on the establishment of task-based status order (Berger & Conner, 1969; Fişek et al., 1991). Later works developed a more general *behaviour interchange patterns* (BIP) model, which maintained that the status of members can be determined dynamically: when member A performs “an initial action that is potentially influential” on the task outcome, other members may either agree or disagree with that action. For the members who agreed with A’s initial action, A is status-superior to them. Whereas for those who disagreed with A’s initial action, they become status-superior to A. In this way, members form expectations through behaviour cycles, and their expectations will in turn structure a *power-and-prestige order* (Fişek et al., 1991, 1995). This uneven distribution of expectations will lead to differential participation, and will shape the perception of leadership, deference, agreement, evaluation of others (Webster Jr, Hysom, & Fullmer, 1998; Bonito et al., 1999).

On the other hand, the expectancy violations theory (EVT) is looking at the degree to which expectations modulate the perceiver’s behaviours (Burgoon, 1978, 1995; Burgoon, White, & Greene, 1997). Echoing **Section 2.2.1**, this theory applies *after* an interaction behaviour has occurred. The EVT holds that a person would assess a behaviour depending on how they expect it to be, and when it violates their expectation, their cognitive process will be activated to make a deeper assessment of the behaviour, its meaning and function. This can result in an intensified response from that person. The EVT also proposes that people shape their expectations and make predictions towards an interpersonal interaction based on three information sources (Burgoon, 1993, 1995), which help them judge which behaviour is relevant and should be expected, and which is not:

1. Communicator characteristics. For instance, the physical appearances, gender,

personality, social skills, language style, task expertise, and socio-demographic background of the communicator. For instance, a negative behaviour committed by a person who was held in high regard would be perceived as less acceptable, compared with a same behaviour of someone who was poorly regarded (Burgoon, 1993).

2. The nature of the relationship. For instance, the similarity, familiarity, attraction, power differentials (i.e. being equal or not) among interactants or group members. One example is that when there is a power asymmetry, the less powerful person is expected to show more respect and deference towards the more powerful person (Burgoon, 1995).

3. Context. For instance, the physical surrounding, culture, privacy, formality, and the nature and the goal of the task. For instance, different cultures or organisations may have different "power distance", and the difference in how people view power relationships can result in different expectations (Hofstede, 1984).

2.2.3 The basis of temporal expectation in cognitive neuroscience

The theories of expectation can inform the study of agency in many aspects. As the *Shneiderman's Eight Golden Rules of Interface Design* says, "[e]xperienced users strongly desire the sense that they are in charge of the interface and that the interface responds to their actions. They don't want surprises or changes in familiar behaviour, and they are annoyed by tedious data-entry sequences, difficulty in obtaining necessary information, and inability to produce their desired result" (Shneiderman, 2010). In this context, "surprises" or "changes in familiar behaviour" are a kind of expectation violation, which can cause a diminishing effect on the user's experience of control. In order to approach the three research questions proposed in **Chapter 1**, this PhD research will explore how *temporal* expectations and the violation of them may impact the user's agency experience during mixed-initiative interaction.

Our brain is a predictor. It can anticipate not only the content or features of forthcoming events, but also their timing (Nobre, Correa, & Coull, 2007). For animals that are as intelligent as human beings or as primitive as rats, there is both psychophysical and neurophysiological evidence to support the theory that brains are able to extract temporal regularity and patterns out of external stimuli, and that they are able to use such patterns to predict the timing of forthcoming events (Barnes & Asselman, 1991; Barnes, Collins, & Arnold, 2005; Shuler & Bear, 2006; Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008).

Temporal expectations are observed across different areas of the brain. The neurons in the *motor-related* areas can fire in synchrony with the temporal regularity of external stimuli. This synchrony can trigger anticipatory saccades and shorten the latency of saccades, as well as facilitate general motor preparation and execution (Reuter-Lorenz, Oonk, Barnes, & Hughes, 1995; Riehle, Grün, Diesmann, & Aertsen, 1997; Barnes et al., 2005). The *sensory* perception areas on the cortex are able to make faster detection and discrimination responses (Niemi & Näätänen, 1981; Nobre et al., 2007) as the variability and uncertainty of event timing decrease. Temporal regularities can be learnt, and the cortical regions responsible for *learning and motivation* are associated with the predictions of the reward delivery time of future events (Barnes et al., 2005; Medina, Carey, & Lisberger, 2005; Shuler & Bear, 2006; Tsujimoto & Sawaguchi, 2005).

Neuronal entrainment towards an external temporal structure is considered as the core mechanism that allows attentional biasing (Schroeder & Lakatos, 2009). One experiment has shown that when external visual stimuli appear in a rhythmic manner, the brains' low-frequency neuronal oscillations can entrain ("phase-lock") to the temporal pattern of that stream. This temporal synchrony will result in an amplified neuronal excitability for task-relevant events and a decreased reaction time, hence optimising the subjects' attentional selection (Lakatos et al., 2008). Another example is that a person's visual perception (e.g. detection and discrimination) can be enhanced by temporal expectation. This is because the oscillatory activities in neurons in the primary visual cortex have entrained to the temporal structure of external signals, and the anticipated temporal structure can improve the quality of sensory information and accelerate the accumulation of relevant evidence. Therefore, temporal expectation can facilitate the sensory processing of the events that occur at a predictable and expected beat and enable a person to make perceptual decisions faster and more accurately (Rohenkohl, Cravo, Wyart, & Nobre, 2012).

It is worth noting that *attention* and *expectation* can both enhance signal detection, facilitate pattern recognition and improve information processing, but they modulate relevant neuronal oscillations in a different manner. Attention is *relevance* driven and increases energy sensitivity for signal-present stimuli, whereas expectation is *probability* driven and increases energy sensitivity for signal-absent stimuli (Fujioka, Trainor, Large, & Ross, 2009; Wyart, Nobre, & Summerfield, 2012). This can be supported by the results from electroencephalography (EEG) and magnetoencephalography (MEG) studies: attention can sharpen the neuronal firing pattern, which possesses a higher signal-to-noise ratio, so it helps achieve neuronal entrainment and reduces internal noises. Expectation, on the other hand, biases the baseline of signal selection, thus constraining the acquisition and interpretation of inputs to a more limited population (Summerfield & Egner, 2009; Rohenkohl et al., 2012).

As a very basic parameter of any interaction, *timing* can not only reflect the dynamics during an interaction but also interfere with it. As introduced in **Section 2.1.3**, a person's timing perception can be warped by their sense of agency and therefore can serve as a valid metric for agency: but does the timing have a potential to affect the sense of agency reversely? For instance, if the timing of events during an interaction is predictable and aligned with the user's expectation, will it preserve the user's experience of agency? If the event's timing violates the user's expectation, will the user feel they are losing control over the interaction, and therefore experience a loss of agency? Since we know that when the timing of an event is predictable and expected, a person would be relieved from extensive sensory processing, thus releasing more cognitive resources for the cognitive processes that occur on a higher order, it does not take a great leap to ask if those released cognitive resources can contribute to the agency attribution/experience process.

2.3 Rhythmic entrainment during interaction

As introduced in **Section 2.2.3**, with the help of temporal expectation, a person can anticipate the timing of forthcoming events in a rhythmic sequence, and be able to notice when there is a missing or wrongly-timed event. In other words, being rhythmic makes a process predictable, and potentially more "under control".

2.3.1 Definition of rhythm

The term *rhythm* is used in many disciplines to describe a process. For instance, rhythm can be found in the sound of a piece of music, the movements in dancing, the steps in walking, the flow of speech, the electrical activity in muscles, the oscillation of neurons and the cycle of our biological clock. The definition of rhythm has yet to come to an agreement across different fields and contexts, but here is one version that is adopted in music, conversation and linguistics studies: *"rhythm is the systematic patterning of sound in terms of timing, accent, and grouping"* (Patel, 2010). To apply this to a broader context, rhythm can be the systematic patterning of *events* in terms of timing, accent, and grouping. Rhythm is often interchanged with the term *periodicity* (Patel, 2010), but "being periodic" is a narrower criterion: it requires the events of a process to repeat regularly in time, and does not take the pattern in accent or grouping into account. Hence all periodic processes are rhythmic, but not all rhythmic processes are periodic.

For a real-life process to be called "rhythmic", the granularity of the time scale and the strictness of criteria vary from discipline to discipline. In chronobiology, "circadian" rhythm refers to a roughly 24-hour cycle of a living organism, and if the recurrence period is longer than an Earth day, the rhythm is "infradian". "Ultradian" rhythm covers a much wider range, from a few minutes to up to 12 hours, an example is the rapid eye movement cycles in sleep, which last for about 90 minutes each. A "supra-ultradian" cycle can be a few seconds or even milliseconds, such as our heart rate and pulse (normally 60∼100 beats/min (Spodick, Raju, Bishop, & Rifkin, 1992)), and spoken languages that contain structural information (e.g. differentiation of consonants) of very fast rhythm (20∼50ms) (Clayton et al., 2005; London, 2012b). Because of the natural fluctuation in biological processes or communication, the rhythms above are not bound by an ultra-precise recurrence period or grouping pattern.

The criteria of being rhythmic in music performance may be higher than other fields, because professional musicians' sensitivity to temporal deviation is significantly greater than that of non-musicians', and their preferred quantities are very close to their perceptual threshold (Sundberg, Friberg, & Frydén, 1991). However, it is well recognised and accepted that "musicians never perform rhythms in a perfectly regular, or mechanical, fashion" (Large & Kolen, 1994), and the timing variability can be either intentional (e.g. for expressive purposes) or unintentional (e.g. lack of experience)

(Shaffer, 1981; Sloboda, 1983; Todd, 1985). In musical ensembles, co-performers may drift towards or away from each other's rhythm, despite the fact that 20∼30ms is the perceptual threshold for detecting asynchronous onsets (Goebl & Parncutt, 2002; Goebl & Palmer, 2009). Given those temporal deviations and variabilities, both musicians and listeners are still able to perceive meter in rhythms in music performances (Large & Kolen, 1994; Leman, 2012; Nowicki et al., 2013).

According to psychoacoustic research, a person's auditory perception of rhythm and their ability to adapt to it vary with the length of intervals (Drake & Botte, 1993), which falls roughly into three categories: very short (<400ms), moderate (400∼1500ms) and long (>1500ms) (London, 2012b). Our ability to hear rhythmic patterns is also limited to a range between 100ms and around 6 seconds, within which regular intervals of around 600ms will form a "maximal pulse salience" zone (London, 2012a). For very short intervals (100∼300ms), it is difficult to distinguish individual beats but we are still able to judge the difference in durations and in numerosity, and for very long acoustic intervals (>1500ms), we are less able to hear them in a coherent cycle but tend to perceive them as isolated acoustic events instead.

In addition, a person does not form their rhythmic perception passively. They will subjectively differentiate the accentuation of identical beats or tones, and actively group them into twos, threes or more (London, 2012b). This process is called "subjective metricization" (London, 2012a). *Grouping* requires a person to organise the durations of intervals; specifically, the intervals between sound onsets. It is often measured as the *inter-onset interval* (IOI). Accentuation, loudness, timbre, pitch and/or duration are all cues that one can use when determining the boundaries of groups (London, 2012b).

2.3.2 The role of rhythm: an attribute and a design resource

One common perspective that researchers from different disciplines take is to study rhythm as an attribute of a process or a series of events. For instance, the rhythmic structure of music can be a robust classifier for music genres. In classification tasks, computer algorithms can achiever higher accuracy than human participants by calculating the regularity of the rhythm, the temporal features in short/medium/long timeframes, the relation of the main beats to the sub-beats, and the relative strength of beats (Tzanetakis & Cook, 2002; Meng, Ahrendt, Larsen, & Hansen, 2007). Linguists and phonologists have long been studying how to classify languages based on their rhythmic properties. A classical framework claims that Germanic languages are stress-timed (e.g. English, Dutch), Romance languages are syllable-timed (e.g. French, Spanish), and other languages are mora-timed (e.g. Japanese). Researchers are still working on designing newer temporal parameters for classification (David, 1967; Dauer, 1983; Barry, Andreeva, Russo, Dimitrova, & Kostadinova, 2003). Rhythm also plays an important role in language perception and acquisition. Two studies found that infants are able to discriminate between their native language and foreign languages using rhythmic information (Nazzi, Bertoncini, & Mehler, 1998; Ramus, Nespor, & Mehler, 1999). In system security research, the keystroke dynamics of the user, especially the rhythm in it, have been used as a non-intrusive biometric authentication (Karnan, Akila, & Krishnaraj, 2011; Banerjee & Woodard, 2012).

Apart from being a static classifier, rhythm can also play a more active and dynamic role. As introduced in **Section 2.2.3**, neuronal oscillations can synchronise with rhythmic stimuli in the external world, which can facilitate sensory processing and reduce a person's cognitive load. Rhythm can also energise and co-ordinate bodily movements as well as stir up or soothe emotions (Panksepp & Bernatzky, 2002). One example is that people may find it difficult not to nod their heads or sway their bodies when listening to dance music with strong rhythms (e.g. a Strauss waltz, marches, techno beat) (Scherer & Zentner, 2001). It was also found that Argentinian lullabies could significantly decrease infant listeners' heart rates, and their breathing rhythm would synchronise with the lullaby rhythm, whereas jazz music could not have such effects (Kneutgen, 1970). This effect was further supported by a recent study on young healthy adults, who showed instantaneous corresponding cardiovascular/respiratory responses that "mirrored" the changes of music profiles (Bernardi et al., 2009). In addition, auditory rhythm can have therapeutic effect on stroke patients by balancing their muscle activation temporally thereby improving their stride symmetry (Thaut, McIntosh, Prassas, & Rice, 1993). Rhythm is often found in speeches that are perceived as overwhelming and persuasive (Woodall & Burgoon, 1981). In recent decades, more research in musicology and social psychology started to focus on the interaction between two or more rhythmic processes. It was found that the performers or interactants involved tend to establish a mutual adaptive timing pattern, and that rhythmic interaction can positively influence their co-ordination and co-operation, enhance empathic perception and facilitate pro-social behaviours (Spiro

& Himberg, 2012; Cross, 2013; Hawkins et al., 2013).

Now that we know that rhythm can be used to express one's own emotions and to produce emotions in others, and it has co-ordinating effect on cognitive, physical and social levels, it is a natural step forward to postulate that rhythm may also play a significant role in human-computer interaction. As discussed in **Section 2.1.4**, in order to give the user a sense of agency, traditional human-computer interaction research has intensively investigated the most appropriate ways to allow the user to take actions or to present them with desirable outcomes. However, very few studies have looked into the temporal aspects of human-computer interaction, and even fewer have considered the rhythmic dynamics of the interaction as a manipulatable factor. Therefore, this dissertation will establish rhythm as a design resource, and investigate its effect on the interaction between human and computer, particularly in mixed-initiative interaction.

2.3.3 Entrainment effects between rhythmic processes

The mutual adaptive timing pattern between two or more rhythmic processes mentioned in **Section 2.3.2** can be explained by the *entrainment* theory. Entrainment refers to a process in which two or more rhythmic processes interact with and adapt to each other, and eventually act in a relatively stable synchrony (Clayton et al., 2005). In other words, the two or more oscillators *lock up* to each other. This effect is often referred as "co-ordination", "alignment", "convergence" and "synchronisation" in different areas of literature (Pearson et al., 2006).

The study of entrainment phenomena originated from physics (Rosenblum, Pikovsky, & Kurths, 1996), and mathematical models were built to describe how two or more chaotic oscillators become coupled through phase synchronisation. This concept has since been generalised and applied in musicology, neuropsychology and social psychology (Auer, Couper-Kuhlen, & Müller, 1999; McGrath & Kelly, 1986). For example, human's social interaction could be considered as a rhythmic process that carries meaning and conveys intentionality (Lenneberg, 1967). Entrainment in this context may occur in many different forms on a spatiotemporal dimension without the co-actors' intentional effort. This includes mirroring each other's postural dynamics during a conversation (Shockley, Santana, & Fowler, 2003), co-ordinating handheld pendulum swinging while solving co-operative puzzle tasks together (Richardson, Marsh,

& Schmidt, 2005), synchronising gait when walking side-by-side (Nessler & Gilliland, 2009), dancing together (Leman, 2012), rocking chairs together (Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007), tapping a finger with others (Repp, 2005; Himberg, 2006; Spiro & Himberg, 2012), aligning to each other's speech prosody (e.g. energy, pitch, speaking rate) and word choice (Richardson et al., 2005; Levitan & Hirschberg, 2011; Levitan, Gravano, & Hirschberg, 2011), as well as adapting to each other during musical performance - either among the musicians themselves (Goebl & Palmer, 2009), or between the performers and the audience (Large & Kolen, 1994; Clayton et al., 2005), such as the rhythmic applause in concert halls (Néda, Ravasz, Vicsek, Brechet, & Barabási, 2000).

Entrainment is not just a co-ordination between sensorimotor systems of each actor in the interaction. It can also be a mutual agreement between cognitive processes that involves perceptual synchronisation and adjustment. Jones and Boltz (1989) proposed a three-stage cognitive mechanics of temporal entrainment. In the first stage, the listener forms *perception*, which prepares them to shape expectations. Then if the expectations are met, the listener enters the *synchronisation* stage. If the expectations are violated, the listener enters the *adjustment/assimilation* stage instead. Jones and Boltz (1989) also established a *dynamic attending* model, which claims that a person has two attending modes: one is *future-oriented* attending, the other is *analytic* attending. Future-oriented mode will be adopted when the temporal structure of external stimuli is coherent and predictable. This mode supports anticipatory behaviours on a relatively long time span. Analytic mode is switched on when the temporal structure of the stimuli is incoherent and complex, and it is hard to form expectations. This mode focuses on grouping and counting adjacent events in a relatively shorter time span (Clayton et al., 2005).

It has been found that social entrainment on a cognitive level can produce *intersubjectivity* - a common ground that supports "the sharing of subjective states by two or more individuals" (Jones & Boltz, 1989; Schegloff, 1992). Furthermore, entrainment is closely related with the ability and experience of *empathy*, because both call on a process of mirroring and simulating another mental system (Gallese, 2001). Empathy is a trait that needs to be defined on both a cognitive and emotional dimension (Davis, 1980, 1983). Its cognitive dimension emphasises the capability and experience of *perspective taking* (i.e. "seeing the world from another person's perspective"), while its emotional component can recognise and share others' emotions

(Gallese, 2001, 2003). Through the experience of entrainment and empathy, people are able to understand and share others' intentions, emotions and actions automatically and promptly.

Rhythmic entrainment during interaction lays the foundation for mutual trust and predictability (Keller et al., 2007; Pecenka & Keller, 2011; Nowicki et al., 2013), which can not only make an interaction smooth and successful, but also allow interactants to "relax to a state of equilibrium" (Richardson et al., 2005; Clayton et al., 2005). An interaction with such relaxation will allow interactants to release cognitive resources for other things such as problem solving and to experience a sense of enjoyment (Hawkins et al., 2013; Repp & Su, 2013; Keller, Novembre, & Hove, 2014; Gallotti, Fairhurst, & Frith, 2017). Entrainment can further enhance one's memory for relevant features of the interaction (Macrae, Duffy, Miles, & Lawrence, 2008), boost their pro-sociality and positive affect (Spiro et al., 2013), and build the sense of mutual affiliation and rapport (Hove & Risen, 2009; Miles et al., 2009).

Two processes can be synchronised and locked up together when they are proceeding in parallel or in turn. The latter case is also known as *alignment*, which is the relative juxtaposition of the components in a rhythmic system along a same timeline. It has been found in early conversational studies that under certain circumstances, entrainment may occur "across turn boundaries". For instance, when it comes to the listener's turn to talk, they may speak in a rhythm that is precisely aligned to the rhythm set by the previous speaker (Couper-Kuhlen, 1993; Auer et al., 1999). Furthermore, if the listener is engaged in the conversation, their utterance of a preferred answer (e.g. confirmation, agreement) will land on the beat that is aligned with the rhythmic accent pattern of the speaker's speech, while the utterance of dispreferred response (e.g. hesitation, disagreement) will be misaligned or delayed (Hawkins et al., 2013). In a broader sense, such alignment can serve as a cue for the response matching (Tognoli, 1969) and conversation repair (Schegloff, 1992) mechanism, which can resolve ambiguity during the interaction. Whilst adept entrainment in communication is correlated with positive social affects and evaluation and strong interpersonal attraction, overly rigid and precise entrainment may not be as effective - in fact, moderate rhythmic interaction that is not too "perfect" is perceived and evaluated most positively (Clayton et al., 2005; Warner, Malloy, Schneider, Knoth, & Wilder, 1987).

Those findings and implications have not yet been fully recognised by or applied to human-computer interaction design or research. Hence, this dissertation will investigate what will happen when a computer entrains to (or goes astray from) the user's rhythmic operations in mixed-initiative interaction, and observe the user's agency experience during this convergence or divergence in terms of rhythm.

2.3.4 Rhythm and entrainment in human-computer interaction

While the role of rhythm during human-human interaction has been investigated thoroughly, the work on the rhythmic aspects in human-computer interaction is relatively thin. Keystrokes and mouse clicks are the most typical interaction behaviours that can appear rhythmic, because those events can be organised in terms of timing (e.g. the timestamp of mouse clicks and keystrokes), accent (e.g. the pressure of mouse clicks and keystrokes), and grouping (e.g. double clicks, combination of keys). There are three major lines of research topics related with the temporal aspects of keystrokes and mouse clicks: the first focuses on the usability of certain software or devices, the second is affective computing, and the third is biometric authentication.

In usability studies, the user's typing and clicking behaviours are recorded. Their speed, inter-stroke interval length and variation, task performance, success and error rates, and subjective ratings on the tasks, the devices and the contexts are frequently selected as either the measurements for analyses or the objectives for optimisations (Akamatsu & MacKenzie, 1996; Chertoff, Byers, & LaViola Jr, 2009).

In affective computing, more physiological and psychological measurements are adopted. When the user is interacting with a computer system, we can collect and track the pressure and timing parameters of their keystrokes and mouse clicks, their ratings on psychological metrics, their facial expressions, their electroencephalography (EEG) and electromyography (EMG), their skin conductivity, heart rate and breath rate, eye gaze in real time. The combination of some of those parameters can effectively reflect the user's affective states, such as being stressed or confused, being relaxed or satisfied (Epp, Lippold, & Mandryk, 2011; Kołakowska, 2013; Picard & Picard, 1997).

Biometric authentication leverages the uniqueness of individuals' typing pattern, such as the finger pressure distribution and the rhythm/timing dynamics, as a biometric fingerprint. After sufficient training using machine learning algorithms, keystroke-based biometric authentication can be more reliable in verifying the user identity compared with traditional password authentication. An additional advantage is that this method is non-intrusive for the user, and it does not add much cost to implement on current computer systems - because almost all personal computers have at least a keyboard and/or a mouse (Karnan et al., 2011). Most of the research effort on this subject is dedicated to finding better statistical classifiers that can reduce the likelihood of authentication errors (e.g. the false rejection rate as a type I error, the false acceptance rate as a type II error) (Yu & Cho, 2004) and minimising the amount of training data the algorithm will ask for (Kang, Park, Hwang, Lee, & Cho, 2008).

As we can see, all three lines of research use the rhythm of keystrokes and mouse clicks as a static attribute of the user's interaction behaviours or status, and have not considered the rhythm as a design resource. However they can still inform this research by providing standard methods and parameters to record and measure the rhythm, which will be manipulated as an independent variable in the experiment design in the later chapters.

The term "entrainment" has been used in human-computer interaction research, but it was in fact first used to describe a well studied phenomenon in social psychology, the "chameleon effect" (Chartrand & Bargh, 1999). During human-human interaction, interactants will unintentionally mimic their interaction partner's verbal or nonverbal behaviours, such as facial expressions, postures and mannerisms. Similarly, in some human-computer interaction studies, participants appeared to adapt or accommodate to the computer system, which can be as plain and simple as a text-dialogue interface or as vivid and complex as an anthropomorphic robot, by altering their own verbal and non-verbal behaviours. This includes their choice of words (Pearson et al., 2006), the length and complexity of the phrases and sentences they use, their facial expressions, their affective responses, their vocal interaction features such as the prosodic contours of speech (Breazeal, 2002), and their gaze and gesture (Iio et al., 2011).

While the majority of research in human-computer entrainment has not considered the requirements on temporal aspects, such as the alignment of rhythm between two series, there are a few studies that paid attention to the timing. Breazeal (2002)'s study on human-robot interaction did observe the temporal entrainment between their participants and the robot *Kismet*; specifically, their participants gradually entrained to the robot by adjusting the timing of their turn-taking, in order to avoid interruptions or awkward pauses in the conversation. However, from the time codes (based on video recordings) presented in the paper, we can see that they adopted a loose and qualitative criterion to define temporal entrainment, such as having longer pauses to wait for each other's response, using shorter phrases, and producing a smoother flow with less mutual interruptions and pauses. Moreover, this line of research still viewed the entrainment phenomenon as just a *product* of the interaction, rather than as a *design resource* that can be manipulated and have an active effect on the interaction process and experience.

Another line of studies on temporal entrainment aim to improve the sociability of the computer counterpart in human-computer interaction. Some of the recent research projects took a much closer look at the effect of temporal entrainment of an embodied conversational agent (ECA) or a robot on the user's experience (Inden et al., 2013; Inden, Malisz, Wagner, & Wachsmuth, 2014). Those studies compared different algorithmic models for generating the timing of an ECA's or a robot's backchannels (e.g. visual cues like head nodding, verbal cues like "um", "yeah", etc.). The first one is fully randomised; the second one is copying the timing of the human participant; the third is entrained to macro-timing distributions based on the human participant's utterances or pauses; the fourth is entrained to micro-timing distributions based on the nearest rhythmic event from the human participant, such as the onset of a vowel or an eye blink; and the fifth is a combination of the third and the fourth. The results were mixed: while the manipulation of timing did not influence the perceived rapport of the ECA, the ECA that copied the human participant's timing was perceived as having higher attractiveness than the ECA that made responses at random times. When the ECA copied the timing of the human participants or entrained to the macro-timing of their speech, it was perceived as missing fewer opportunities of giving timely backchannels (like nodding or saying "um") (Inden et al., 2013). Another human-robot interaction study showed that the user would like a song more if their robot companion was swaying "on-beat" with that song (compared with swaying "off-beat") (G. Hoffman & Vanunu, 2013). Although the authors of that paper discussed their findings in terms of robotic social referencing and perceived similarity, those phenomena may also be explained on the basis of entrainment: the user may have entrained themselves to the rhythm of the music, thus they perceived the on-beat movements of the robot as entraining to their rhythm, hence the rhythmic co-ordination may have produced the sense of enjoyment (Repp & Su, 2013; Keller et al., 2014; Gallotti et al., 2017).

This PhD dissertation will complement the existing literature and push the boundaries of the HCI discipline by firstly, proposing rhythmic entrainment as a design resource in human-computer interaction, and secondly, investigating how rhythmic entrainment can influence the user's sense of agency in mixed-initiative interaction.

CHAPTER 3

Research framework

After reviewing the relevant literature in the last chapter, I now present the research framework that guides the development of my PhD research.

In order to answer the three questions proposed in **Chapter 1**, four sets of hypotheses are formulated in this chapter based on existing theories of rhythmic entrainment in social psychology and the theories of the sense of agency and temporal expectation in cognitive neuroscience. The hypotheses have been tested empirically in controlled laboratory experiments, which will be reported in the following three chapters. The main considerations of adopting an empirical research approach are as follows:

1. The causal link between the timing of mixed-initiative interaction (MII) and the user's sense of agency may be confirmed through the manipulation of the timing and the measurements of its effects on users' agency experience in controlled experiments, which can exclude confounding factors and guarantee a satisfying level of internal validity. The work by Coyle et al. (2012) has shown that the metrics of the sense of agency established in cognitive neuroscience are valid and informative when studying the user's experience of control in HCI. However, previous research in mixed-initiative interaction, as reviewed in **Chapters 1** and **2**, has not specifically studied what factors can affect the user's sense of agency, or provided any standardised empirical measures for agency. This research will fill this gap and confirm the causality with empirical evidence.

2. This research demonstrates that the timing of interaction can be a design resource, and aims to provide empirically supported quantitative standards for it. Very little research has deliberately manipulated the timing of mixed-initiative interaction, and existing design assumptions on interaction timing are rarely formulated on the basis of empirical evidence. For instance, quantitative assertions like "[t]o be interactive the training part of the loop must take less than five seconds and generally much faster" (Fails & Olsen Jr, 2003) are still based on the belief "the *faster* the better."

3. Through hypothesis testing, this research can provide HCI researchers with specific and applicable design guidelines. In the existing body of literature, design guidelines or principles for mixed-initiative interaction are expressed in a broad and vague manner. For example, "developing automated services that are performed *in line with* (emphasis in original) a user's activity, allowing users to take advantage of contributions provided by a system while they work in a natural manner" (Horvitz, 1999b) and "[s]ometimes activity should occur at a certain time, rather than in response to an external event" (Wolber & Myers, 2001). Statements such as these can hardly be translated into practical advice during implementation.

4. During controlled experiments, a large amount of behavioural and subjective data of good quality can be obtained. In addition to hypothesis testing results, the data also provides opportunities for post-hoc analysis, which may offer broader insights to and trigger further discussions in the HCI community.

3.1 Perceived control from predictable rhythm

As reviewed in **Section 2.1.1**, one theory holds that a sense of agency arises from a retrospective inference process, which calls on available cognitive resources. Consequently, if the mind is "pre-occupied" with a high cognitive load, a person may experience a reduced sense of agency (Hon et al., 2013). Also as reviewed in **Section 2.2.3**, if external stimuli appear in a predictable rhythm, a person can form temporal expectations that allow more efficient sensory processing and boost attentional selection (Lakatos et al., 2008; Rohenkohl et al., 2012), hence more cognitive resources can be released. In other words, a person's experience of agency may be impaired by high

cognitive load, while rhythmic stimuli can lower their cognitive load. Drawing on the two stances above, I propose the first hypothesis as follows:
$\boldsymbol{H_{MII} - 1}$: Predictable rhythm in mixed-initiative interaction will preserve the user's perceived control, whereas irregular time intervals will impair their perceived control.

Section 2.1.3 introduced two ways to measure the user's sense of agency empirically. It can be measured explicitly by simply collecting a person's subjective report of perceived control, such as letting them give ratings on a numerical scale (Wegner & Wheatley, 1999; Aarts et al., 2005). This method is used in all three experiments in the later chapters.

The sense of agency can also be measured implicitly by assessing the degree of distortion in a person's subjective experience of time. This metric comes from the "intentional binding" phenomenon (Moore & Obhi, 2012), where a person will perceive an involuntary action as happening earlier than it actually did (conversely, an intentional action is perceived as happening later), while an unintended outcome is perceived as occurring later than its actual time (conversely, an intended outcome is perceived as happening sooner). A standard paradigm for measuring the intentional binding effect is using the Libet clock as shown in Figure 2.2 (Libet et al., 1983; Haggard, Clark, & Kalogeras, 2002), which is adopted in Experiment 2.

3.2 Perceived rhythmic entrainment

In a task-oriented group, individual members need to refer to other members' actions to achieve joint anticipatory control (Knoblich & Jordan, 2003). Similarly, in a music ensemble, co-performers will adjust their rhythmic behaviours when they notice others' deviations from their temporal expectations, in order to achieve a perceived coherent performance (Keller et al., 2007; Pecenka & Keller, 2011; Nowicki et al., 2013). Such action or temporal co-ordination is the core of an entrainment process. Because a more rhythmic pattern is more predictable thanks to humans' ability in forming temporal expectation (Nobre et al., 2007), adaptation during entrainment should require fewer cognitive resources. Hence I extend the findings above in the context of mixed-initiative interaction and propose the following hypothesis:
$\boldsymbol{H_{MII} - 2}$: Predictable rhythm in mixed-initiative interaction is more likely to induce

the user's entrainment behaviours, while unpredictable irregular timing is less likely to induce their entrainment behaviours.

Research in mutual adaptive tapping uses cross-correlation and auto-correlation coefficients to quantify and measure entrainment effects (Nowicki et al., 2013). Cross-correlation measures the "similarity of two interacting series as a function of the displacement of one relative to the other" (Boker, Rotondo, Xu, & King, 2002). It ranges between 0 and 1, and a larger positive value indicates a stronger temporal similarity between the two series. Assuming the two series have global stability, their cross-correlation coefficients can be calculated on the whole interval series. When the series are not stationary and only local stability can be assumed, which is often the case for the data produced in psychological experiments, their *windowed* cross-correlation coefficients should be calculated (Boker et al., 2002), and the window size (i.e. the number of observations within a window), window increment (i.e. the number of observations between adjacent windows, or the time lapse between one window movement), and the lag increment (i.e. the interval of time between the two windows truncated from the two series of interest) are determined by specific experiment designs. Another measure is the auto-correlation of a series. It is also called serial correlation. This is the correlation of a series with itself at different time points. Hence, the auto-correlation coefficient represents the "similarity between observations" of a signal itself. Previous research uses the joint lag 1 auto-correlation of one series of intervals. A positive value ($0 \sim 1$) suggests a greater tendency for temporal assimilation, whereas a negative value ($-1 \sim 0$) indicates a tendency for compensation (Nowicki et al., 2013).

3.3 Perceived level of stress

Studies in social psychology have shown that rhythmic entrainment can provide a basis for mutual trust and predictability, resulting in better anticipatory control (Keller et al., 2007; Pecenka & Keller, 2011). In addition, entrainment can induce a sense of intersubjectivity (i.e. a sense of "being together") (Schegloff, 1992; Gill, 2012) and establish mutual rapport and affiliation (Hove & Risen, 2009; Miles et al., 2009) between the interactants, thus facilitating interpersonal communication or joint problem solving activities, as reviewed earlier in **Section 2.3.3**. Furthermore, interpersonal

co-ordination on a motor and/or cognitive level can produce a sense of empathy (Spiro et al., 2013), smoothness (Gallotti et al., 2017), relaxation (Richardson et al., 2005; Clayton et al., 2005) and enjoyment (Repp & Su, 2013; Hawkins et al., 2013; Keller et al., 2014). In mixed-initiative interaction, this may result in a reduced sense of stress and mental effort. Therefore the next pair of hypotheses are:
$\boldsymbol{H_{MII} - 3.1}$: Predictable rhythm in mixed-initiative interaction can reduce the user's perceived effort, whereas unpredictable irregular timing can increase the their perceived effort.
$\boldsymbol{H_{MII} - 3.2}$: Predictable rhythm in mixed-initiative interaction can reduce the user's perceived level of stress, whereas unpredictable irregular timing can increase their perceived level of stress.

The Task Load Index (TLX) ratings system developed by the National Aeronautics and Space Administration of the United States (NASA) is a standardised work load and stress instrument (Hart & Staveland, 1988). The sub-scales include: task participants' mental demand, physical demand, temporal demand, perceived success/failure in task performance, perceived amount of effort devoted in the task, and perceived frustration during the task. Each sub-scale corresponds to a 7-point scale with 21 gradations.

While an overall TLX workload score is often calculated by adding up the weighted score for each sub-scale, it is common practice to look into individual TLX sub-scales during the analysis in order to answer more focused questions and to obtain more detailed insights. Previous studies in cognitive ergonomics, for example, on people's sensory and cognitive vigilance with task stimuli displayed on a digital monitor (Deaton & Parasuraman, 1993), or on how people's physical exertion levels can be affected by mental demands (Mehta & Agnew, 2011), have reported and compared participants' ratings on each individual sub-scale. In affective computing studies, for instance, when validating whether or not heart rate could be used as a physiological indicator of the user's mental state, the TLX "mental demand" sub-scale was singled out during the analysis (Rowe, Sibert, & Irwin, 1998) in order to answer specific research questions. When evaluating simulations for medical operations, researchers focused on participants' ratings on the TLX "mental demand" and "physical demand" sub-scales to investigate possible causes of human errors (Yurko, Scerbo, Prabhu, Acker, & Stefanidis, 2010). Likewise, all three experiments reported in this dissertation have adopted the six TLX sub-scales, and the ratings on each sub-scale were analysed individually.

3.4 Task performance

According to the review in **Section 2.2.3**, our brain is able to extract temporal patterns for a series of random external stimuli or events and form temporal expectations (Barnes & Asselman, 1991; Barnes et al., 2005; Shuler & Bear, 2006; Nobre et al., 2007; Lakatos et al., 2008). Such expectation allows attentional biasing and improves both the efficiency and the quality of sensory processing (Fujioka et al., 2009; Rohenkohl et al., 2012; Arnal & Giraud, 2012). Hence, it is easier for a person to predict and respond to random stimuli that occur regularly than those that occur at irregular times. As a result, when the sequence of stimuli has more than one dimension of uncertainty, such as its semantic content, spatial distribution and temporal attributes, a predictable rhythm can greatly simplify the temporal dimension of the target sequence. Therefore, predictable timing may allow people to devote more cognitive resources to handling the information carried by other dimensions (Rohenkohl et al., 2012). This effect has been further supported by recent studies, confirming that the temporal periodicity of random stimuli can not only improve the accuracy of complex decision making, but also allow people to make decisions faster based on less information without sacrificing accuracy (Greatrex, 2018). Hence the last pair of hypotheses are formulated as follows:
$\boldsymbol{H_{MII} - 4.1}$: Predictable rhythm in mixed-initiative interaction can help the user achieve better task performance, whereas unpredictable irregular timing can impair their task performance.
$\boldsymbol{H_{MII} - 4.2}$: Predictable rhythm in mixed-initiative interaction can make the user feel more confident in their own performance, whereas unpredictable irregular timing can impair their confidence in their own performance.

Working memory is one of the higher-order cognitive constructs (Schmiedek, Lövdén, & Lindenberger, 2014). In experimental psychology and cognitive neuroscience, N-back task paradigm is usually used as a valid indicator of working memory (Schmiedek et al., 2014). In an N-back task, participants are required to “monitor the identity or location of a series of verbal or nonverbal stimuli and indicate when the currently presented stimulus is the same as the one presented n trials previously” (Owen, McMillan, Laird, & Bullmore, 2005). This paradigm is applied in Experiment 1, which is reported in **Chapter 4**. The number of correct recalls of the shape and location of random stimuli were recorded and compared. In all experiments reported in this dissertation, participants were also asked to rate how confident they were, and how successful they perceived their performance to be on the TLX sub-scales (Hart & Staveland, 1988).

CHAPTER 4

Perceived agency and the timing of visual targets - Experiment 1

In the existing body of HCI literature, as reviewed in **Chapter 2**, very little work has been done on the effect of timing on the user's sense of agency, and there is no standard empirical paradigm of manipulating timing in mixed-initiative interaction research. Therefore, the two motivations of Experiment 1 of this dissertation are, first and foremost, to test the hypotheses proposed in **Chapter 3**, and secondly, to find a set of effective settings when manipulating the temporal structure of mixed-initiative interaction in controlled experiments.

This experiment adapted a simple type of stimulus-response paradigm that is widely used in ergonomics and cognitive psychology studies, in which sequences of user-initiated actions are conventionally followed by visual prompts initiated by the system (Simon & Wolf, 1963; Shanks et al., 1989; Kornblum, Hasbroucq, & Osman, 1990; Worringham & Beringer, 1998; Rohenkohl et al., 2012). The temporal aspects of the system-initiated events in this experiment were manipulated in a controlled manner as a first step, in order to preclude potential confounding factors that might be introduced by more realistic HCI task scenarios.

4.1 Method

This experiment was designed to investigate how different timing patterns of the presentation of visual stimuli can influence the user's sense of control ($\boldsymbol{H_{MII} - 1}$), entrainment behaviours ($\boldsymbol{H_{MII} - 2}$), perceived effort and stress level ($\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$) and their task performance ($\boldsymbol{H_{MII} - 4.1}$ and $\boldsymbol{H_{MII} - 4.2}$). Therefore, the experiments tasks should have the following characteristics:

1. The tasks should be repetitive or have repetitive steps. This will allow different temporal structures (i.e. rhythmic, random, entrained) to be imposed on the tasks or the steps.

2. The tasks should have a "turn-taking" dynamics. There should be a mix of user-initiated actions and system-initiated actions in the tasks or the steps in order to emulate realistic mixed-initiative interaction.

3. The tasks should require a reasonable amount of cognitive resources such as working memory. This is because $\boldsymbol{H_{MII} - 1}$, $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 4.1}$ and $\boldsymbol{H_{MII} - 4.2}$ were proposed based on the theories that "an occupied mind feels less control" (Hon et al., 2013) and "predictable rhythm can spare cognitive resources" (Lakatos et al., 2008; Rohenkohl et al., 2012), as cited in **Sections 3.1** and **3.4**, hence the cognitive load of the tasks should not be too high or too low, so that participants do not feel too occupied to feel any control, or too idle to have differentiating performances under different temporal structures.

4. The tasks should exert a reasonable amount of pressure on participants, so that participants' ratings for their stress level under different temporal structures will not be too high or too low to be compared when testing $\boldsymbol{H_{MII} - 3.2}$.

5. The visual stimuli should be clear and simple, without spatial or semantic ambiguity. This is to minimise the systematic errors caused by unpredictable and uncontrollable factors, particularly when the timestamp of participants' actions is a crucial measurement of their entrainment behaviours ($\boldsymbol{H_{MII} - 2}$): any confusions or hesitations during the tasks may impair the quality of the timestamps.

6. Participants' task performance should be measurable, so that performance data can be obtained to test $\boldsymbol{H_{MII} - 4.1}$.

Based on the characteristics above, **Section 4.1.1** describes the design of the tasks in Experiment 1.

4.1.1 Task design and procedures

Before starting an experiment session, all participants agreed to sign an informed consent form (Appendix A.1). Each experiment session consisted of a practice stage and a formal stage. The sample tasks in the practice stage were designed and presented in exactly the same way as the formal tasks, in order to give participants an opportunity to practise the actions that would be involved in the formal tasks and become familiar with the interaction. In both stages, participants were asked to do five types of task. Each task required them to click on multiple simple geometric targets on a computer screen using a mouse. In the first kind of task (Task 0), a prompt icon (a white cross inside a blackened square-shape area with Gaussian blur effect, sized 80×80 pixels, displayed as about 22×22 mm^2) as shown in Figure 4.1 would appear at a certain location on the screen, and participants were asked to click on the icon. Once the icon was clicked on, it would change its location, and participants needed to follow it and click on it again. The icon was programmed to appear only at one of the four fixed locations at a time. The locations were the four corners of a 330×330 pixels (displayed as about 90×90 mm^2) pre-defined blank square area in the centre of the screen. The starting point was the top-left corner. Upon each click, the icon disappeared and re-appeared immediately in the adjacent corner of that blank area in a clockwise direction. Participants were asked to click on those prompts at a rate that they were comfortable with for thirty rounds without rushing. Task 0 served two purposes in this experiment. The first purpose was that it primed participants to attend and react to the prompts or targets that would also appear on those four locations in a clockwise direction in the other four kinds of task. The second was that the average length of each participant's clicking intervals in Task 0 was then used as a default natural rhythm customised for each of them in later tasks.

The other four kinds of task consisted of the same number of rounds of interaction, and each round had a *Prompt* phase, a *Target* phase, and a *Recall* phase, hence meeting

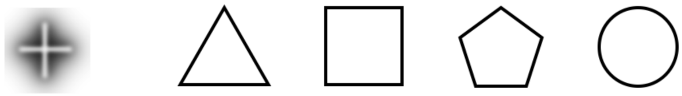


Figure 4.1: The prompt icon (a Gaussian-blurred cross) and target shapes (four simple geometric shapes) used in Experiment 1

the 1st and 2nd characteristics. Every round adopted the N-back paradigm ($N = 4$, meeting the 3rd and 4th characteristics) introduced in **Chapter 3**. The phases were designed as follows:

1. In the *Prompt* phase, a blurred cross (identical to the ones in Task 0) would appear in each corner of the blank square area sized 330×330 pixels (about 90×90 mm^2) in a clockwise sequence. In all kinds of task, the blurred cross would appear just four times during the *Prompt* phase of any given round, as shown in Figure 4.4. In Task 1 and Task 2, starting in the top-left corner of the blank square area, the blurred cross would appear then disappear in a corner, and then instantly re-appear in the next corner on its own, and participants did not need to click on it. In Task 3 and Task 4, the blurred cross would also first appear in the top-left corner of the blank square area, but it would not disappear at its current location and re-appear in the next corner until participants clicked on it. Hence, the intervals were determined by participants' clicking actions.

2. Next, in the *Target* phase, one of four random simple geometric targets (a triangle, a square, a pentagon, or a circle as shown in Figure 4.1, hence meeting the 5th characteristic) sized 80×80 pixels (about 22×22 mm^2) would appear in each of the four corners where the blurred cross had just appeared. Similarly, the random target would start in the top-left corner and only appear four times in a clockwise direction during the target phase of any given round, as illustrated in Figure 4.4. The geometric target was programmed to be random when it re-appeared, so the shape may or may not be the same at different locations. In Task 1, Task 2 and Task 3, the target would appear, disappear and re-reappear on its own following system-determined temporal intervals, and participants were only required to observe the target closely. In Task 4, participants were asked to not only observe the target but also manually click it so that the target would

disappear and re-appear in the next corner.

3. Finally, in the *Recall* phase, a horizontal bar (sized 80×320 pixels and displayed as about about 22×90 mm^2, as shown in Figure 4.2) would appear in each of the four corners of the same 330×330-pixel (about 90×90 mm^2) area in a clockwise manner as for the blurred cross and the random targets. The horizontal bar consisted of four clickable buttons, and from left to right the foreground of each button was a triangle, a square, a pentagon, or a circle, as illustrated in Figure 4.4. In the *Recall* phase of a given round in all of the tasks, participants were required to recall which geometric shape appeared during the *Target* phase in each of the four corners, and click the button the foreground of which corresponded to the recalled shape, hence meeting the 4^{th} and 6^{th} characteristics.

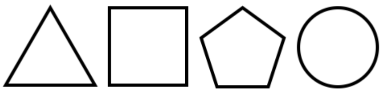
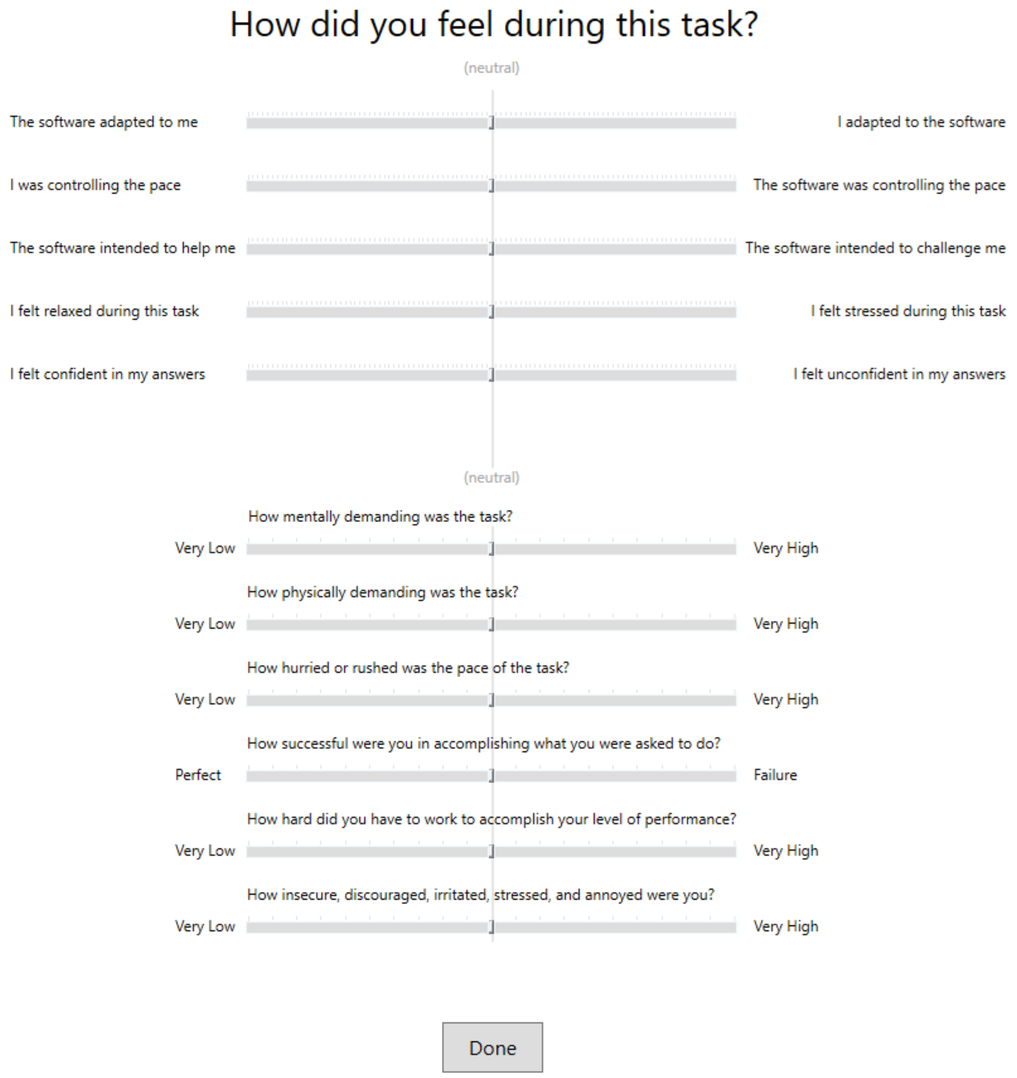


Figure 4.2: The horizontal choice bar used in the *Recall* phase in Experiment 1

In the practice stage, each of the four tasks had three rounds. In the formal stage, each task consisted of thirty rounds. After each task, participants reported their subjective ratings for their perceived sense of control and stress during the task they just completed using two sets of slider bars on the screen, as shown in Figure 4.3. As shown in Appendix A.5.1, the sequence of the four tasks was randomised for each participant in order to mitigate learning effect.

Considering that experimental demand and prior expectation may cause participants to hold subjective biases, all participants were told that this experiment would study *"how people follow various sequences of events on a screen"*. The term "timing" or "rhythm" was not mentioned during either the recruitment message or the task briefing. The script that was used during the experiment introduction can be found in Appendix A.3.2.

After each participant completed the experiment, they were given a debrief, informing them that in addition to their task performance, this study was also interested in how different timing patterns of the presentation of visual stimuli might have affected their subjective perception of control. Each experiment session lasted for 20-30 minutes, and a small gift (valued £6∼£8) was given in appreciation of their time. This experiment was reviewed and approved by the ethics committee of the Computer Laboratory, University of Cambridge.



How did you feel during this task?

(neutral)

The software adapted to me	I adapted to the software
I was controlling the pace	The software was controlling the pace
The software intended to help me	The software intended to challenge me
I felt relaxed during this task	I felt stressed during this task
I felt confident in my answers	I felt unconfident in my answers

(neutral)

How mentally demanding was the task?
Very Low — Very High

How physically demanding was the task?
Very Low — Very High

How hurried or rushed was the pace of the task?
Very Low — Very High

How successful were you in accomplishing what you were asked to do?
Perfect — Failure

How hard did you have to work to accomplish your level of performance?
Very Low — Very High

How insecure, discouraged, irritated, stressed, and annoyed were you?
Very Low — Very High

Figure 4.3: The slider bars used to collect participants' subjective ratings after each task in Experiment 1

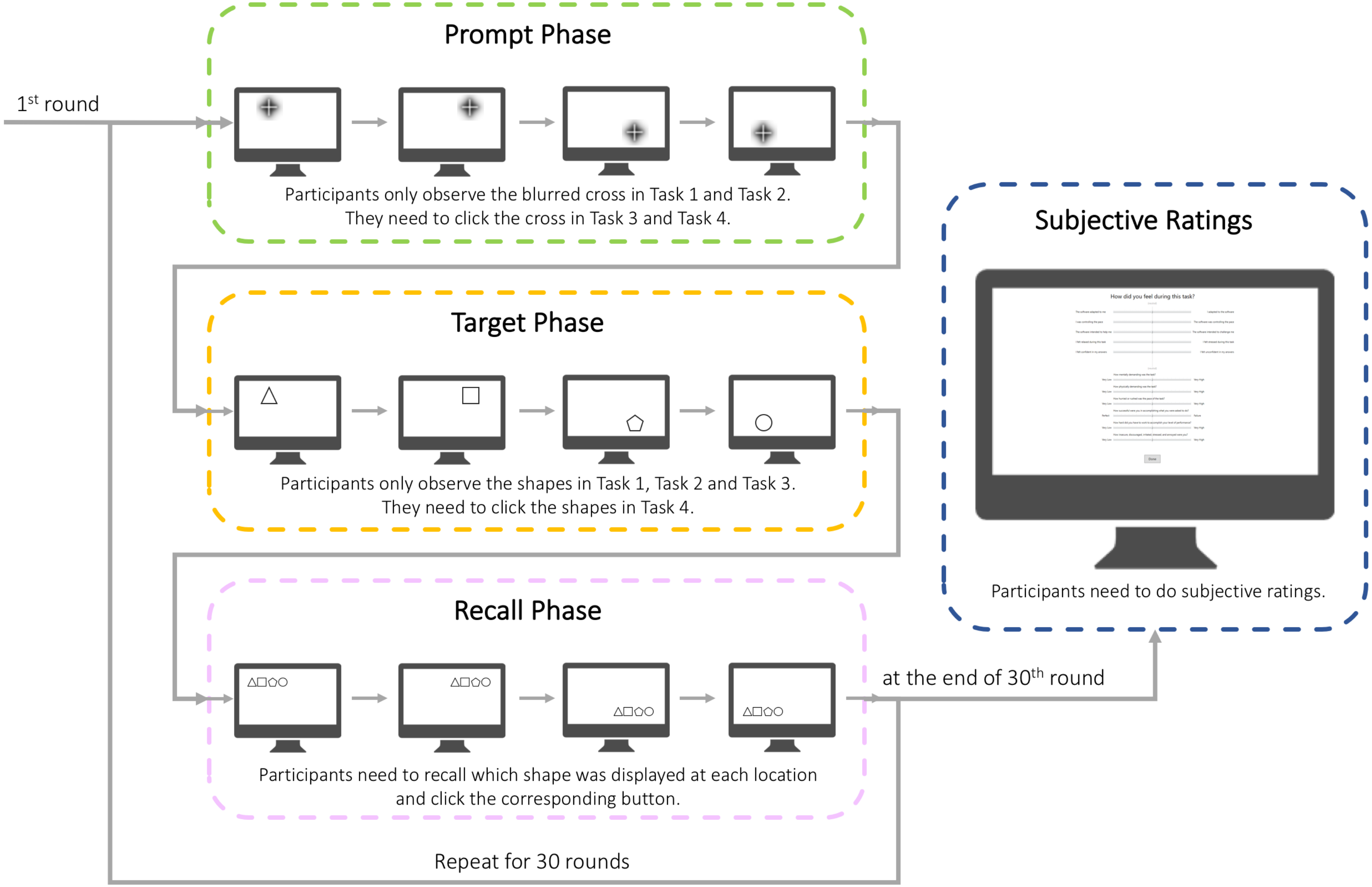

Prompt Phase
1st round
Participants only observe the blurred cross in Task 1 and Task 2.
They need to click the cross in Task 3 and Task 4.
Target Phase
Participants only observe the shapes in Task 1, Task 2 and Task 3.
They need to click the shapes in Task 4.
Recall Phase
△□⬠○
Participants need to recall which shape was displayed at each location and click the corresponding button.
Repeat for 30 rounds
at the end of 30th round
Subjective Ratings
How did you feel during this task?
Done
Participants need to do subjective ratings.

Figure 4.4: The illustration of the design of formal tasks in Experiment 1

4.1.2 Independent variable and manipulation

This experiment used a within-subject design. The only independent variable in this experiment was the imposition of either predictable rhythmic intervals or randomised arrhythmic intervals throughout an experimental task. The four conditions are shown in Table 4.1. Each condition corresponded to a different method of initiating an action and setting the pace. Taking the rows of the table from ***Sys-ii*** down to ***Usr-r***, it can be seen that the rhythmic character of the initiative-taking during the interaction becomes more and more predictable and under the user's control.

Independent variable	*Description of treatment*	*Abbreviation*
Irregular intervals	***Sys***tem takes the initiative at ***i***rregular ***i***ntervals	***Sys-ii***
Predictable rhythm	***Sys***tem takes the initiative in a ***p***redictable ***r***hythm	***Sys-pr***
	Use***r*** takes the initiative, ***Sys***tem aligns	***Usr-Sys***
	Use***r*** takes the initiative in their own ***r***hythm	***Usr-r***

Table 4.1: The independent variable and its settings in Experiment 1

The rationale for choosing the four temporal structures above is:

1. ***Sys-ii***: In cognitive neuropsychology studies, a series of visual stimuli are often presented by computer devices at random times during the experiment tasks, and participants need to respond and recall the stimuli (Lakatos et al., 2008; Rohenkohl et al., 2012). Therefore in Task 1 of this experiment, a series of random visual prompts displayed by the system at random times (***Sys-ii***) represent a worst-case scenario in mixed-initiative interaction, in which the user has absolutely no control over the timing of system-initiated actions, and no control over the content of the stimuli to be displayed either.

2. ***Sys-pr***: Presenting visual stimuli in a rhythmic manner is often used to contrast with the random timings above in the same studies (Lakatos et al., 2008; Rohenkohl et al., 2012). In Task 2 of this experiment, a series of random visual prompts displayed by the system rhythmically (***Sys-pr***) represent an improved

scenario in mixed-initiative interaction, in which the user has no control over the timing or the content of system-initiated actions but at least the timing is regular and predictable.

3. ***Usr-Sys***: In existing studies in social psychology, entrainment can be found in a conversation where the listener's utterance would fall on the beats aligning precisely to the speaker's rhythm, and the speech rhythm can be carried over across the turn-taking boundary between speakers (Couper-Kuhlen, 1993; Auer et al., 1999; Hawkins et al., 2013). In Task 3 of this experiment, a series of random visual prompts were displayed by the system in a manner that strictly mirrored the pace of the user's actions (***Usr-Sys***). This represents a rigid entrainment scenario in mixed-initiative interaction, in which the user has some control over the timing but not the content of system-initiated actions. Participants were not informed of this pace-mirroring dynamics either.

4. ***Usr-r***: In human-computer interaction and cognitive psychology studies (Grossman & Balakrishnan, 2005; Schmiedek et al., 2014), it is common to ask participants to interact with visual targets on a screen (e.g. mouse clicks, key presses, finger taps etc.). In Task 4 of this experiment, a series of random visual prompts were only displayed by the system when the user triggered them manually in their own time (***Usr-r***). This represents another extreme case scenario opposite to ***Sys-ii*** in mixed-initiative interaction, in which the user has the most control over the timing and can deal with system's action in their own pace, though not the content of system-initiated actions.

The design of each type of task is illustrated in Figure 4.4. As introduced in **Section 4.1.1**, each experiment session always started with a preparation task (Task 0), in which participants needed to click on the blurred prompt cross that appeared in order at four locations on the screen for thirty rounds. All of their between-click intervals were recorded, and the average length of those intervals (denoted as M_i for Participant i) was later used to set the rhythm for Task 1 and Task 2.

In both Task 1 (***Sys-ii***) and Task 2 (***Sys-pr***), the screen first displayed the blurred prompt cross in sequence at four locations on the screen, then four randomised geometric shapes in the same order at the same four locations. In the ***Sys-pr*** condition (Task 2), the time interval between the presentation of every two successive visual stimuli had a fixed length customised for each participant, which was M_i (i stands for

Participant i) observed in Task 0, as shown in Figure 4.5. In the ***Sys-ii*** condition (Task 1), the length of the time intervals between stimuli was randomised. The series of random intervals was generated in MathWorks MATLAB R2015b [1], with the mean value set as M_i for Participant i and the interval length ranged between $\frac{1}{2}M_i$ and $\frac{3}{2}M_i$ following continuous uniform distribution, see the intervals labelled with $M_i(RAN)$ in Figure 4.5. Every two adjacent intervals had at least a 25-millisecond difference in length, in order to reach the threshold that people could notice the temporal variation (Goebl & Parncutt, 2002; Goebl & Palmer, 2009). All random intervals in Task 1 were generated right after Task 0 then preloaded into the experimental system before Task 1 started. A sample of random intervals that were used in this experiment are presented in Appendix A.4.1.

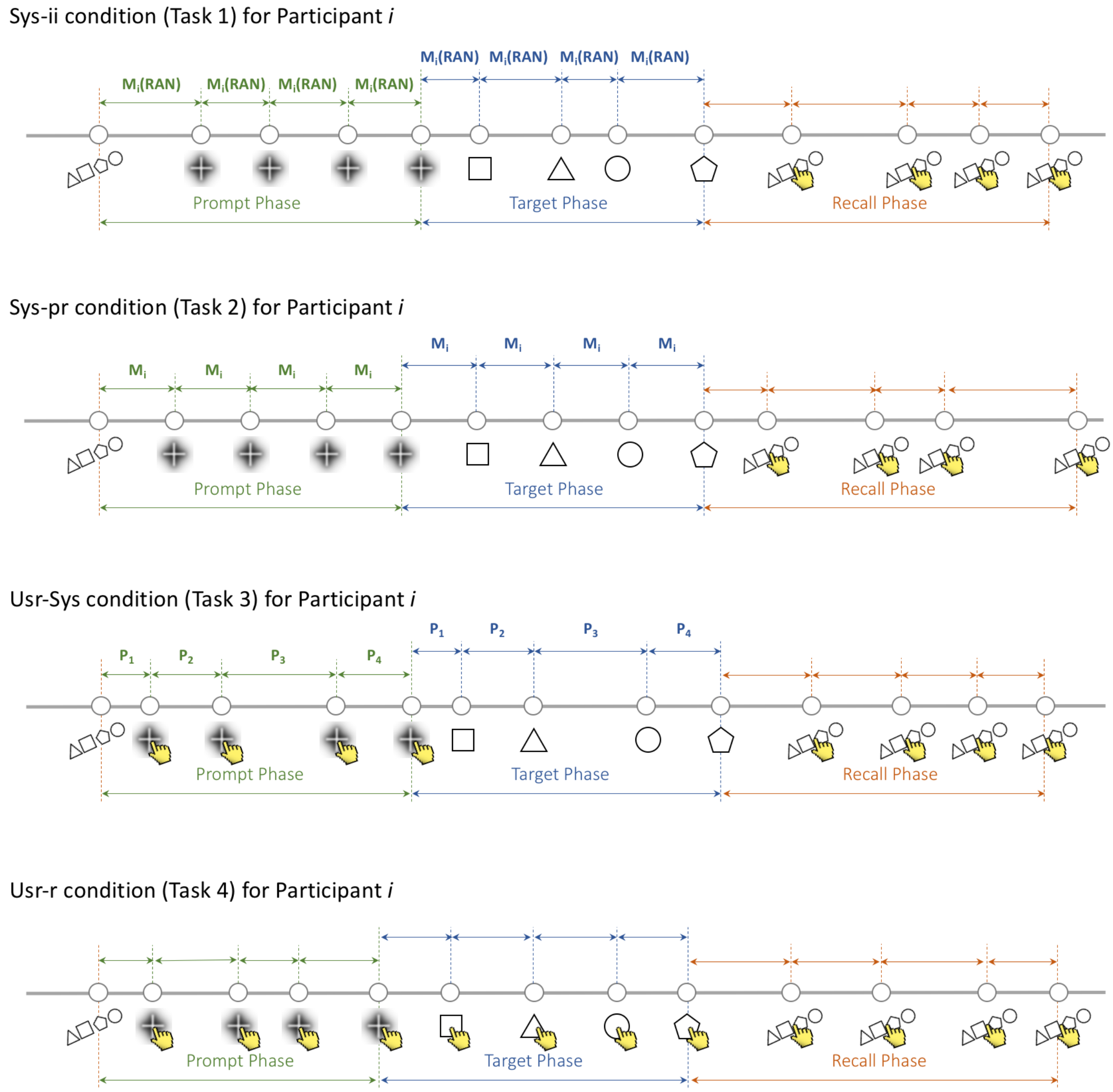
In the ***Usr-Sys*** condition (Task 3), participants needed to first click on the prompt cross, then waited and observed the display of four randomised shapes without clicking. The time intervals between presentation of the shapes mirrored exactly the intervals of participants' own clicking on the prompt cross, as can be seen from Figure 4.5. In the ***Usr-r*** condition (Task 4), participants were asked to click on the prompt cross at the same four locations, then click on four randomised shapes, all at their own preferred rate.

4.1.3 Dependent variables and measures

The intervals of interaction events such as stimulus presentation and participants' mouse clicks were recorded in real time by the experimental system. As shown in Figure 4.5, there were twelve intervals in each round, falling into the three phases introduced in **Section 4.1.1**. For the k^{th} round in a task (r_k), the first four were the intervals before every *Prompt* cross was presented: $I(r_k, P_1)$, $I(r_k, P_2)$, $I(r_k, P_3)$, $I(r_k, P_4)$. The next four were the intervals before a random geometric *Target* presentation: $I(r_k, T_1)$, $I(r_k, T_2)$, $I(r_k, T_3)$, $I(r_k, T_4)$. The final four were the intervals between *Recalls*: $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$.

Based on these intervals, three dependent variables introduced in **Chapter 3**

[1] The code in this experiment was adapted from an open-source MATLAB function used in Experiment 5 and 6 of Greatrex (2018)'s doctoral research to generate sequences of random and aperiodic intervals. The original code can be found via this link: *https://github.com/dcgreatrex-phd/experiment_5/blob/master/private/computeIOIarray.m.*



Sys-ii condition (Task 1) for Participant *i*

Sys-pr condition (Task 2) for Participant *i*

Usr-Sys condition (Task 3) for Participant *i*

Usr-r condition (Task 4) for Participant *i*

Figure 4.5: The illustration of the temporal structure within one round in each of the four treatments in Experiment 1

were calculated to describe the rhythmic entrainment over time: the auto-correlation coefficient of participants' *Recall* intervals during the *Recall* phase of two successive rounds; the cross-correlation coefficient between the *Prompt* intervals and the *Recall* intervals within one round; and the cross-correlation coefficient between *Target* intervals and *Recall* intervals within one round.

For the auto-correlation coefficient, the calculation procedures are as follows. The *Recall* intervals of two successive rounds were considered here, for instance, the k^{th} and $k+1^{th}$ round, hence the lag was 1 round, and the *Recall* intervals involved in the calculation were $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$ and $I(r_{k+1}, R_1)$, $I(r_{k+1}, R_2)$, $I(r_{k+1}, R_3)$, $I(r_{k+1}, R_4)$, as shown in Figure 4.6. Therefore the auto-correlation between the *Recall* intervals in the k^{th} and $k+1^{th}$ round could be calculated using the following formula:

$$AC_{r_k} = \frac{1}{4}\sum_{j=1}^{4} \frac{\Big(I(r_k, R_j) - \overline{I(r_k, R)}\Big) \times \Big(I(r_{k+1}, R_j) - \overline{I(r_{k+1}, R)}\Big)}{std\big(I(r_k, R)\big) \times std\big(I(r_{k+1}, R)\big)},$$

where $\overline{I(r_k, R)}$ and $std\big(I(r_k, R)\big)$ are the mean and the standard deviation of $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$, and $\overline{I(r_{k+1}, R)}$ and $std\big(I(r_{k+1}, R)\big)$ are the mean and the standard deviation of $I(r_{k+1}, R_1)$, $I(r_{k+1}, R_2)$, $I(r_{k+1}, R_3)$, $I(r_{k+1}, R_4)$. Every task had thirty rounds, hence twenty-nine coefficients AC_{r_k} could be computed. Considering that participants might not have been ready in the first couple of rounds, the first auto-correlation coefficient AC_{r_1} was removed from analysis. The mean value of the rest twenty-eight auto-correlation coefficients were the average auto-correlation of one task. Each participant would have four mean auto-correlation coefficients calculated in this way, one for each task.

The cross-correlation coefficients between the *Prompt* intervals and the *Recall* intervals within one round were calculated in a similar manner. For instance, in the k^{th} round as shown in Figure 4.7, the *Prompt* intervals were $I(r_k, P_1)$, $I(r_k, P_2)$, $I(r_k, P_3)$, $I(r_k, P_4)$, and the *Recall* intervals were $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$, hence the cross-correlation between those two series could be obtained using this formula:

$$CC(P\&R)_{r_k} = \frac{1}{4}\sum_{j=1}^{4} \frac{\Big(I(r_k, P_j) - \overline{I(r_k, P)}\Big) \times \Big(I(r_k, R_j) - \overline{I(r_k, R)}\Big)}{std\big(I(r_k, P)\big) \times std\big(I(r_k, R)\big)},$$

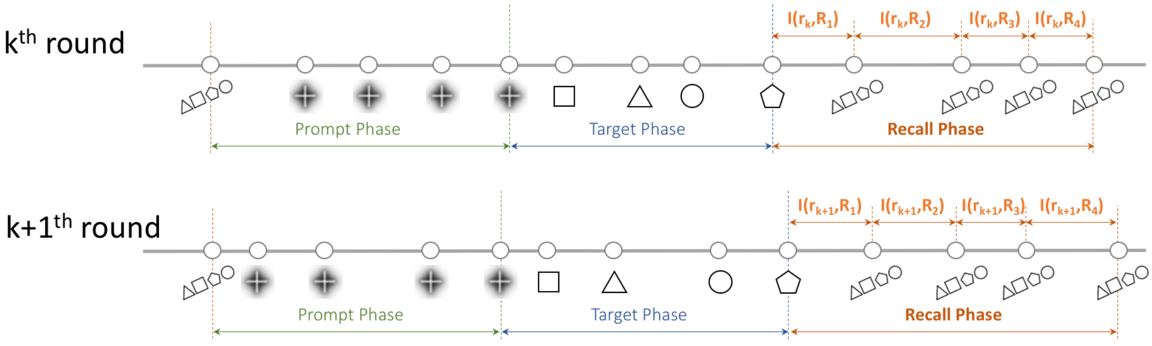


Figure 4.6: *Recall* intervals that were used for auto-correlation calculation in Experiment 1

where $\overline{I(r_k, P)}$ and $std\big(I(r_k, P)\big)$ are the mean and the standard deviation of $I(r_k, P_1)$, $I(r_k, P_2)$, $I(r_k, P_3)$, $I(r_k, P_4)$, and $\overline{I(r_k, R)}$ and $std\big(I(r_k, R)\big)$ are the mean and the standard deviation of $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$. Every task had thirty rounds, hence thirty coefficients $CC(P\&R)_{r_k}$ could be computed. Again the cross-correlation coefficients in the first two tasks were removed during analysis, and the mean value of the remaining twenty-eight rounds were the average cross-correlation of one task. Each participant would have four mean cross-correlation coefficients between *Prompt* and *Recall* intervals calculated in this way, one for each task.

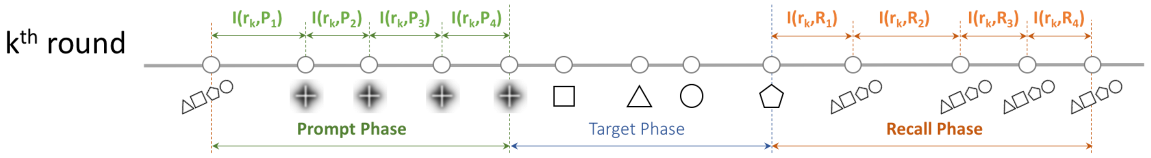


Figure 4.7: *Prompt* and *Recall* intervals that were used for cross-correlation calculation in Experiment 1

Similarly, the cross-correlation coefficients between *Target* intervals and *Recall* intervals within one round could be calculated using the same method:

$$CC(T\&R)_{r_k} = \frac{1}{4}\sum_{j=1}^{4}\frac{\Big(I(r_k, T_j) - \overline{I(r_k, T)}\Big) \times \Big(I(r_k, R_j) - \overline{I(r_k, R)}\Big)}{std\big(I(r_k, T)\big) \times std\big(I(r_k, R)\big)},$$

where $\overline{I(r_k, T)}$ and $std\big(I(r_k, T)\big)$ are the mean and the standard deviation of the four

Target intervals $I(r_k, T_1)$, $I(r_k, T_2)$, $I(r_k, T_3)$, $I(r_k, T_4)$ in the k^{th} round, and $\overline{I(r_k, R)}$ and $std\big(I(r_k, R)\big)$ are the mean and the standard deviation of the *Recall* intervals $I(r_k, R_1)$, $I(r_k, R_2)$, $I(r_k, R_3)$, $I(r_k, R_4)$, as illustrated in Figure 4.8. Once again only the last twenty-eight rounds were considered during analysis.

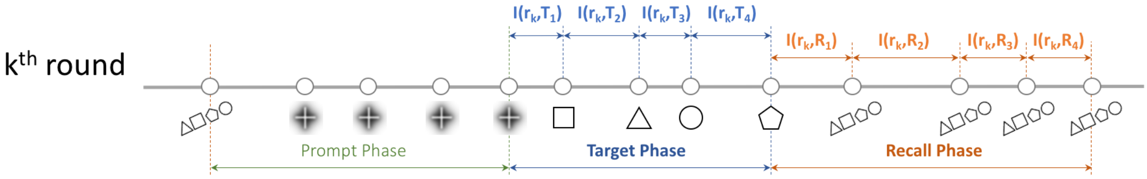


Figure 4.8: *Target* and *Recall* intervals that were used for cross-correlation calculation in Experiment 1

Participants' choices of shape and location during the *Recall* stage were recorded by the experimental system, and the number of accurate recalls in each task was counted as a dependent variable.

After each task, in order to collect subjective ratings, participants were presented with two sets of slider bars, all initialised to the mid position, with paired opposite statements at each end. As introduced in **Section 3.3**, the NASA-TLX sub-scales were adopted to assess participants' mental demand, physical demand, temporal demand, performance, effort and frustration (Hart & Staveland, 1988; Deaton & Parasuraman, 1993; Rowe et al., 1998). The arrangement of each sub-scale is shown in Figure 4.3. On the mental demand, physical demand, temporal demand, effort and frustration sub-scales, the "very low" label is on the left hand side, "very high" is on the right hand side. On the performance sub-scale, the "perfect" label is on the left, the "failure" label is on the right.

Participants were also asked to give a numeric rating on the following items:

1. *"The software adapted to me"* vs. *"I adapted to the software"*
2. *"I was controlling the pace"* vs. *"The software was controlling the pace"*
3. *"The software intended to help me"* vs. *"The software intended to challenge me"*
4. *"I felt relaxed during this task"* vs. *"I felt stressed during this task"*
5. *"I felt confident in my answers"* vs. *"I felt unconfident in my answers"*

As commonly done in previous studies (Deaton & Parasuraman, 1993; Rowe et al., 1998; Yurko et al., 2010; Mehta & Agnew, 2011), participants' ratings on each individual sub-scale were contrasted among the four task conditions in order to test different hypotheses.

4.1.4 Participants

Twenty-two participants (age $M = 27.6$, $\sigma = 4.61$; 8 females) were recruited for this experiment. Their background information was collected using the form in Appendix A.2. Three participants are left-handed. In each experimental session, participants were allowed to use the computer mouse on their preferred side. Five participants have normal vision and the other seventeen have corrected-to-normal vision. One participant has "light red/green" colour blindness.

Participants' education level ranged from PhD to high school, the breakdown is as follows: seven participants have obtained a PhD degree, fourteen with a Masters degree, and one participant graduated from high school and did not pursue a higher degree. Fifteen participants were studying STEM subjects (e.g. computer science, engineering, biology, radiology, psychology), four were studying humanities subjects (e.g. law, classics, history, literature), one in education, one in international business administration, and one did not specify.

As shown in Table 4.2, eleven participants reported that they had received music training such as instrument playing, singing and composing for 2∼10 years, eight participants had undertaken training in dancing or gymnastics ranging between 3 months and 12 years, and nine participants reported that they had 6∼20 years of video game playing experience.

4.1.5 Apparatus

All experiment sessions were carried out in Office SS08 of the Computer Laboratory, University of Cambridge. All participants used the same desktop computer (System: Windows 10 Pro, 64-bit; CPU: 2.80GHz; RAM: 8.00GB) with the same computer monitor (Samsung, SM2443BW 24-inch Black Widescreen LCD, 1920×1200) and the

Table 4.2: Participants' background information in Experiment 1

No.	*Age*	*Gender*	*Handedness*	*Vision*	*Colour-blindness*	*Education*	*Music training*	*Dance/ gymnastics*	*Game playing*
1	35	Female	Right	Normal	-	PhD in computer science	-	-	-
2	22	Male	Right	Corrected to normal	-	PhD in computer science	2yrs, 5-10hr/week Instrument	4yrs, 2hr/week Ballroom, Latin	14yrs, 20hr/week Mouse/keyboard/ playstation handle
3	26	Male	Right	Corrected to normal	-	PhD in computer science	-	-	-
4	27	Male	Right	Normal	-	Masters in molecular biomedine	-	-	-
5	25	Male	Right	Normal	-	PhD in international business administration	-	-	-
6	24	Male	Right	Corrected to normal	-	Masters in engineering	6yrs, 3hr/week Instrument	-	18yrs, 7hr/week Touch screen/mouse/keyboard
7	25	Male	Right	Corrected to normal	-	Masters in computer science	5yrs, 3hr/week Instrument	-	20yrs, 6hr/week Touch screen/mouse/keyboard
8	24	Female	Right	Normal	-	Masters in computer science	-	3mons, 2hr/week, Ballroom	10yrs, 3-4hr/week Touch screen/mouse/keyboard
9	26	Female	Right	Corrected to normal	-	PhD in psychology	10yrs, 1hr/week Instrument	1.5yrs, 2hr/week, Ballet	-
10	25	Male	Left	Corrected to normal	-	Masters in law	-	-	20yrs, 1hr/week Playstation handle/Gameboy
11	24	Female	Right	Corrected to normal	-	Masters in classics	7yrs, 5-7hr/week Instrument	4yrs, 4hr/week Gymnastics	-
12	32	Male	Right	Corrected to normal	-	Masters in radiology	-	-	6yrs, 4-6hr/week Mouse/keyboard/playstation handle
13	28	Male	Right	Normal	Red/green (light)	Masters in history and English literature	10yrs, 4hr/week Instrument	6mons, 2hr/week Walz, Tango, Cha-cha-cha	-
14	32	Male	Right	Corrected to normal	-	Masters in computer science	-	-	15yrs, 5hr/week Strategy games
15	25	Male	Right	Corrected to normal	-	Masters in computer science	5yrs, 1hr/week Singing, instrument	12yrs, 2hr/week, Radio dancing	10yrs, 3hr/week Touch screen/mouse/keyboard
16	23	Male	Right	Corrected to normal	-	Masters in computer science	-	-	15yrs, 4hr/week Mouse/keyboard/motion sensing/ handheld console
17	33	Female	Right	Corrected to normal	-	PhD in electrical-electronics engineering	-	-	-
18	38	Female	Left	Normal	-	High school	-	-	-
19	-	Male	Right	Corrected to normal	-	Masters in law	5yrs, 2hr/week Singing/composing/ instrument	5yrs, 1hr/week Ballet	-
20	36	Male	Right	Normal	-	PhD in computer science	5yrs, 10hr/week Instrument	-	-
21	27	Female	Right	Corrected to normal	-	Masters in education	8yrs, 3-4hr/week Instrument	17mons, 2-3hr/week Zumba/Rub/Ballroom/Salsa	-
22	28	Female	Right	Normal	-	Masters in stem cell biology and medicine	2yrs, 6hr/week Instrument	-	-

same optical mouse (Microsoft IntelliMouse Optical 1.1A).

The experimental system was implemented using C# as a Windows Presentation Foundation (WPF) application. The software was developed completely by the author of the dissertation, and they can be found via this link:

https://github.com/ChristineGuoYu/PhD_Experiment_1

During every experiment session, the programme ran in Visual Studio Community 2015 environment (Version 14.0.23107.0 D14REL).

4.2 Result analysis

The experiment results were analysed using the following procedures:

Step 1: In order to test a given hypothesis, an omnibus test was employed as the first step to detect whether or not the rhythm setting had caused a significant overall difference among the four experiment conditions in **Section 4.1.2**. If the data was normally distributed or approximately normally distributed, a repeated-measure one-way ANOVA was used as the omnibus test. Otherwise, the non-parametric Friedman Test was used as the omnibus test.

Step 2: If the omnibus test results confirmed that the rhythm setting did cause a significant overall effect across four conditions, a *planned* contrast analysis was applied to reveal the effects within the omnibus test. As instructed in Rosenthal, Robert, and Rosnow (1985)'s and Abdi and Williams (2010)'s guide, when conducting contrast analysis, the hypothesis under investigation was expressed (or "translated") into one or several sets of contrast weights (or "contrast coefficients"), each set was defined in the following format shown in Table 4.3,

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weight (λ)	λ_1	λ_2	λ_3	λ_4

Table 4.3: Weights in contrast analysis

where $\Sigma\lambda = 0$. Since there were four conditions in this experiment, hence three sets of orthogonal contrasts could be assigned to reflect three kinds of trend across the conditions: linear (↗ or ↘), quadratic (↗↘ or ↘↗) or cubic (↗↘↗ or ↘↗↘). For each kind of trend, the weights (λ_i) were defined in the standard table for orthogonal polynominal-based contrasts in Rosenthal, Rosnow, and Rubin (2000, p. 151)'s book and Haans (2018, p. 7)' guide.

Step 3: With the contrast weights assigned as above, the sum of squares (SS) and the mean square (MS) for the contrast could be calculated,

$$MS_{contrast} = \frac{SS_{contrast}}{df} = \frac{L^2}{n\Sigma\lambda^2},$$

in which $df = 1$, n is the number of participants in each condition, λ_i is defined in Table 4.3 based on the given hypothesis, and L is the sum of λ-weighted condition totals T_i ($L = \sum_{i=1}^{4} T_i\lambda_i$). With the error term ($SS_{error}$ and MS_{error}) calculated for the within-subjects contrast, an F test would be carried out against the hypothesised contrast:

$$F = \frac{MS_{contrast}}{MS_{error}}$$

When several sets of contrasts (e.g. k) against the same set of data were tested at the same time, the alpha level for each F test would be adjusted using the *Bonferroni* correction method ($\alpha = 0.05/k$) in order to control the probability of Type I errors.

Step 4: Following the results of the omnibus test and the contrast analysis above, incidental *post-hoc* pairwise comparisons would be carried out to reveal further insights. Similarly, when several pairwise comparisons were analysed at the same time, the *Bonferroni* correction was applied on the alpha level to reduce Type I errors.

4.2.1 Sense of control

In order to test the effectiveness of the manipulation of the independent variable and hypothesis $\boldsymbol{H_{MII} - 1}$, participants' subjective ratings for their sense of control during different tasks are analysed. The data did not pass the Shapiro-Wilk Normality Test, therefore the non-parametric Friedman Test was used to analyse the overall effect of rhythm setting across four conditions. As shown in Table 4.4, the rhythm setting did have a significant overall effect on participants' ratings for their sense of control, and the manipulation of the independent variable was effective.

Measurement	*N*	χ^2	*df*	*Sig.*
Sense of control	22	43.340	3	**.000***

Table 4.4: Omnibus test for sense of control (Friedman Test)

As hypothesised in $\boldsymbol{H_{MII}-1}$, predictable rhythm in mixed-initiative interaction will increase the user's sense of agency, while irregular timing will result in a decrease. In this experiment, as the rhythm became increasingly predictable and under participants' control across four conditions (***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***), an upward linear trend ("↗") in the ratings for sense of control should be expected if the hypothesis were true. The weights for the predicted linear trend were defined in Table 4.5.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 4.5: Weights in contrast analysis for ratings for sense of control (Hypothesis $\boldsymbol{H_{MII}-1}$: "↗" trend)

The contrast analysis was carried out in SPSS (Rosenthal et al., 1985; Haans, 2018), and the results are presented in Table 4.6. A significant linear trend (p<0.001) was found in the ratings for sense of agency as predicted. Figure 4.9 [2] demonstrates how the data followed the hypothesised trend: participants tended to report a stronger sense of control of the interaction pace as they took more initiative in rhythm setting (e.g. ***Usr-r*** condition); when the system took the initiative in both the ***Sys-ii*** and the ***Sys-pr*** conditions, participants also reported a stronger sense of control when the system did it in a predictable and rhythmic manner, compared with an unpredictable

[2] The slider bar for the rating on the sense of control had the statement *"I was controlling the pace"* on its left end, and *"The software was controlling the pace"* on its right end, hence the lower the original value of the rating, the stronger the sense of control participants were reporting, or vice versa. During the analysis, the value of the ratings were calculated inversely, hence in Figure 4.9, a higher rating value suggests that participants perceived themselves as more in control, and a lower value as less in control.

and arrhythmic manner. Therefore, hypothesis $\boldsymbol{H_{MII} - 1}$ is supported, and the manipulation of temporal structures in four task conditions is further confirmed to be effective.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↗"	Rhythm	715322.227	1	715322.227	50.805	**.000***
	Error (Rhythm)	295674.772	21	14079.751		

Table 4.6: Contrast analysis for ratings for sense of control (Hypothesis $\boldsymbol{H_{MII} - 1}$: "↗" trend)

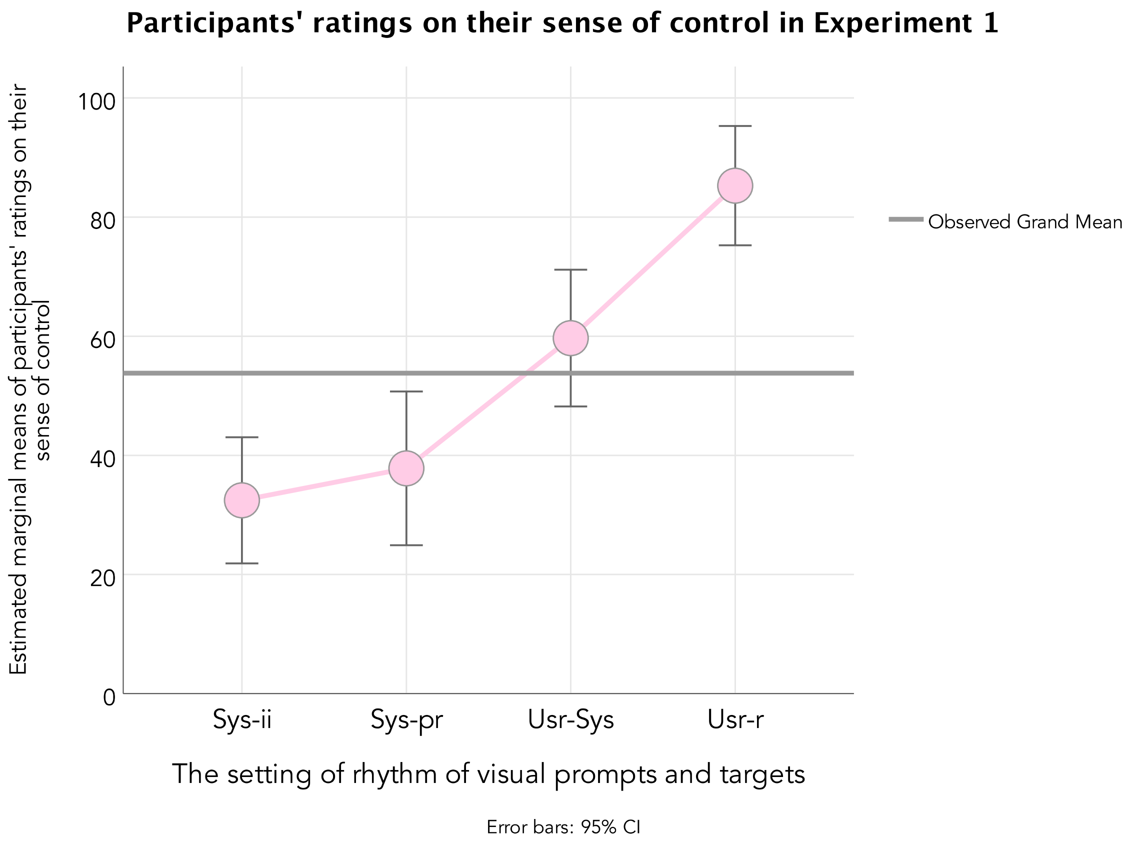


Figure 4.9: Participants' subjective ratings for their sense of control in Experiment 1

After confirming the upward trend in participants' ratings for their sense of control, pairwise comparisons were carried out with the alpha level adjusted to $0.05/6 = 0.0083$ using the *Bonferroni* correction method. Given that the rating data was not normally distributed, the Wilcoxon Signed Ranks Test was adopted. As shown in Table 4.7, four out of six pairs were significantly different. Difference was observed among the other two pairs but not to a significant level after the *Bonferroni* correction.

Pair	***Sys-ii<*** ***Sys-pr***	***Sys-ii<*** ***Usr-Sys***	***Sys-ii<*** ***Usr-r***	***Sys-pr<*** ***Usr-Sys***	***Sys-pr<*** ***Usr-r***	***Usr-Sys<*** ***Usr-r***
Z	-2.017	-3.059	-4.109	-2.433	-4.107	-3.528
Sig.	.044	**.002***	**.000***	.015	**.000***	**.000***

Table 4.7: Pairwise comparisons for ratings for sense of control (Wilcoxon Signed Ranks Test)

In summary, a significant overall effect was confirmed in the omnibus test and a significant upward linear trend was found as predicted in hypothesis $\boldsymbol{H_{MII} - 1}$. The results of pairwise tests (with *Bonferroni* correction) were not as significant, and the limitations will be discussed in **Section 4.3.2**. Therefore $\boldsymbol{H_{MII} - 1}$ is supported but should be interpreted and generalised with caution and limited confidence.

4.2.2 Perceived stress level

Participants' ratings on six TLX sub-scales were not normally distributed, hence the non-parametric Friedman Test was adopted again to test the main overall effect of rhythm setting across four conditions. Significant effect was found on the ratings for the perceived physical demand, the perceived successfulness/failure of task performance and the perceived amount of effort devoted to the tasks, as shown in Table 4.8.

Measurement	*N*	χ^2	*df*	*Sig.*
TLX mental demand	22	4.432	3	.218
TLX physical demand	22	12.277	3	**.006***
TLX temporal demand	22	5.691	3	.128
TLX success	22	13.206	3	**.004***
TLX effort	22	9.332	3	**.025***
TLX frustration	22	5.983	3	.112

Table 4.8: Omnibus test for ratings on TLX sub-scales (Friedman Test)

Due to the task design, the more initiative in rhythm setting participants took, the more clicking actions were required. Hence an upward linear trend ("↗") in the ratings for physical demand should be seen in four task conditions (***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***), and the weights for contrast analysis were assigned in Table 4.9 accordingly. The results of contrast analysis confirmed that there was a significant linear trend as expected (p=0.002), as shown in Table 4.10 and in Figure 4.10.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = -1$	$\lambda_3 = 0$	$\lambda_4 = 2$

Table 4.9: Weights in contrast analysis for ratings on TLX physical demand sub-scale (hypothesis: "↗" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↗"	Rhythm	1298.227	1	1298.227	13.153	**.002***
	Error (Rhythm)	2072.773	21	98.703		

Table 4.10: Contrast analysis for ratings on TLX physical demand sub-scale (hypothesis: "↗" trend)

Hypothesis $\boldsymbol{H_{MII}} - \mathbf{4.2}$ predicts that a predictable rhythm in mixed-initiative interaction would make the user feel more confident in their own task performance, while irregular timing would have the opposite effect. If $\boldsymbol{H_{MII}} - \mathbf{4.2}$ were true, an upward linear trend ("↗") should exist in participants' ratings for how successful they felt during the tasks. According to the original TLX questionnaire design, the sub-scale for task successfulness was labelled *"Perfect"* on its left end, and *"Failure"* on its right, hence the lower the original value of the rating, the more successful participants perceived their task performance to be, or vice versa. During the contrast analysis, the rating value on this sub-scale was calculated inversely, so that the direction of the data trend can reflect hypothesis $\boldsymbol{H_{MII}} - \mathbf{4.2}$.

Table 4.11 shows the weights assigned for each condition, and the results of contrast analysis are presented in Table 4.12. It was confirmed that there was a significant upward linear trend (p=0.018) in participants' ratings for their perceived success across four conditions (***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***) as predicted.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -2$	$\lambda_2 = -1$	$\lambda_3 = 0$	$\lambda_4 = 3$

Table 4.11: Weights in contrast analysis for ratings on TLX success sub-scale (Hypothesis $\boldsymbol{H_{MII} - 4.2}$: "↗" trend)

The rating data is illustrated in Figure 4.10, where a higher rating value indicates that participants perceived their task performance as more perfect, and a lower value as a failure. The results above support $\boldsymbol{H_{MII} - 4.2}$ that a more predictable pattern of intervals during mixed-initiative interaction can give participants a stronger sense of confidence in their task performance.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↗"	Rhythm	721.636	1	721.636	6.611	**.018***
	Error (Rhythm)	2292.363	21	109.160		

Table 4.12: Contrast analysis for ratings on TLX success sub-scale (Hypothesis $\boldsymbol{H_{MII} - 4.2}$: "↗" trend)

As hypothesised in $\boldsymbol{H_{MII} - 3.1}$, if it were true that a predictable rhythm in mixed-initiative interaction could reduce the user's perceived effort, then participants' ratings for their perceived amount of effort devoted to accomplish the tasks would be lower in the tasks that had predictable rhythms. In other words, there might be either a downward linear trend ("↘") across the four conditions in the ratings on the TLX "effort" sub-scale. At the same time, given that the ***Usr-Sys*** condition represents a rigid entrainment (i.e. the system's pace strictly copied participants' pace), it was uncertain whether it would be definitely better than the ***Sys-pr*** condition (e.g. the system's pace was perfectly rhythmic). This was because on the one hand, more user initiative combined with system entrainment should have been beneficial, on the other hand, strict entrainment such as the one in this experiment was found to be less positively perceived than not-so-perfect entrainment in previous social psychology studies (Warner et al., 1987; Clayton et al., 2005). Hence, the ***Usr-Sys*** condition may cause a "dip" in the trend in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. Therefore, an alternative cubic trend ("↘↗↘") in which the ratings in

Sys-pr and ***Usr-r*** conditions were lower than in ***Usr-Sys*** might also exist. Two sets of contrasts were therefore assigned to match each trend, as shown in Table 4.13.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -3$
Contrast weights for "↘↗↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 1$	$\lambda_2 = -3$	$\lambda_3 = 3$	$\lambda_4 = -1$

Table 4.13: Weights in contrast analysis for ratings on TLX effort sub-scale (Hypothesis $\boldsymbol{H_{MII} - 3.1}$: "↘" or "↘↗↘" trend)

The results of contrast analysis are shown in Table 4.14. With the alpha level corrected using the *Bonferroni* method ($0.05/2 = 0.025$), a significant cubic trend was confirmed (p=0.006). This trend is visualised in Figure 4.10: participants reported to have devoted less effort in ***Sys-pr*** and ***Usr-r*** conditions where the rhythm of interaction was predictable, compared with ***Sys-ii*** and ***Usr-Sys*** conditions where the timing was not as regular. Hence hypothesis $\boldsymbol{H_{MII} - 3.1}$ is supported.

Direction	*Sourc*Direction*e*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↘"	Rhythm	320.727	1	320.727	2.187	.154
	Error (Rhythm)	3079.273	21	146.632		
"↘↗↘"	Rhythm	744.727	1	744.727	9.335	**.006***
	Error (Rhythm)	1675.273	21	79.775		

Table 4.14: Contrast analysis for ratings on TLX effort sub-scale (Hypothesis $\boldsymbol{H_{MII} - 3.1}$: "↘" or "↘↗↘" trend)

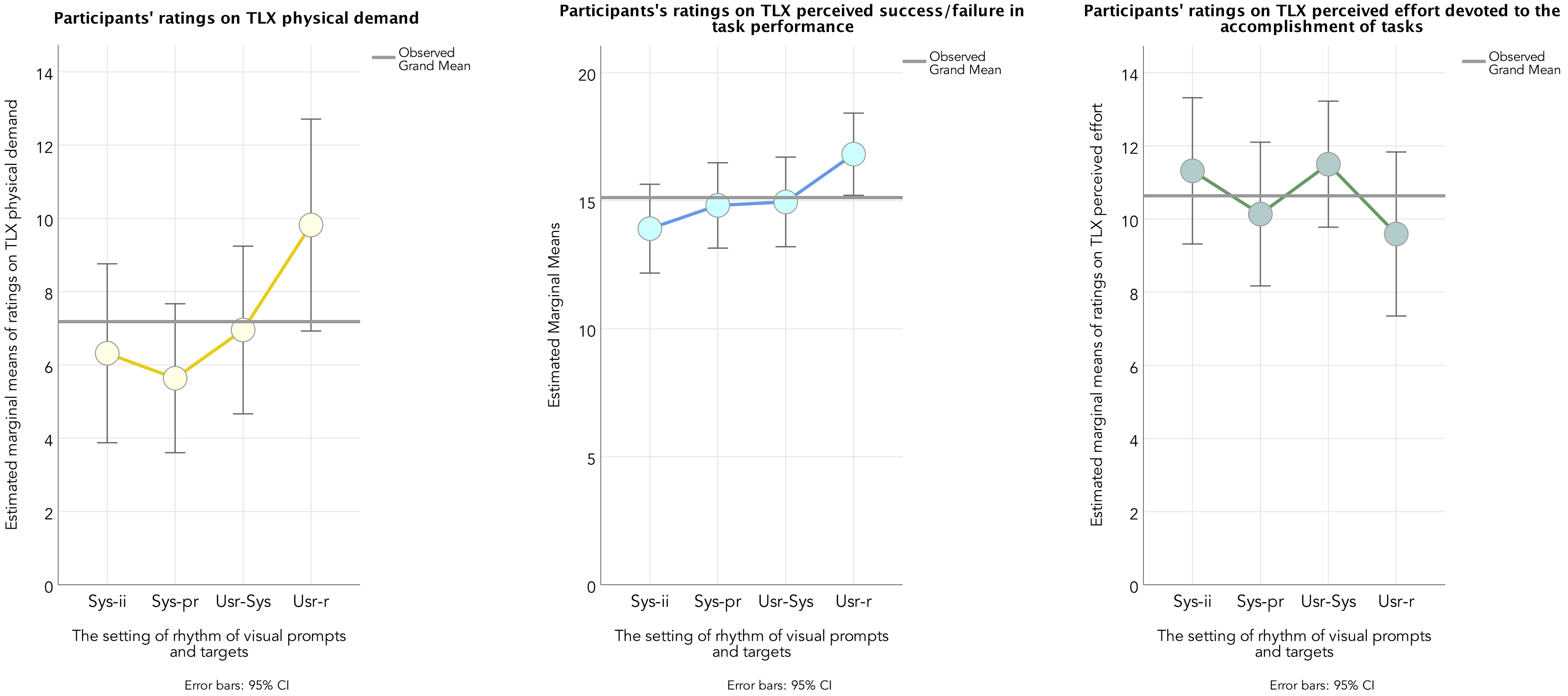


Figure 4.10: Participants' ratings on the TLX scale (physical demand, perceived effort devoted to the task, and perceived success/failure in task performance) in different tasks in Experiment 1

Following the confirmation that significant trends exist in the ratings on "physical demand", "success" and "effort" as predicted, pairwise comparisons were then conducted among four conditions. The Wilcoxon Signed Ranks Test was adopted because the data did not pass the normality test earlier, and the results are presented in Table 4.15. As the alpha level had been corrected to 0.05/6 = 0.0083 using the *Bonferroni* method, significant difference was only found between ***Sys-ii***, ***Sys-pr*** and ***Usr-r*** on the TLX "physical demand" sub-scale and between ***Usr-Sys*** and ***Usr-r*** on the "success" sub-scale. Difference was observed among other pairs of conditions though not to a significant level after the *Bonferroni* correction.

Pair	***Sys-ii***< ***Sys-pr***	***Sys-ii***< ***Usr-Sys***	***Sys-ii***< ***Usr-r***	***Sys-pr***< ***Usr-Sys***	***Sys-pr***< ***Usr-r***	***Usr-Sys***< ***Usr-r***
TLX physical demand						
Z	-0.791	-1.068	-2.664	-2.045	-3.202	-2.401
Sig.	.429	.285	**.008***	.041	**.001***	.016
TLX success						
Z	-0.242	-1.976	-1.950	-1.463	-3.202	-2.954
Sig.	.808	.048	.051	.143	.043	**.003***
TLX effort						
Z	-2.229	-.242	-2.103	-2.199	-0.404	-2.075
Sig.	.026	.809	0.035	0.028	0.686	.038

Table 4.15: Pairwise comparisons for ratings on TLX sub-scales (Wilcoxon Signed Ranks Test)

To summarise, according to the omnibus test, the rhythm setting had caused a significant overall effect on participants' ratings on three TLX sub-scales. Hypotheses $\boldsymbol{H_{MII}-3.1}$ and $\boldsymbol{H_{MII}-4.2}$ are supported by the results of contrast analysis, in which a significant upward linear trend was confirmed to exist in the rating data for "success" and a significant cubic trend in "effort" rating. However, pairwise difference was not as significant in every two conditions under the *Bonferroni* correction. Therefore while the results supported $\boldsymbol{H_{MII}-3.1}$ and $\boldsymbol{H_{MII}-4.2}$, caution is required when interpreting the findings. Their limitations will be discussed in **Section 4.3.2** later.

4.2.3 Entrainment behaviours

The cross-correlation and auto-correlation coefficients between intervals, as introduced in **Chapter 3**, were calculated and analysed in order to test hypothesis $\boldsymbol{H_{MII} - 2}$.

Because the recorded intervals between the appearance of every two prompt crosses and those between every two random geometric targets were identical within any round in the ***Sys-pr*** condition, the standard deviation of the intervals was always 0, hence the cross-correlation formula in **Section 4.1.4** was not applicable to the ***Sys-pr*** condition. The within-round cross-correlation coefficients between *Prompt* intervals and *Recall* intervals in the other three conditions passed the Shapiro-Wilk Normality Test and did not violate Mauchly's Sphericity assumption ($Mauchly's\ W$=0.879, χ^2=2.587, *DoF*=2, *p*=0.274). Therefore ANOVA with repeated measures was adopted, and the results in Table 4.16 confirmed that the overall main effect of rhythm on the coefficient was significant (p<0.001).

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Cross-correlation	1.465	2	0.733	31.630	**.000***
Error (Rhythm)	0.973	42	0.023		

Table 4.16: Omnibus test for cross-correlation coefficients between *Prompt* intervals and *Recall* intervals (ANOVA with repeated measures)

As hypothesised in $\boldsymbol{H_{MII} - 2}$, a more predictable rhythm in mixed-initiative interaction would be more likely to induce the user's entrainment behaviours, while unpredictable and irregular timing would have the opposite effect. Assuming $\boldsymbol{H_{MII} - 2}$ were true, participants' entrainment behaviours would be stronger in the ***Usr-Sys*** condition (where the system took the initiative following the rhythm set by the user) than either ***Sys-ii*** or ***Usr-r***. In other words, a "↗↘"-shaped quadratic trend should be expected in the within-round cross-correlation coefficients between *Prompt* intervals and *Recall* intervals, since the coefficients would be larger in the ***Usr-Sys*** condition than in ***Sys-ii*** or ***Usr-r***. In addition, because the rhythm in ***Usr-r*** would be more predictable than in ***Sys-ii***, an upward linear trend ("↗") might be seen in the order of ***Sys-ii*** → ***Usr-r*** → ***Usr-Sys***. Hence two sets of polynomial contrast weights were assigned accordingly in Table 4.17.

Conditions (Tasks)	***Sys-ii***	***Usr-Sys***	***Usr-r***
Contrast weights for "$\nearrow\searrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_3 = 2$	$\lambda_4 = -1$
Contrast weights for "$\nearrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_3 = 1$	$\lambda_4 = 0$

Table 4.17: Weights in contrast analysis for cross-correlation coefficients between *Prompt* intervals and *Recall* intervals (Hypothesis $\boldsymbol{H_{MII} - 2}$: "$\nearrow\searrow$" or "$\nearrow$" trend)

As shown in Table 4.18, the contrast analysis results for both trends were very significant (p<0.001), after the alpha level being corrected using the *Bonferroni* method ($0.05/2 = 0.025$). As pictured in Figure 4.11(a), the average cross-correlation coefficient between *Prompt* intervals and *Recall* intervals was larger in the ***Usr-Sys*** condition than that of the ***Sys-ii*** and ***Usr-r*** conditions, while the coefficient in ***Usr-r*** was also larger than ***Sys-ii***. Since a greater value of the cross-correlation coefficient suggests a stronger tendency to entrainment, the results above support $\boldsymbol{H_{MII} - 2}$, that participants entrained their *Recall* intervals with predictable *Prompt* intervals and *Target* intervals, but did not entrain as much when the intervals of system-initiated events were irregular.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"$\nearrow\searrow$"	Rhythm	15.989	1	15.989	78.717	**.000***
	Error (Rhythm)	4.265	21	0.203		
"$\nearrow$"	Rhythm	3.245	1	3.245	72.195	**.000***
	Error (Rhythm)	0.944	21	0.045		

Table 4.18: Contrast analysis for cross-correlation coefficients between *Prompt* intervals and *Recall* intervals (Hypothesis $\boldsymbol{H_{MII} - 2}$: "$\nearrow\searrow$" or "$\nearrow$" trend)

Similar procedures were applied during the analysis of the average cross-correlation coefficient between the *Target* intervals and *Recall* intervals within one round in the ***Sys-ii***, ***Usr-Sys*** and ***Usr-r*** condition. The data passed the Shapiro-Wilk Normality Test but did not pass Mauchly's Sphericity test (*Mauchly's W*=0.669, χ^2=8.054, *DoF*=2, p=0.018), hence non-parametric Friedman Test was used, which revealed that there was also a significant difference in the coefficient among the three conditions, as shown in Table 4.19.

Measurement	*N*	χ^2	*df*	***Sig***
Cross-correlation	22	26.455	2	**.000***

Table 4.19: Omnibus test for cross-correlation coefficients between *Target* intervals and *Recall* intervals (Friedman Test)

Under the same hypothesis ($\boldsymbol{H_{MII}-2}$), a similar “↗↘”-shaped quadratic trend should be expected in the within-round cross-correlation coefficients between *Target* intervals and *Recall* intervals, a similar upward linear trend (“↗”) might be seen in the order of ***Sys-ii*** → ***Usr-r*** → ***Usr-Sys***. Hence the two sets of polynomial contrast weights assigned in Table 4.20 were the same as Table 4.17.

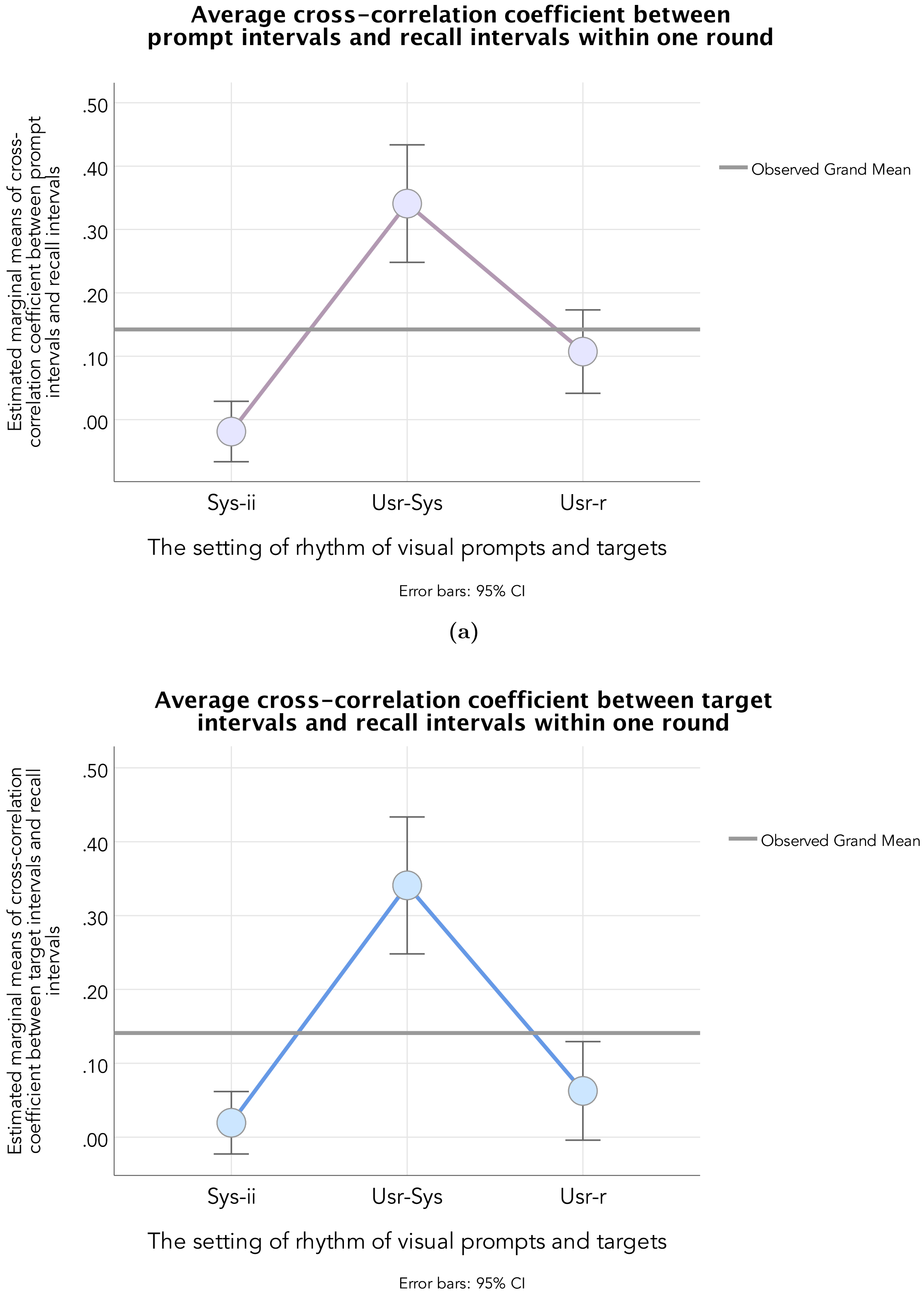
Conditions (Tasks)	***Sys-ii***	***Usr-Sys***	***Usr-r***
Contrast weights for “↗↘” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_3 = 2$	$\lambda_4 = -1$
Contrast weights for “↗” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_3 = 1$	$\lambda_4 = 0$

Table 4.20: Weights in contrast analysis for cross-correlation coefficients between *Target* intervals and *Recall* intervals (Hypothesis $\boldsymbol{H_{MII}-2}$: “↗↘” or “↗” trend)

The contrast analysis results are shown in Table 4.21. Once again both the quadratic and the linear trends were very significant (p<0.001), with the alpha level being corrected using the *Bonferroni* method ($0.05/2 = 0.025$). Figure 4.11(b) demonstrates that the average cross-correlation coefficient between *Target* intervals and *Recall* intervals was larger in ***Usr-Sys*** than ***Sys-ii*** and ***Usr-r*** conditions, and larger in ***Usr-r*** than ***Sys-ii***. Therefore, hypothesis $\boldsymbol{H_{MII}-2}$ is further supported.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
“↗↘”	Rhythm	12.855	1	12.855	75.883	**.000***
	Error (Rhythm)	3.558	21	0.169		
“↗”	Rhythm	2.675	1	2.675	70.226	**.000***
	Error (Rhythm)	0.794	21	0.038		

Table 4.21: Contrast analysis for cross-correlation coefficients between *Target* intervals and *Recall* intervals (Hypothesis $\boldsymbol{H_{MII}-2}$: “↗↘” or “↗” trend)



(a)

(b)

Figure 4.11: Average cross-correlation coefficient between (a) *Prompt* intervals and *Recall* intervals (b) *Target* intervals and *Recall* intervals within one round in Experiment 1

To further investigate the nature of the entrainment behaviours caused by different rhythm settings, pairwise comparisons were carried out using paired samples t tests, given that the cross-correlation coefficients were normally distributed. The alpha level was corrected to $0.05/3 = 0.017$ using the *Bonferroni* method. As shown in Table 4.22, all three pairs for *Prompt-Recall* and two pairs for *Target-Recall* interval cross-correlation were significantly different. Hence the results of contrast analysis are further supported, as is hypothesis $\boldsymbol{H_{MII} - 2}$.

Pair	***Sys-ii< Usr-Sys***	***Sys-ii< Usr-r***	***Usr-Sys> Usr-r***
Cross-correlation (Prompt vs. Recall intervals))			
t	-7.292	-3.402	4.661
df	21	21	21
Sig.	**.000***	**.003***	**.000***
Cross-correlation (Target vs. Recall intervals))			
t	-7.772	-1.384	5.281
df	21	21	21
Sig.	**.000***	.181	**.000***

Table 4.22: Pairwise comparisons for cross-correlation coefficients 1) between *Prompt* intervals and *Recall* intervals and 2) between *Target* intervals and *Recall* intervals (paired samples t test)

The analysis of the average auto-correlation coefficient of participants' *Recall* intervals between two successive rounds provide strengthened support for $\boldsymbol{H_{MII} - 2}$. The coefficients in all four conditions and in Task 0 (fully self-paced clicking on the prompt cross) passed the Shapiro-Wilk Normality Test and Mauchly's Sphericity test ($Mauchly's\ W$=0.643, χ^2=8.563, DoF=9, p=0.480), hence the data was analysed using ANOVA with repeated measures in SPSS. The results in Table 4.23 show that the manipulation of the setting of the rhythm produced a very significant effect ($p<0.001$) on different tasks.

As hypothesised in $\boldsymbol{H_{MII} - 2}$, when the rhythm of mixed-initiative interaction is predictable, the user is more likely to entrain to that rhythm. This also means that they are less likely to adopt a self assimilation time-keeping strategy, hence there may be a downward linear trend ("↘") in the auto-correlation coefficient. The weights for such a trend were assigned in Table 4.24.

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Auto-correlation	1.815	4	0.454	18.702	**.000***
Error (Rhythm)	2.038	84	0.024		

Table 4.23: Omnibus test for auto-correlation coefficients of participants' *Recall* intervals (ANOVA with repeated measures)

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***	***Free***
Contrast weights for "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 2$	$\lambda_2 = 1$	$\lambda_3 = 0$	$\lambda_4 = -1$	$\lambda_5 = -2$

Table 4.24: Weights in contrast analysis for auto-correlation coefficients of participants' *Recall* intervals (Hypothesis $\boldsymbol{H_{MII} - 2}$: "↘" trend)

The results of contrast analysis shown in Table 4.25 confirmed that this linear trend was very significant (p<0.001). As can be seen in Figure 4.12, the average auto-correlation coefficient of participants' free pace clicking intervals in Task 0 was lower than the auto-correlation of participants' *Recall* intervals in two successive rounds in the other four task conditions, and among the four conditions, the auto-correlation increased as the rhythm became less predictable or under control. This suggests that participants might have been making an effort to maintain their own rhythm and not to entrain with the system-imposed rhythm when the system took all of or part of the initiative.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↘"	Rhythm	16.678	1	16.678	49.769	**.000***
	Error (Rhythm)	7.037	21	0.335		

Table 4.25: Contrast analysis for auto-correlation coefficients of participants' *Recall* intervals (Hypothesis $\boldsymbol{H_{MII} - 2}$: "↘" trend)

To further examine the auto-correlation coefficients, two groups of pairwise comparisons were carried out. The first group had 6 pairs, formed among the four formal task conditions, while the second group had 4 pairs, formed between Task 0 and each of the four conditions. The alpha level was adjusted to 0.05/6 = 0.0083 and

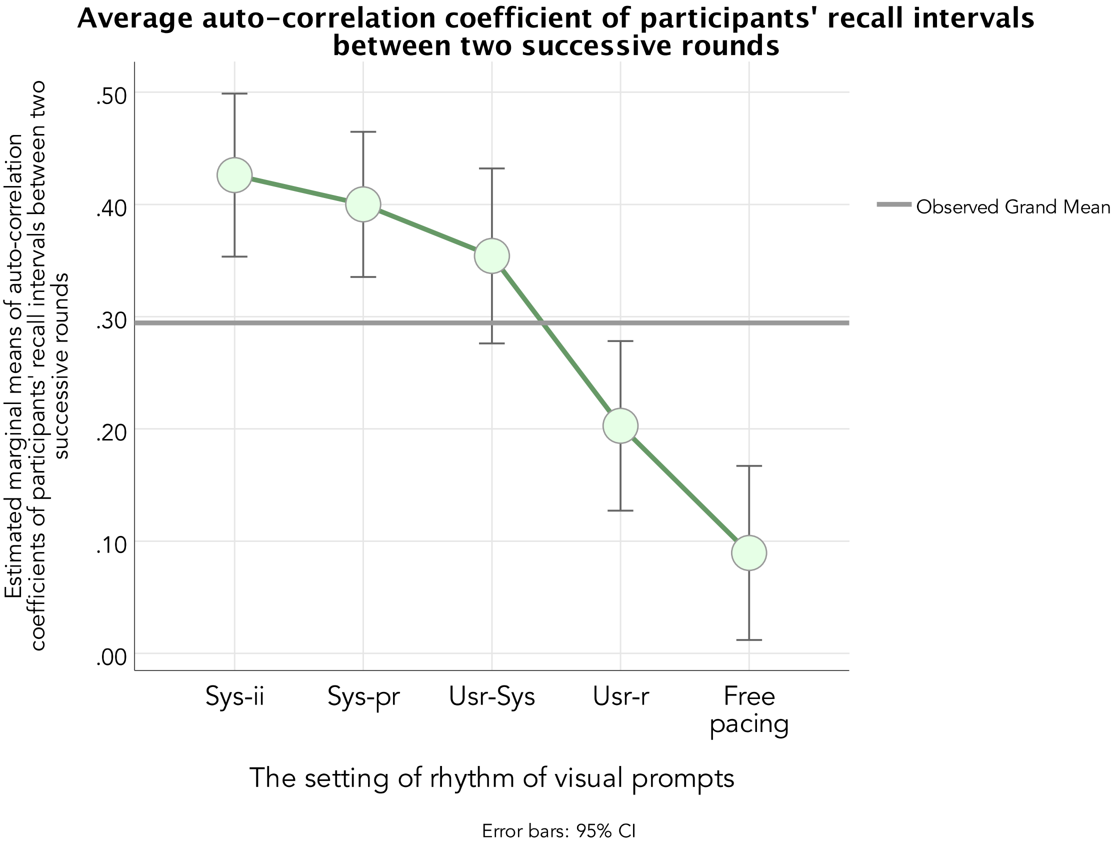


Figure 4.12: Average auto-correlation coefficient of participants' *Recall* intervals in two successive rounds in Experiment 1

$0.05/4 = 0.0125$ respectively with the *Bonferroni* correction. The results are shown in Table 4.26. The average auto-correlation coefficient of participants' free pace clicking intervals in Task 0 was significantly lower than the auto-correlation of participants' *Recall* intervals in two successive rounds in the other four formal tasks. The coefficient in the **Usr-r** condition was also significantly weaker than the other conditions. This suggests that participants exhibited as much self assimilation in the ***Sys-ii*** condition as they did in ***Sys-pr*** and ***Usr-Sys***, which indicates that they might have tried to maintain their own rhythm and did not entrain with the system-imposed rhythm when the system took all of or part of the initiative.

In short, the rhythm setting caused a significant overall effect on the cross-correlation and auto-correlation coefficients according to the omnibus tests, and the results of contrast analysis confirm that participants exhibited a significantly higher

Pair	***Sys-ii*>** ***Sys-pr***	***Sys-ii*>** ***Usr-Sys***	***Sys-ii*>** ***Usr-r***	***Sys-pr*>** ***Usr-Sys***	***Sys-pr*>** ***Usr-r***	***Usr-Sys*>** ***Usr-r***
t	.620	1.520	4.950	1.288	4.194	3.342
df	21	21	21	21	21	21
Sig.	.542	.143	**.000***	.212	**.000***	**.003***

Pair	***Self*<** ***Sys-ii***	***Self*<** ***Sys-pr***	***Self*<** ***Usr-Sys***	***Self*<** ***Usr-r***
t	-6.212	-6.412	4.674	2.548
df	21	21	21	21
Sig.	**.000***	**.000***	**.000***	.019

Table 4.26: Pairwise comparisons for auto-correlation coefficients of participants' *Recall* intervals (paired samples *t* test)

level of entrainment when the rhythm of mixed-initiative interaction was predictable, therefore $\boldsymbol{H_{MII} - 2}$ is supported. Not all condition pairs in the pairwise comparisons were found to be significantly different under the *Bonferroni* correction, therefore the findings should be interpreted with caution. The limitations will be discussed in **Section 4.3.2**.

4.2.4 Number of correct recalls of shape and location

As analysed in **Section 4.2.2**, a significant linear trend was confirmed in participants' ratings on the TLX "success" sub-scale, meaning that they perceived their task performance to be more perfect when the rhythm of the mixed-initiative interaction was more predictable and under their control. In order to support hypothesis $\boldsymbol{H_{MII} - 4.1}$, the number of accurate recalls made by the participants in each task condition was investigated.

The non-parametric Friedman Test was used to detect the overall main effect of rhythm because the numbers of correct recalls did not pass the Shapiro-Wilk Normality test (p<0.001). The main effect was significant (p=0.037) across four conditions, as shown in Table 4.27.

If hypothesis $\boldsymbol{H_{MII} - 4.1}$ were true, the number of participants' accurate recalls in each task condition should follow an upward linear trend ("↗") across four conditions

Measurement	*N*	χ^2	*df*	***Sig***
Number of accurate recalls	22	8.497	3	**.037***

Table 4.27: Omnibus test for number of accurate recalls (Friedman Test)

just as the ratings on TLX “success” sub-scale. The weights were assigned accordingly in Table 4.28.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for “$\nearrow$” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 4.28: Weights in contrast analysis for number of accurate recalls (Hypothesis $\boldsymbol{H_{MII}}-$**4.1**: “$\nearrow$” trend)

Despite the significant overall effect revealed in the omnibus test, the results of contrast analysis in Table 4.29 show that the hypothesised linear trend is not significant.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
“$\nearrow$”	Rhythm	192.045	1	192.045	1.446	.243
	Error (Rhythm)	2788.955	21	132.807		

Table 4.29: Contrast analysis for number of correct recalls (Hypothesis $\boldsymbol{H_{MII}} -$ **4.1**: “$\nearrow$” trend)

In order to find further insight, a *post-hoc* analysis was carried out. Surprisingly, if the weights were reassigned so that the linear trend would follow the direction of ***Usr-Sys*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-r*** as in Table 4.30, then a significant upward linear trend (“$\nearrow$”, p=0.017) could be found in the contrast analysis, with alpha level corrected using the *Bonferroni* method ($0.05/2 = 0.025$). These results are presented in Table 4.31 and Figure 4.13.

The results above mean that, although it was unexpected that participants made less accurate recalls in the ***Usr-Sys*** condition than expected[3], their performance did improve following the predicted upward direction (“$\nearrow$”) in the other three conditions

[3] A potential explanation to this exception is provided in **Sections 4.2.2** and **6.4.2**.

Conditions (Tasks)	***Usr-Sys***	***Sys-ii***	***Sys-pr***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_3 = -3$	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_4 = 3$

Table 4.30: New weights in *post-hoc* contrast analysis for number of accurate recalls (*post-hoc*: "↗" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↗"	Rhythm	1910.227	1	1910.227	6.732	**.017***
	Error (Rhythm)	5958.773	21	283.751		

Table 4.31: *Post-hoc* contrast analysis for number of correct recalls (*post-hoc*: "↗" trend)

(***Sys-ii*** → ***Sys-pr*** → ***Usr-r***), in which the rhythm of mixed-initiative interaction was becoming more predictable and under control. Hence hypothesis $\boldsymbol{H_{MII} - 4.1}$ is partially supported. In other words, a predictable rhythm, either set by the computer system or by participants themselves, in the presentation of visual prompts and targets might have helped participants perform better in tasks that require higher-order cognitive constructs such as working memory.

Following the above analysis, pairwise comparisons were carried out on the number of participants' accurate recalls between different conditions. Because the data was not normally distributed, the Wilcoxon Signed Ranks Test was adopted. As shown in Table 4.32, differences were observed among conditions but not to a significant level after the alpha level was corrected to 0.05/6 = 0.0083 using the *Bonferroni* method.

Pair	***Sys-ii***< ***Sys-pr***	***Sys-ii***< ***Usr-Sys***	***Sys-ii***< ***Usr-r***	***Sys-pr***< ***Usr-Sys***	***Sys-pr***< ***Usr-r***	***Usr-Sys***< ***Usr-r***
Z	-1.288	-1.264	-1.976	-1.936	-0.598	-2.446
Sig.	.198	.206	.048	.053	.550	.014

Table 4.32: Pairwise comparisons for number of correct recalls (Wilcoxon Signed Ranks Test)

In summary, based on the results of the omnibus test, the rhythm setting did cause a significant overall effect on the number of participants' accurate recalls. In

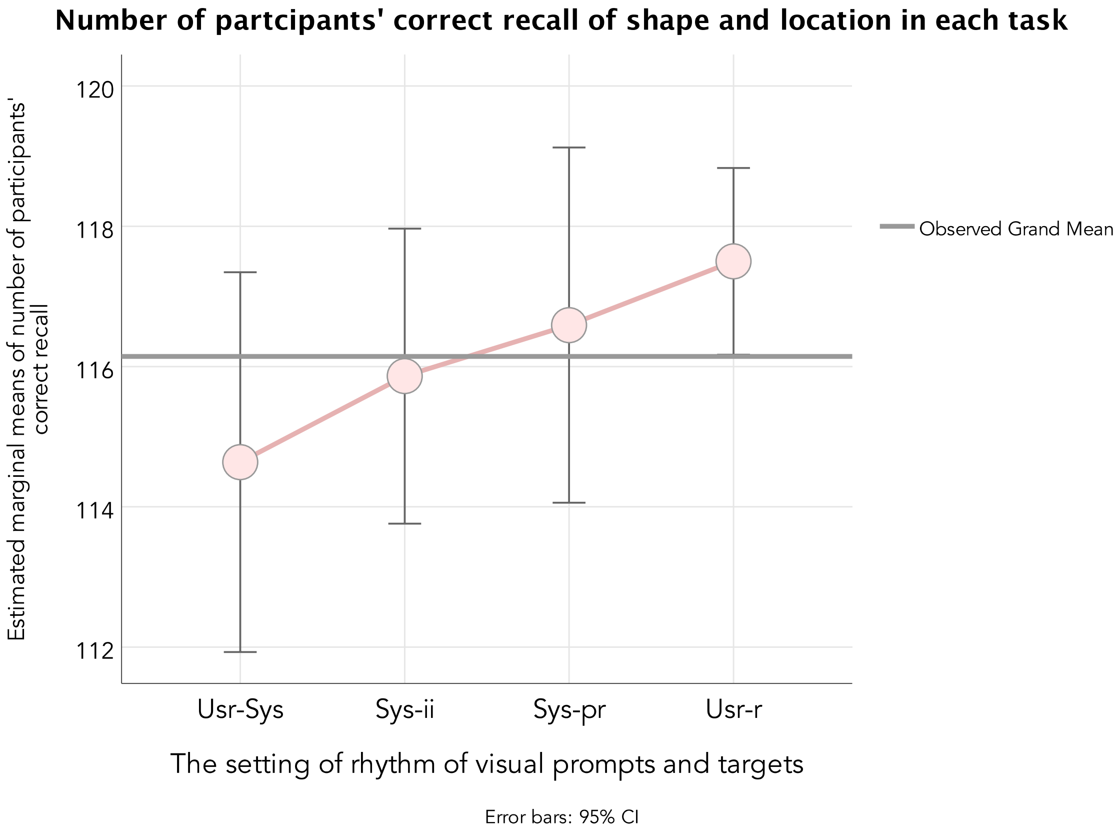


Figure 4.13: The number of participants' correct recalls of shape and location of the visual targets in Experiment 1

contrast analysis, it was found that a predictable rhythm, either set by the computer system (***Sys-pr***) or by participants themselves (***Usr-r***), in the presentation of visual prompts and targets had helped participants perform better in the experiment tasks, hence hypothesis $\boldsymbol{H_{MII} - 4.1}$ is partially supported. However, the findings above should be only accepted with limited confidence and due caution, because the results of pairwise comparisons with the *Bonferroni* correction failed to support the hypothesis. The limitations will be discussed in **Section 4.3.2**.

4.3 Further analysis and discussion

4.3.1 Design implications

In this experiment, as shown in the contrast analysis in **Section 4.2.1**, when participants had more manual control over the interaction rhythm, their reported sense of control significantly increased in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. They also reported that their own performance got significantly better in this direction, as shown in **Section 4.2.2**. Furthermore, despite the fact that they reported the task as being significantly more physically demanding in this direction, they reported to have devoted significantly less effort in it, as shown in **Section 4.2.1**. From the findings above, the first design implication is drawn as follows:

> ***Design implication*** **1.1**: During mixed-initiative interaction, greater reliance on manual control at a relatively micro level can reduce the user's perceived effort and stress. The user may prefer devoting more physical effort and taking more initiative in exchange for a stronger sense of control to letting the system take more initiative.

As reported in **Section 4.2.4**, during contrast analysis, a significant "↗" trend was found in the number of accurate recalls made by participants following the direction of ***Usr-Sys*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-r***. Interestingly, the difference between ***Sys-pr*** and ***Usr-r*** was not as significant under pairwise comparison. This may imply that even though participants were not in any control of the rhythm of the interaction and the system was rhythmically initiating all the visual stimuli in Task 2 (***Sys-pr***), their task performance was almost as good as that of Task 4 (***Usr-r***).

Supporting evidence to the observation above was presented in **Section 4.2.2**. A significant "↘↗↘" trend was found on participants' rating on the TLX effort sub-scale in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. This indicates that participants perceived that they devoted much less effort in Task 2 (***Sys-pr***) and Task 4 (***Usr-r***), compared with when the system took the initiative arrhythmically in Task 1 (***Sys-ii***) or when participants took half of the initiative in Task 3 (***Usr-Sys***).

The two phenomena above can be explained by studies in both cognitive psychology and neuroscience: according to the *dynamic attending theory* proposed by Large and Jones (1999), when external stimuli appear in a regular temporal pattern, the human brain can form temporal expectation and dynamically concentrate its "attentional energy" to future moments that are metrically aligned with the pattern. According to the *diffusion model* proposed by Rohenkohl et al. (2012), the temporal regularity of visual stimuli can shape temporal expectations that sharpen the rate of the accumulation of sensory evidence, which can boost neuronal excitability and increase "the speed and accuracy of perceptual decisions" (Rohenkohl et al., 2012). Therefore the second design implication is:

> ***Design implication* 1.2**: During mixed-initiative interaction, when the system is taking more initiative and the user does not have enough manual control, repetitive system-initiated events (on a micro level) should happen regularly rather than randomly in time, so that the user can form temporal expectation. This can help them focus their attention to important information and facilitate the processing of information.

Interestingly, when participants had full control over the interaction pace during Task 0 (free pacing) and Task 4 (***Usr-r***), they did not maintain a stable rhythm over time. However, when the system was more involved in the loop and taking more initiative (i.e. ***Usr-Sys***, ***Sys-pr*** and ***Sys-ii***) in the interaction, it seemed that participants started to regulate the rhythm of their own actions. This can be seen from the contrast analysis of the auto-correlation coefficients in **Section 4.2.3**: the value of auto-correlation coefficient became significantly lower ("↘") in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** → ***Task 0 (free pacing)***.

In addition, the contrast analysis in **Section 4.2.3** confirmed a significant "↗↘" trend in the cross-correlation coefficients in the direction of ***Sys-ii*** → ***Usr-Sys*** → ***Usr-r***. This indicates that participants did not entrain with arrhythmic system-initiated *Prompt* intervals and *Target* intervals in the same way that they did with system-entrained intervals or self-initiated intervals.

The two observations above suggest that participants tended to manifest a higher level of self-assimilation in time in their own clicking sequences when the system had more control over the interaction pace, as if struggling against external unpredictability posed by the system-initiated actions. A plausible interpretation of this phenomenon is that participants' maintenance of their own pace could have been a strategy to assert control over the interaction, even though it was just an attempt to preserve their own agency and they did not in fact change the timing of the system's behaviours. Hence the third design implication can be drawn as follows:

> ***Design implication* 1.3**: During mixed-initiative interaction, the involvement of system-initiated events may trigger the user to keep a more stable rhythm in their own actions. The user's tendency to maintaining temporal regularity may have been a gesture or strategy to assert control.

It was also not surprising that participants experienced the least sense of control and confidence, most effort and worst accuracy when the system set an arrhythmic pace in Task 1 (the ***Sys-ii*** condition), as the results of contrast analysis support hypotheses $\boldsymbol{H_{MII}-1}$, $\boldsymbol{H_{MII}-2}$, $\boldsymbol{H_{MII}-3.1}$, $\boldsymbol{H_{MII}-4.1}$, $\boldsymbol{H_{MII}-4.2}$. Considering their loose pace in Task 0 (free pacing) and Task 4 (the ***Usr-r*** condition), maintaining a high level of temporal regularity on their own against an unpredictable external temporal structure imposed by the system may have contributed to their perceived effort.

Hypothesis $\boldsymbol{H_{MII}-3.1}$ can be further supported by the *post-hoc* analysis of the average raw asynchronies in participants' mouse-clicking intervals between every two successive rounds. Similar to the study by Nowicki et al. (2013), in this experiment the raw asynchrony between any two successive rounds was calculated by subtracting the onset time of each clicking event in the *Recall* phase of the $k+1^{th}$ round from the corresponding click in the *Recall* phase of the k^{th} round, as illustrated in Figure 4.14.

The data passed the Shapiro-Wilk Normality Test (DoF=22, p>0.05), so ANOVA with repeated measures was carried out in SPSS to test whether or not the rhythm had caused a significant overall effect on the average raw asynchrony across the conditions. As shown in Table 4.33, the overall effect of rhythm was found to be significant (p=0.022).

Given that "human movement timing is inherently variable" and the asynchrony should accumulate over time (Nowicki et al., 2013), therefore the average raw asynchrony in Task 0 (free-pacing) should be greater in its absolute value among the other conditions. Also given that the user may be more likely to maintain their sense of control by keeping

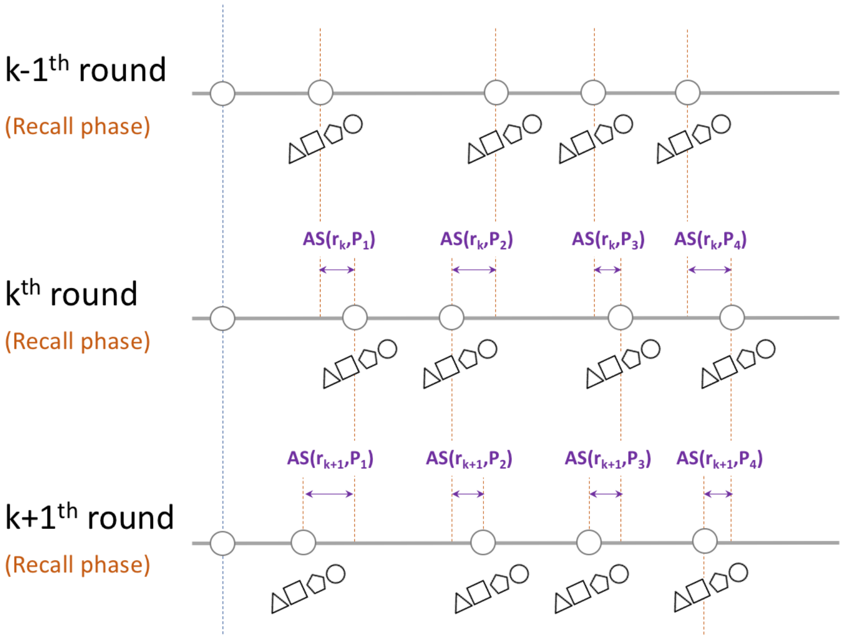


Figure 4.14: Illustration of the raw asynchrony of participants' *Recall* intervals between two successive rounds in Experiment 1

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Avg raw asynchrony	145208.636	4	36302.159	3.030	**.022***
Error (Rhythm)	1006523.98	84	11982.428		

Table 4.33: Omnibus test for average raw asynchrony (ANOVA with repeated measures)

their own rhythm (e.g. the self-assimilation strategy (Nowicki et al., 2013)), and more likely to entrain to flexible and responsive external timing (e.g. a compensation/error-correction strategy (Nowicki et al., 2013)), combining with the analysis in **Section 4.2.3**, the average raw asynchrony in the ***Sys-ii*** condition should also be expected to be greater in its absolute value, and relatively smaller in ***Usr-Sys***. Hence a quadratic trend ("↗↘") should be expected in the average raw asynchrony data, and the weights were assigned accordingly in Table 4.34.

The results of contrast analysis confirmed that the predicted quadratic ("↗↘") trend was significant (p=0.035), as shown in Table 4.35. In Figure 4.15, it can be seen

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***	***Free***
Contrast weights for “$\nearrow\searrow$” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -2$	$\lambda_2 = 1$	$\lambda_3 = 2$	$\lambda_4 = 1$	$\lambda_5 = -2$

Table 4.34: Weights in contrast analysis for average raw asynchrony (hypothesis: “$\nearrow\searrow$” trend)

that the average raw asynchrony in Task 0 (free pace) was more negative (i.e. greater in absolute value) compared with other four tasks, while the average raw asynchrony of the ***Usr-Sys*** condition is surprisingly close to zero.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
“$\nearrow\searrow$”	Rhythm	1019260.233	1	1019260.33	5.104	**.035***
	Error (Rhythm)	4193430.078	21	199687.127		

Table 4.35: Contrast analysis for raw asynchrony (hypothesis: “$\nearrow\searrow$” trend)

The results above suggest that very existence of the system’s involvement throughout the rounds of tasks in the experiment may have induced participants’ sustained effort in a temporal-error-correction process, which inhibited the accumulation of asynchrony. It not only supports ***Design implication*** **1.3**, but also provides a further explanation as to why participants reported that they devoted much more effort when interacting with the system in the ***Sys-ii*** condition even though they did not need to click any of the prompts and targets, than in the ***Usr-r*** condition when they needed to manually click every single stimulus on the screen. This leads to the fourth design implication:

> ***Design implication*** **1.4**: During mixed-initiative interaction, when the user is exerting control through regulating the rhythm of their own actions against the irregular timing of system-initiated events, the attentional energy they devoted to the sustained rhythm-keeping process may contribute to a higher level of perceived effort. Therefore when the system detects a higher temporal regularity from the user’s actions, handing over more initiative to the user may release them from the role of rhythm-keeping and the accompanying effort and stress.

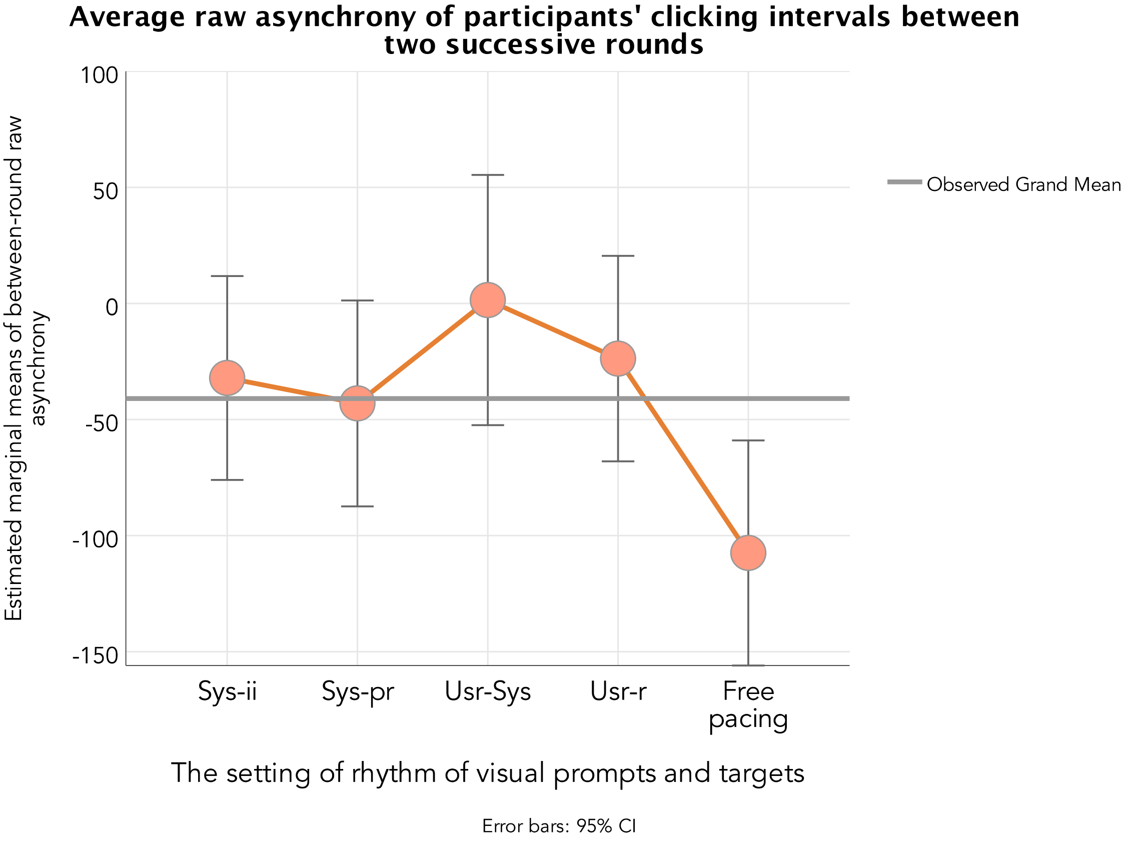


Figure 4.15: Average raw asynchrony in each task in Experiment 1

4.3.2 Limitations

While the findings in Experiment 1 have supported the four sets of hypotheses proposed in **Chapter 3**, there were limitations in the following four aspects:

- First, the experiment was carried out in a highly controlled environment and strictly followed the designed protocol. The visual stimuli presented on the screen were simple, though randomised, geometric shapes that carried about the same amount of information and required the same level of attention. However, in realistic applications, the user may need to deal with a series of mixed-initiative tasks that have mixed levels of difficulty and require different amount of attention, effort and processing time. In addition, the information may not often be as straightforward as simple shapes: the user may need to attend to the semantic, mathematical or logical dimension of the information, which may require higher-order cognitive skills such as more complex decision-making or problem-solving processes in addition to the demand on working memory. Experiment 3 reported in **Chapter 6** aims to address this limitation.

- Second, the average length of interaction intervals in each participant's experiment session ranged from 600ms to 1100ms (with the shortest interval occurred in a session as 257ms), which was a relatively rapid pace and a low granularity in human-computer interaction. Hence further evidence is required in order to determine whether the findings in this experiment will be applicable on a greater timescale: Experiment 3 is designed to investigate this issue. Furthermore, in Tasks 1, 2 and 3, the timing of system-initiated events were either predetermined through averaging and randomisation or determined solely based on the user's clicking intervals, rather than being determined by the system's runtime. However, when interacting with an actual mixed-initiative interface, such as using a Programming-by-Example application, the system may have an intelligent back-end algorithm that runs and updates continuously along with the user's demonstrative behaviours. It may take the system an uncertain and fluctuating amount of time before it can actually be ready to get back to the user. Hence when manipulating the interaction rhythm on a real application, the system's runtime should be considered when determining the time frame between the latest and the next system-initiated events.

- Third, all participants in this experiment are adults who are educated, familiar with interacting with a piece of software using a computer mouse, physically able to make controlled and precise movement when clicking rapid-changing geometric shapes on a computer screen, and capable of maintaining their focus on repetitive tasks for 20-30 minutes. Therefore, the findings and design implications in this experiment need to be generalised with caution and limited confidence, especially when the potential user population covers a wider range of capabilities (e.g. people with impaired mobility or cognitive capacity may find it challenging to attend to or catch up with a relatively fast-paced interaction (Boucugnani & Jones, 1989; Hutchinson et al., 1989; Murata, 2006)) and preferences (e.g. the amount of control people prefer to assume can vary with their age and personal competence (Rodin, 1986; Schieman & Campbell, 2001)), hence more human factors should be considered in future studies and design on mixed-initiative interaction.

- Fourth, although the results of contrast analysis in **Sections 4.2.1**, **4.2.2**, **4.2.3** and **4.2.4** confirmed that the predicted trends were very significant in the experiment data, as hypothesised in $\boldsymbol{H_{MII} - 1}$, $\boldsymbol{H_{MII} - 2}$, $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 4.1}$ and $\boldsymbol{H_{MII} - 4.2}$, a significant difference between each pair of task conditions was not always confirmed in pairwise comparisons. Therefore, the findings and the design implications in this chapter should be interpreted with due caution and accepted with limited confidence. There are, however, two arguments to ease the concerns over insignificant pairwise comparison results. Firstly, all pairwise tests in this chapter were subject to the *Bonferroni* correction, hence the difference between some pairs of conditions that would have been considered as significant otherwise was not accepted after the alpha level was reduced to only $0.05/6 = 0.0083$. The *Bonferroni* correction is recognised as a simple but very conservative method, while it does reduce the likelihood of Type I errors, it also impairs the power of the statistical test and leads to more Type II errors (Nakagawa, 2004), and can be "deleterious to statistical inferences" (Perneger, 1998). Armstrong (2014) argued that the *Bonferroni* correction should only be applied when a large number of *unplanned independent* pairwise tests are carried out without hypotheses. In this experiment, all comparisons were carried out in *planned* tests under sound theoretical hypotheses, and the pairwise comparisons between task conditions were *dependent* to each other, hence the *Bonferroni* correction may have been excessively strict on the results of this experiment. Secondly, when the data failed the pairwise comparisons or even an omnibus test like ANOVA, it does not mean that we can conclude that the independent variable was not effective, as shown in the example in Rosenthal et al. (1985, p. 2)'s book. This is because omnibus tests and independent pairwise tests can reveal a diffused effect on observations, but they entirely disregard the arrangement of different levels of the independent variable that constitute the whole spectrum (Rosenthal et al., 1985; Furr & Rosenthal, 2003; Abdi & Williams, 2010). In this experiment, as described in **Section 4.1.2**, the independent variable had four levels, among which the rhythmic character of the initiative-taking during the interaction became increasingly more predictable and under the user's control. Therefore, contrast analysis can allow us to draw more precise conclusions (i.e. testing the contrast weight coefficients that correspond to the prediction/hypothesis under investigation) than omnibus or pairwise tests can, and the results of contrast analysis should be respected nevertheless.

4.4 Summary

In this chapter, I explored the three research questions in **Chapter 1** and tested the four sets of hypotheses formulated in **Chapter 3** by designing and carrying out a controlled experiment. As reported in **Section 4.1**, I adapted a conventional stimulus-response experiment paradigm in cognitive psychology, and developed an experimental system that can manipulate the timing of system-initiated visual stimuli in a simplified mixed-initiative interaction.

The results of this experiment have provided strong evidence to support my hypotheses. As analysed in **Section 4.2.1**, participants reported a significantly higher sense of control when they took more initiative and had more controlled the pace of the interaction, compared with when the system took part of or all of the initiative. When the system took the initiative in a rhythmic manner or took half of the initiative by emulating the participants' rhythm, participants perceived themselves as significantly more in control than when the system took the initiative at unpredictable and irregular times. Hence the hypothesis $\boldsymbol{H_{MII} - 1}$ is supported.

The analysis of auto-correlation and cross-correlation coefficients in **Section 4.2.3** suggests that participants tended to entrain with the system's rhythm more when it was aligned with their own rhythm, and did not entrain with the system when it was taking the initiative irregularly. Therefore the hypothesis $\boldsymbol{H_{MII} - 2}$ is supported. In addition, the more initiative taken by the system during the interaction, the stronger the tendency participants exhibited to maintaining their own rhythm. Their rhythm-keeping inclination was much weaker when they took the initiative themselves and had full control of the pace.

According to participants' subjective reports, as compared in **Section 4.2.2**, when the system took the initiative in a rhythmic manner, or when they took the initiative and set the pace themselves, they reported the perception that they devoted a significantly less amount effort compared with when the system took the initiative at unpredictable and irregular times. The results support hypothesis $\boldsymbol{H_{MII} - 3.1}$.

Furthermore, as reported in **Sections 4.2.2** and **4.2.4**, when participants took the initiative, they were more confident in their answers and perceived the task as being accomplished more successfully, and they did achieve the best performance.

When the system took the initiative rhythmically, participants' actual performance was as successful as when they controlled the pace themselves, while system's irregular intervals had resulted in a worse performance. Therefore hypothesis $\boldsymbol{H_{MII}} - \mathbf{4.2}$ is supported, and hypothesis $\boldsymbol{H_{MII}} - \mathbf{4.1}$ is supported with the exception in the ***Usr-Sys*** condition.

Based on the above findings, I have provided a basic answer to the three research questions: During mixed-initiative interaction, letting the user take the initiative in their own pace or letting the system take the initiative in a predictable and rhythmic manner can preserve the user's sense of agency and improve their task performance, hence both are appropriate. Also, when the system aligns with the user's rhythm, the user exhibits a stronger tendency to entrainment and appreciates the sense of control. On the other hand, it is inappropriate to let the system take the initiative in an unpredictable and irregular manner, which can impair the user's sense of agency and task performance and cause a higher level of perceived effort.

I also drew four design implications from my experiment findings in **Section 4.3.1**. The first emphasises that in mixed-initiative interaction, the user may be happy with devoting more physical effort in exchange of a stronger sense of control. The second suggests that if the system is to take more initiative, it should do it in a rhythmic manner. The third deduces that the system's initiative taking may cause the user to put extra effort in keeping their own rhythm in order to assert control, and hence the fourth implication is that when the system detects a strong rhythm-keeping tendency from the user's actions, giving the user more opportunities to take back the initiative may release them from the struggle of asserting control.

As discussed in **Section 4.3.2**, the findings of this experiment are limited because of the simplification of mixed-initiative interaction in a controlled experiment, the low granularity of the interaction timescale, the ideality of the participants recruited for this experiment, and not all results that were significant in both omnibus tests and planned contrast analysis were identified in pairwise tests with the *Bonferroni* correction. I will further address and discuss the first two aspects in **Chapter 6**.

CHAPTER 5

Perceived agency and the timing of auditory targets - Experiment 2

The results of Experiment 1 not only show that the manipulation of the timing structure was effective, but also support the hypotheses proposed in **Chapter 3**. It was confirmed that when the user is interacting with visual stimuli presented on a computer screen, a predictable interaction rhythm can preserve the user's sense of control, facilitate their entrainment behaviours, reduce their perceived stress level and enhance their performance in the tasks that require higher-order cognitive resources such as working memory.

Experiment 2 was designed and carried out to further explore the findings above. There were three motivations. First, to see how rhythmic aspects of system-initiated actions can influence users' perception of time (i.e. the intentional binding effect), which can serve as an implicit measure of users' experience of agency that was not measured in Experiment 1. Second, to examine whether or not there is a link between users' internal experience of agency and their explicit report of perceived control. Third, to investigate whether or not the manipulation of temporal structures in Experiment 1 would cause similar effects on users' experience of control and stress when the interaction takes place in the auditory modality in Experiment 2.

5.1 Method

Experiment 2 used the same structure as Experiment 1, in which the temporal structures of the interaction actions between the user and the system was manipulated differently in several tasks. Participants were asked to attend to auditory stimuli (e.g. beep tone) instead of visual ones (e.g. simple shapes) (Haggard, Aschersleben, Gehrke, & Prinz, 2002; Moore, Wegner, & Haggard, 2009; Moore, Lagnado, et al., 2009).

Because this experiment aims to investigate how different timing patterns of the presentation of auditory stimuli can influence the user's experience of internal agency and external control ($\boldsymbol{H_{MII} - 1}$), perceived stress level ($\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$), and confidence in their performance ($\boldsymbol{H_{MII} - 4.2}$), the experiments tasks should have the following characteristics:

1. Just as the tasks in Experiment 1, the tasks in Experiment 2 should also be repetitive or have repetitive steps, on which different temporal structures (i.e. rhythmic, random, entrained) can be imposed.

2. The tasks in Experiment 2 should also have a "turn-taking" dynamics, with a mix of user-initiated actions and system-initiated actions in the tasks or the steps so as to emulate realistic mixed-initiative interaction.

3. As with Experiment 1, the tasks in Experiment 2 should require a reasonable amount of cognitive resources such as working memory. This is to ensure participants are not too occupied to feel any control, or too idle to have differentiating performances under different temporal structures, so that their subjective ratings and time estimation can effectively reflect the phenomena as predicted in $\boldsymbol{H_{MII} - 1}$ and $\boldsymbol{H_{MII} - 4.2}$.

4. The tasks should also avoid using on-screen elements that can distract participants' visual attention away from the Libet Clock, which is a commonly used paradigm in the studies of intentional binding, as reviewed in **Section 2.1.4**.

5. The tasks should exert an appropriate level of pressure on participants, so that participants' ratings for their stress level under different temporal structures will not be too high or too low to be compared when testing $\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$.

6. The auditory stimuli should be clear and simple, without semantic ambiguity or differences in pitch, timbre, or accent. This is to minimise the systematic errors caused by unpredictable and uncontrollable factors, which is particularly important given that participants' time perception is a crucial measure of their experience of agency ($\boldsymbol{H_{MII} - 1}$) and is susceptible to confusion or hesitation.

The tasks in Experiment 2 were designed in keeping with the characteristics above, and **Section 5.1.1** describes the design of the tasks, and how each of the characteristics listed as above was met by the design.

5.1.1 Task design and procedures

The participants of this experiment were the same people who were recruited for Experiment 1, and their background information can be found in Table 4.2 in **Section 4.1.2**. In every experiment session, each participant went through a practice stage before starting with the formal tasks. The practice stage consisted of six kinds of task. Five of the tasks were designed and presented in the same way as the ones in the formal stage, so that participants could walk through all the procedures and become familiar with the interface and the actions involved.

In the first kind of task (Task 0) in both stages, participants were asked to wear a pair of enclosed overhead headphones, from which they could hear a sequence of beeps (frequency: 3600Hz, duration: 50ms) that were played repeatedly at a fixed interval (interval length: 660ms [1]). At the same time a horizontal slider bar was presented on the computer screen, as shown in Figure 5.1, and participants were asked to use the slider bar to adjust the rate of the beeps they were listening to until they found the rate comfortable. The thumb of the slider bar was initialised to the midpoint, which corresponded to the pre-set 660ms interval (i.e. 90 beats per minute). The slider bar was labelled with "slower" on its left end and "faster" on its right end, corresponding to a range of interval length that was between 260ms and 1060ms (i.e. 56∼230 beats per minute). The gradation of the slider bar allowed 1ms-level precision. After participants chose their preferred rate of beeps on the slider bar, the "Confirm" button would only be activated after the beep had been played at that chosen rate for

[1] As reviewed in **Section 2.3.1**, regular intervals of around 600ms will form "maximal pulse salience" zone (London, 2012a)

twelve times, in order to guarantee that participants were sure about their decision. This chosen interval was used to set a default comfortable rhythm customised for each participant in later tasks.

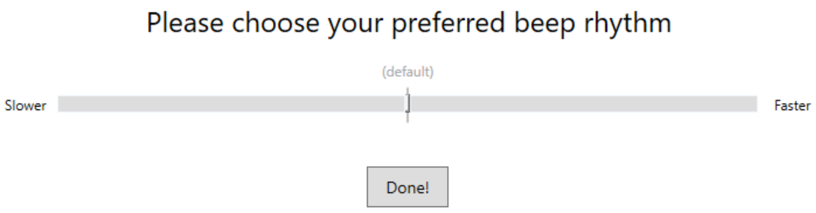

Please choose your preferred beep rhythm

Figure 5.1: The slider bar participants used to select a comfortable rhythm in Experiment 2

Each of another four kinds of task in both stages required participants to listen to a randomised number of identical beeps (frequency: 3600Hz, duration: 50ms, meeting the 6^{th} characteristic) wearing the headphones while observing a standard Libet clock on the computer screen as reviewed in **Section 2.1.3** (Libet et al., 1983), with the clock face located at the centre of the screen sized around 80×80 pixels (displayed as about 22×22 mm^2 on the screen), and its clock hand being 2×40 pixels (about 0.6×11 mm^2), meeting the 4^{th} characteristic.

As with Experiment 1, every task had the same number of rounds of interaction. Participants needed to click the “Ready” button under the Libet clock to start each round, and the Libet clock would start rotating from its initial point (12 o’clock) upon clicking. Each round had a *Prompt* phase, an *Attention* phase, and a *Recall* phase, hence meeting the 1^{st} and 2^{nd} characteristics.

1. In the *Prompt* phase, four identical beeps (frequency: 3600Hz, duration: 50ms) would be played from the headphones. As illustrated in Figure 5.3, in Task 1 and Task 2, the beeps would be triggered by the experimental system, and participants only needed to listen to the beeps while observing the rotating clock. The time intervals of the beeps were controlled by the system. In Task 3 and

Task 4, participants needed to click a grey gradient round button that says "Click!" (size: 66×66 pixels) right under the Libet clock four times to trigger four beeps, as shown in Figure 5.2, hence the beeping intervals were determined by participants' own clicking actions. Upon each click the grey gradient button would disappear for 50ms then re-appear to prevent accidental double clicks made by the participants. The purpose of using a button that was rendered with a gradient background rather than with a sharp outline was to give a cue to the participants, that they needed not to aim the button with precision or place their mouse cursor right in the middle of the button when they clicked.

2. In the *Attention* phase of any given round in Task 1, Task 2 and Task 3, participants only needed to listen to the beeps coming out of the headphones while observing the rotating Libet clock. The number of beeps were randomised, and it could be three, four, five or six. In Task 4, participants needed to keep clicking the same grey gradient round button under the Libet clock to produce beeps while observing its rotating hand. Again the number of clicks and beeps was randomly chosen among three, four, five or six, but participants were able to determine the beep intervals through their own clicking actions. The randomisation of number of beeps met the 5^{th} characteristic. The hand of the Libet clock kept rotating for a random amount of time (500∼1000ms) after the last beep, either produced by the system or the participant.

3. Finally, in the *Recall* phase of any round in all four kinds of task, an empty text input box would appear under the Libet clock, see Figure 5.3. Participants needed to recall and report the position of the clock hand when they heard the last beep by typing numbers into that text box, hence the 3^{rd} and 5^{th} characteristics were met. They were encouraged to make their estimations as accurately as they could, and they were allowed to input either integers or a number with up to two decimal places. The "Next" button under the clock would be activated when the system detected a valid input in the text box, and participants had to click on it in order to move to the next round.

During an experiment session, a participant would first practice each type of task described above for three rounds, then completed thirty rounds of each task during the formal stage. The number of rounds in both stages were determined during pilot studies. Participants were asked to provide subjective ratings after each task.

Just as in Experiment 1, the sequence of the four kinds of task in Experiment 2 was randomised for each participant, as shown in Appendix A.5.2, in order to mitigate the learning effect.

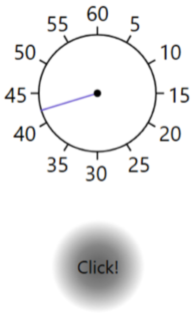
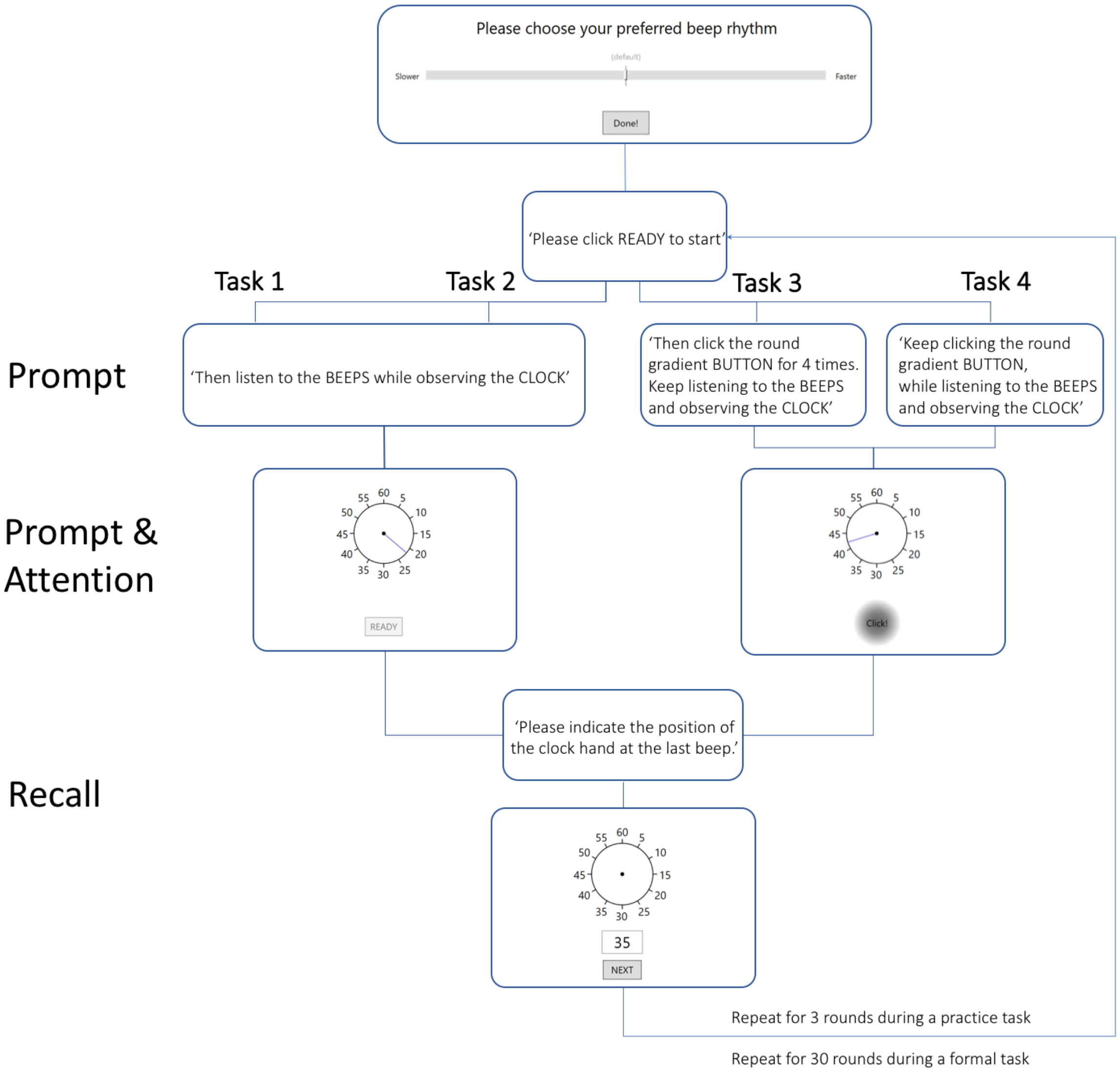


Figure 5.2: The Libet clock and the gradient button used in Experiment 2

The additional one task was a baseline task, which participants needed to complete before all other tasks in the practice stage. This task was used to measure the outcome baseline error for each participant, and also adopted the Libet clock paradigm. In each of the twenty rounds in this task, participants were asked to observe a rotating Libet clock on the screen, while attending to a single beep (frequency: 3600Hz, duration: 50ms) that was generated at a random time (2000∼6000ms) by the experimental system coming from the headphones. The Libet clock kept rotating for a random amount of time (500∼1000ms) after that beep. Again participants needed to report the perceived position of the clock hand when the beep occurred using either integers or numbers with up to two decimal places using a keyboard.

Participants were told that the purpose of this experiment was to explore *"how people follow various sequences of sounds from a computer"*, and again the term "timing" or "rhythm" was avoided in both the recruitment message and the task briefing script in order to minimise participants' bias caused by their prior expectation. The full introduction script for this experiment is included in the Appendix A.3.3. Each participant was also given a debrief when they completed the experiment, explaining that the purpose of this experiment was not only to study how they follow auditory stimuli given by a computer system, but also to explore if the rhythmic aspect of the stimuli presentation had affected their temporal perception and their experience of agency and stress during the interaction. Each session of Experiment 2 lasted for 20∼30 minutes, and the same gift (valued £6∼£8) was given as a reward. This experiment



Please choose your preferred beep rhythm

‘Please click READY to start’

‘Then listen to the BEEPS while observing the CLOCK’

‘Then click the round gradient BUTTON for 4 times. Keep listening to the BEEPS and observing the CLOCK’

‘Keep clicking the round gradient BUTTON, while listening to the BEEPS and observing the CLOCK’

‘Please indicate the position of the clock hand at the last beep.’

Repeat for 3 rounds during a practice task

Repeat for 30 rounds during a formal task

Figure 5.3: Illustration of the task procedures in Experiment 2

was reviewed and approved by the ethics committee of the Computer Laboratory, University of Cambridge.

5.1.2 Independent variable and manipulation

Independent variable	*Description of treatment*	*Abbreviation*
Irregular intervals	***Sys***tem takes the initiative at ***i***rregular ***i***ntervals	***Sys-ii***
Predictable rhythm	***Sys***tem takes the initiative in a ***p***redictable ***r***hythm	***Sys-pr***
	Use***r*** takes the initiative, ***Sys***tem aligns	***Usr-Sys***
	Use***r*** takes the initiative in their own ***r***hythm	***Usr-r***

Table 5.1: The independent variable and its settings in Experiment 2 (same as Experiment 1)

This experiment also adopted a within-subject design, and used the same naming of the independent variable as Experiment 1 (see Table 5.1. As introduced in **Section 5.1.1**, every experiment session began with Task 0, in which participants were asked to adjust the length of interval of a series of rhythmic beeps on a slider bar until they found the rate that was comfortable. After participants had chosen a certain interval length, the interval (denoted as L_i for Participant i) would be used in Task 1 and Task 2.

In Task 1 (***Sys-ii***) and Task 2 (***Sys-pr***), four beeps were first played through the headphones by the system, and participants only needed to observe the rotating Libet clock. In the ***Sys-pr*** condition (Task 2), the inter-beep intervals were of the same length, which was L_i determined earlier by participants themselves in Task 0. In the ***Sys-ii*** condition (Task 1), inter-beep intervals were irregular. Every sequence of random intervals was generated in MathWorks MATLAB R2015b using the same function code as Experiment 1: the mean interval length was set as L_i for Participant i, and all intervals fell into a range between $\frac{1}{2}L_i$ and $\frac{3}{2}L_i$ under continuous uniform distribution. Again, in order to highlight the temporal randomness of the beep sequence to participants, every two adjacent intervals were forced to have a minimum of 25 milliseconds difference in length during the generation process.

In the ***Usr-Sys*** condition (Task 3), after participants clicked the "Ready" button to start each round, they first needed to click the grey gradient button under the Libet clock for four times to trigger four beeps, then they were asked to wait and listen to more beeps coming from the headphones generated by the system while attending to the Libet clock. The intervals between those later beeps fully duplicated those intervals between participants' clicking actions during the *Prompt* phase in the same round. In the ***Usr-r*** condition (Task 4), participants needed to click the grey gradient button at their own rhythm until it disappeared.

In every round in each task, the total number of beeps ranged randomly among seven, eight, nine or ten, with the first four beeps always occurring in the *Prompt* phase and the rest in the *Attention* phase. The sequence of how many beeps would occur in each of the thirty rounds in a task for each participant was also generated randomly and preloaded into the experimental system. Of the thirty rounds, there were nine rounds that had seven, eight and nine beeps, and three rounds with ten beeps. The randomisation of number of beeps in each round was to mitigate the confounding effect that might be introduced due to participants' unconscious reliance on counting and grouping (London, 2012a) and anticipating the last beep. A sample of intervals that were used in this experiment are presented in the Appendix A.4.2.

5.1.3 Dependent variables and measures

Subjective report variables were collected in the same way as for Experiment 1. Participants were asked to rate on six NASA-TLX sub-scales plus another set of questions, as reported in **Section 4.1.3**. As has been done in existing research (Deaton & Parasuraman, 1993; Rowe et al., 1998; Yurko et al., 2010; Mehta & Agnew, 2011), participants' ratings on each individual sub-scale were contrasted among four task conditions when testing hypotheses $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 3.2}$ and $\boldsymbol{H_{MII} - 4.1}$.

Another dependent variable in this experiment was the standard measure of outcome binding as reviewed in **Section 2.1.3** (Coyle et al., 2012). For Participant i, the outcome binding effect in a formal task was calculated using the following formula:

$$Outcome\ binding = Outcome(active\ error) - Outcome(baseline\ error)$$

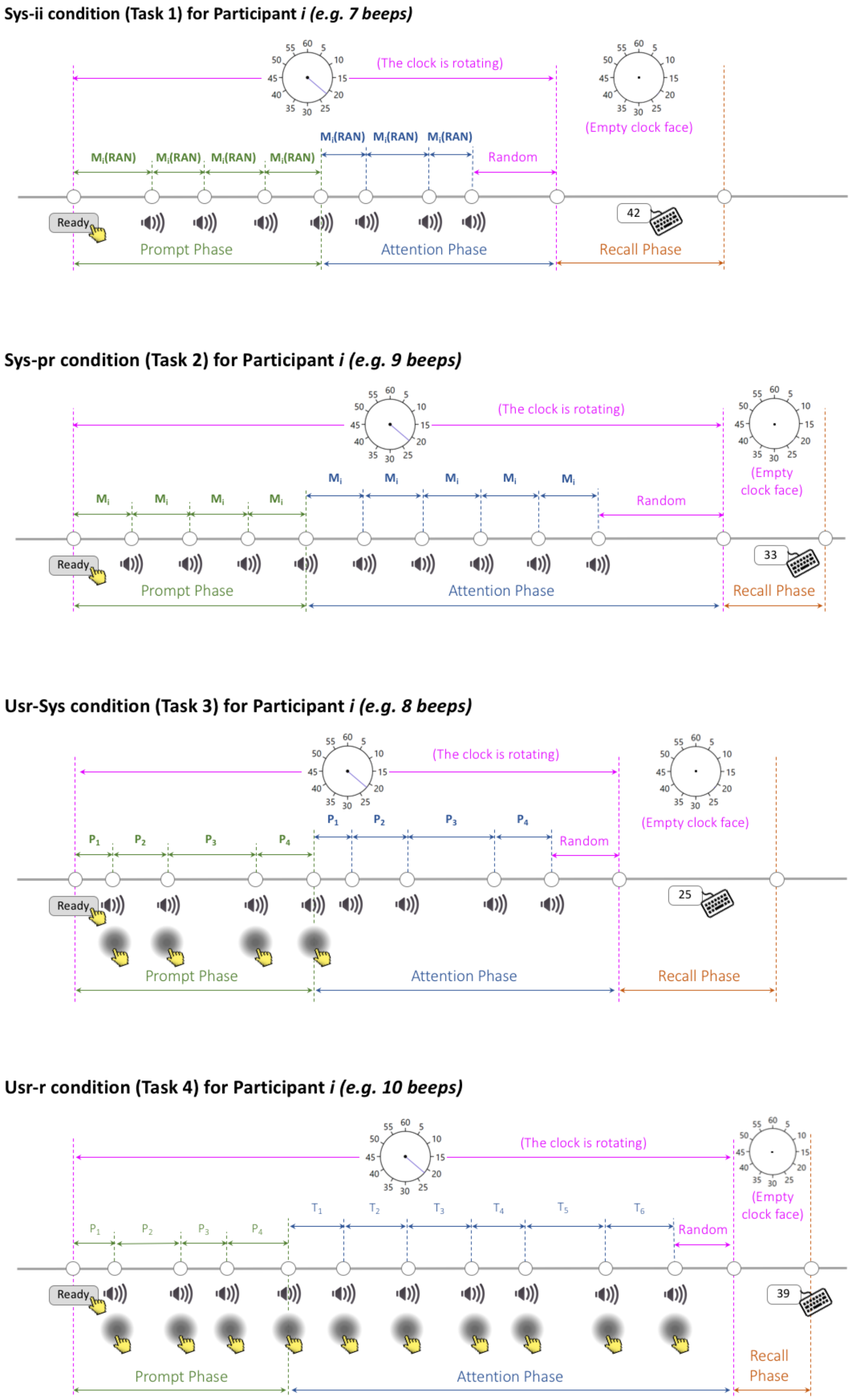


Figure 5.4: Illustration of the temporal structure in a round in each of the four tasks in Experiment 2

where the outcome active error was the average value of the difference between the actual time of the last beep that Participant i was asked to attend to and their reported perceived time of that beep in each of the thirty rounds in the task, while the outcome baseline error was the average value of the difference between the actual time of a random system-generated beep and Participant i's reported perceived time of that beep in the twenty trials in the baseline task during the practice stage. A negative value of the outcome binding effect indicates that participants perceived the beep as occurring later than it did, as shown in Figure 2.1, and a temporal prolonging effect as such correlates with a lower sense of agency. On the other hand, a positive value of outcome binding indicates a temporal attraction effect which happens when participants perceive a higher sense of agency. All components were measured using the Libet clock paradigm and calculated in milliseconds.

5.1.4 Apparatus

All experiment sessions were carried out in Office SS08 in the Computer Laboratory, University of Cambridge. All participants used the same desktop computer (System: Windows 10 Pro, 64-bit; CPU: 2.80GHz; RAM: 8.00GB) with the same computer monitor (Samsung, SM2443BW 24-inch Black Widescreen LCD, 1920×1200) and the same optical mouse (Microsoft IntelliMouse Optical 1.1A). The keyboard used in the experiment was a Logitech Internet 350 Keyboard. The enclosed overhead headphones were a pair of Sennheiser HD 256 Linear Headphones.

The experimental system was implemented using C# as a Windows Presentation Foundation (WPF) application. The software was written completely by the author of the dissertation, and they can be found via this link:

https://github.com/ChristineGuoYu/PhD_Experiment_2

During every experiment session, the programme ran in Visual Studio Community 2015 environment (Version 14.0.23107.0 D14REL).

5.2 Result analysis

The data obtained from Experiment 2 was analysed following the same procedures as Experiment 1. As set out at the beginning of **Section 4.2**, the analysis was carried out in the following way:

Step 1: An omnibus test, either repeated-measure one-way ANOVA or non-parametric Friedman Test depending on the data distribution, was conducted in SPSS first in order to test whether or not the independent variable had caused a significant overall effect across different task conditions.

Step 2: If an overall effect was confirmed, a *planned* contrast analysis was carried out by translating the hypotheses under investigation into one or several sets of contrasts.

Step 3: Using the contrasts defined in Step 2, each set of hypothesised contrasts were analysed in SPSS (Haans, 2018) under an F test ($F = \frac{MS_{contrast}}{MS_{error}}$), following the procedures defined in Rosenthal et al. (1985)'s book on contrast analysis. When multiple sets of contrasts (e.g. k) against the same set of data were tested, the alpha level for each F test was corrected using the *Bonferroni* method ($\alpha = 0.05/k$).

Step 4: To further investigate the results of the omnibus test in Step 1 and of the F test(s) in Step 3, a *post-hoc* pairwise analysis was used.

Just as in the result figures of Experiment 1, the results of the four task conditions are arranged along the horizontal axis of each figure below. From left to right, the rhythm setting of auditory prompts and targets became increasingly predictable to participants, and the level of initiative participants were taking also increased along the axis.

5.2.1 Sense of control

Participants' ratings for their sense of control in Experiment 2 are analysed in order to test hypothesis $\boldsymbol{H_{MII} - 1}$. The rating data was not normally distributed according to the Shapiro-Wilk test (p<0.05), hence the non-parametric Friedman Test was used to test the overall main effect of rhythm setting across four task conditions. As shown in Table 5.2, the rhythm had caused a significant overall effect (p<0.001) on participants' ratings for their sense of control. This indicates that the manipulation of temporal structures in the four task conditions was also effective in Experiment 2.

Measurement	*N*	χ^2	*df*	*Sig.*
Sense of control	22	43.248	3	**.000***

Table 5.2: Omnibus test for ratings for sense of control (Friedman Test)

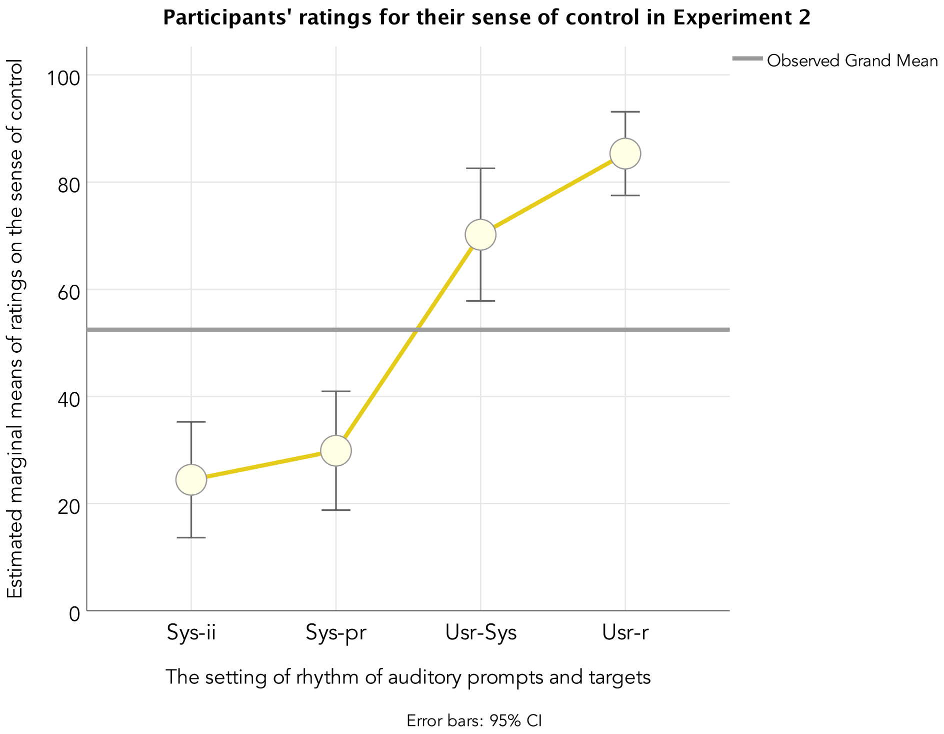
If hypothesis $\boldsymbol{H_{MII} - 1}$ were true, participants would give a higher rating for their sense of control as the rhythm in mixed-initiative interaction became more predictable. In other words, an upward linear trend ("↗") should exist in their rating data following the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. The weights for contrast analysis were assigned accordingly in Table 5.3.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

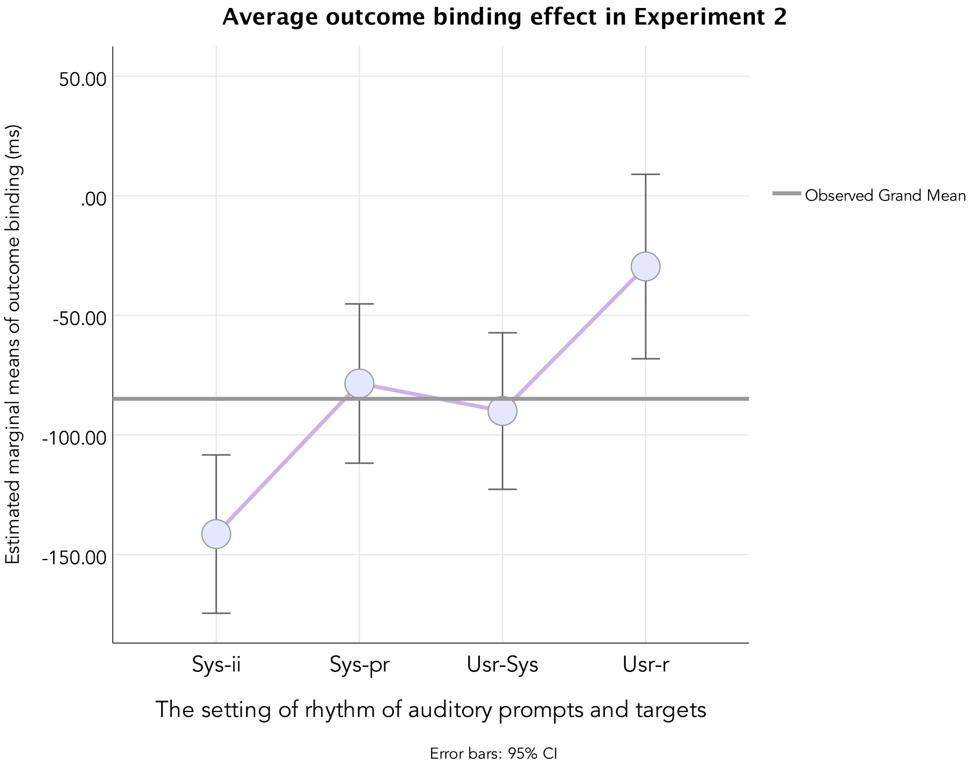
Table 5.3: Weights in contrast analysis for ratings for sense of control (Hypothesis $\boldsymbol{H_{MII} - 1}$: "↗" trend)

The results of contrast analysis in Table 5.4 showed that the hypothesised upward linear trend was significant (p<0.001) in the rating data. As illustrated in Figure 5.5(a), participants reported that they felt increasingly more in control of the pace in the predicted "***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***" direction [2]. The results above further confirm that the manipulation of temporal structures in Experiment 2 had been effective. Meanwhile hypothesis $\boldsymbol{H_{MII} - 1}$ is supported when the timing structures under investigation were presented in the auditory modality in this experiment, just as when the timing structures were presented in the visual modality in Experiment 1.

[2] The slider bar for the "sense of control "rating had the statement *"I was controlling the pace"* on the left, and *"The software was controlling the pace"* on the right, hence the lower the original rating value, the stronger the sense of control participants were reporting, or vice versa. During the analysis, the value of the ratings were calculated inversely, hence in Figure 5.5(a), a higher rating value suggests that participants perceived themselves as more in control, and a lower value as less in control.



(a)



(b)

Figure 5.5: Participants' (a) average rating for sense of control, and their (b) average outcome binding in different tasks in Experiment 2

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↗"	Rhythm	1093146.183	1	1093146.183	59.689	**.000***
	Error (Rhythm)	384595.818	21	18314087		

Table 5.4: Contrast analysis for ratings for sense of control (Hypothesis $\boldsymbol{H_{MII} - 1}$: "↗" trend)

To further investigate the effect of rhythm setting on participants' ratings for their sense of control, pairwise comparisons were carried out. The rating data was not normally distributed, hence the Wilcoxon Signed Ranks Test was adopted. The alpha level was reduced to $0.05/6 = 0.0083$ using the *Bonferroni* method. The results are shown in Table 5.5. Participants reported a stronger sense of control in the ***Usr-r*** condition than ***Usr-Sys***. They also experienced a stronger sense of control in ***Usr-Sys*** than in ***Sys-ii*** and ***Sys-pr***, though not significantly different between the latter two.

Pair	***Sys-ii<*** ***Sys-pr***	***Sys-ii<*** ***Usr-Sys***	***Sys-ii<*** ***Usr-r***	***Sys-pr<*** ***Usr-Sys***	***Sys-pr<*** ***Usr-r***	***Usr-Sys<*** ***Usr-r***
Z	-0.747	-3.458	-4.044	-3.528	-4.108	-3.665
Sig.	.455	**.001***	**.000***	**.000***	**.000***	**.000***

Table 5.5: Pairwise comparisons for sense of control (Wilcoxon Signed Ranks Test)

In short, the results of the omnibus test showed that the rhythm setting caused a significant overall effect across four task conditions, and the results of contrast analysis confirmed that participants' reported sense of control increased in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, thereby supporting hypothesis $\boldsymbol{H_{MII} - 1}$. Follow-up pairwise comparisons further validated this trend, although one pair was not found to be significantly different, hence the findings should be interpreted with caution. The limitations will be discussed in **Section 5.3.2**.

5.2.2 Outcome binding

The outcome binding effect is analysed to further examine the effectiveness of the manipulation of the independent variable in this experiment and to test hypothesis $\boldsymbol{H_{MII} - 1}$. The data passed the Shapiro-Wilk Test of Normality (p>0.05) and did not violate Mauchly's assumption of sphericity ($Mauchly's\ W$=0.774, χ^2=5.049, DoF=5, p=0.410). Hence, ANOVA with repeated measures was carried out in SPSS, in order to examine whether or not the rhythm setting had caused a significant overall effect on participants' outcome binding across different task conditions. As shown in Table 5.6, the overall main effect was very significant (p<0.001).

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
Outcome binding	139152.603	3	46384.201	9.163	**.000***
Error (Rhythm)	318913.188	63	5062.114		

Table 5.6: Omnibus test for outcome binding (ANOVA with repeated measures)

As hypothesised in $\boldsymbol{H_{MII} - 1}$, a more predictable rhythm in mixed-initiative interaction can increase the user's perceived control, and irregular and unpredictable timing intervals can cause the opposite. If $\boldsymbol{H_{MII} - 1}$ were also true in this experiment, the an upward linear trend ("↗") should exist in the data of participants' average outcome binding in different conditions in Experiment 2, because unpredictable intervals in the ***Sys-ii*** condition should cause a loss of agency and result in a greater outcome binding effect (i.e. a more negative value), while more predictable rhythm structures in other conditions could result in a milder outcome binding effect (i.e. a less negative or even positive value). The above prediction based on $\boldsymbol{H_{MII} - 1}$ was then expressed as a set of weight coefficients, as shown in Table 5.7.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 5.7: Weights in contrast analysis for outcome binding (Hypothesis $\boldsymbol{H_{MII} - 1}$: "↗" trend)

The results of contrast analysis are as shown in Table 5.8. The predicted linear trend was found to be very significant (p<0.001). As shown in Figure 5.5(b), the value of the outcome binding effect on participants' time perception did become less negative in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** as predicted in hypothesis $\boldsymbol{H_{MII} - 1}$.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↗"	Rhythm	2311110.291	1	2311110.291	22.446	**.000***
	Error (Rhythm)	2162256.938	21	102964.616		

Table 5.8: Contrast analysis for outcome binding (Hypothesis $\boldsymbol{H_{MII} - 1}$: "↗" trend)

To further look into the effect of rhythm setting on participants' outcome binding, pairwise comparisons were conducted among the four task conditions. The paired samples t test was used here because the data passed the normality test. Again the alpha level was adjusted to 0.05/6 = 0.0083 using the *Bonferroni* method. As shown in Table 5.9, the value in ***Sys-ii*** was significantly more negative than in the other three conditions, among which the outcome binding was different from each other, though not proved to be significant under the *Bonferroni* correction.

Pair	***Sys-ii< Sys-pr***	***Sys-ii< Usr-Sys***	***Sys-ii< Usr-r***	***Sys-pr< Usr-Sys***	***Sys-pr< Usr-r***	***Usr-Sys< Usr-r***
t	-3.021	-3.255	-4.757	0.512	-2.260	-2.574
df	21	21	21	21	21	21
Sig.	**.007***	**.004***	**.000***	.614	.035	.018

Table 5.9: Pairwise comparisons for outcome binding effect (paired samples t test)

It is worth noting that participants' subjective rating for their sense of control analysed in the last section and the outcome binding effect on their time perception measured by reading a Libet clock analysed in this section shared a very similar and significant "↗" trend in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. In order to further look into their relationship, the Spearman's Rank-order Correlation Test was run in SPSS given that the subjective rating data was not normally distributed. A strong positive correlation was found between the two measures, and it was statistically significant (r=0.249, p=0.019), as shown in Table 5.10.

Correlation test	*Correlation coefficient*	*N*	Sig.
Spearman's rho	**0.249***	88	**.019***

Table 5.10: Correlation test between subjective ratings for sense of control and the outcome binding effect (Spearman's Test)

As reviewed in **Section 2.1.3**, the results of explicit measures of subjective reports of sense of control and implicit measure of binding effect are often but not always congruent. Applying Synofzik et al. (2008)'s theory, the subjective rating in this experiment was measuring the judgement of agency (JoA), while the outcome binding effect reflected the feeling of agency (FoA). The FoA happens on a lower order (i.e. primary level) and relies on sensorimotor authorship cues like proprioceptive influences and direct bodily feedforward when the perceived outcome follows their action within a short time window (i.e. less than 200-250ms) (Stetson et al., 2006; Choi & Scholl, 2006), whereas JoA is formed on a higher order (i.e. conceptual level) and relies on contextual authorship cues like goals, beliefs and social cues (Wegner & Sparrow, 2004), and people will readily judge an event as caused by themselves even after several seconds (Shanks et al., 1989; Ebert & Wegner, 2010). Both JoA and FoA contribute to the overall sense of agency (SoA) but how exactly the two interacts will depend on different task contexts and requirements (Ebert & Wegner, 2010).

In this experiment, the temporal structures of the interaction was manipulated as the independent variable and had caused significant effects on both the explicit JoA measure and the implicit FoA measure. In addition, the results of the two types of measures were significantly and positively correlated. Those findings indicate that the temporal structure of an interaction may have been taken as a cue, which can affect both the FoA and the JoA when people form the experience of agency (Wegner & Sparrow, 2004; Moore, Wegner, & Haggard, 2009), hence the manipulation of the temporal structure can cause highly correlated effects on both aspects. This will be discussed further in **Section 5.3.1** later in this chapter.

To sum up, the rhythm setting caused a significant overall effect on participants' outcome binding according to the omnibus test. The results of contrast analysis confirmed that the outcome binding as stronger when the timing was irregular, and weaker when the rhythm was more predictable, as predicted in $\boldsymbol{H_{MII} - 1}$. In other words, when the system was setting the pace of auditory stimuli randomly, participants experienced a much lower level of control compared with when the system was setting a predictable and rhythmic pace or when participants were setting the pace themselves. In addition, predictable and rhythmic signals that were initiated by the system were as effective as those signals that were first initiated by participants then replicated by the system in mitigating the binding effect and preserving participants' experience of agency. However, it should be noted that not all pairs were confirmed to be significantly different during pairwise comparisons, hence the findings above should be interpreted with caution and limited confidence. Furthermore, a significant positive correlation was found between participants' explicit subject report on sense of control and the implicit binding effect on their time perception, indicating that the temporal structure might have operated as an external cue that affected both the judgement of agency and the feeling of agency. The limitations will be discussed in **Section 5.3.2**.

5.2.3 Entrainment and perceived system adaptation and help

Participants' ratings for how adaptive and helpful the system was during each task are analysed to provide hypothesis $\boldsymbol{H_{MII} - 2}$ with further support. The rating data did not pass the Shapiro-Wilk normality test (p<0.05), therefore the overall main effect of rhythm setting across four task conditions was tested with the non-parametric Friedman Test. The results are shown in Table 5.11, and confirm that the rhythm had caused a very significant overall effect on participants' ratings for both their perceived adaptivity and the perceived helpfulness of the system (both p<0.001).

Measurement	*N*	χ^2	*df*	Sig.
Perceived adaptivity	22	18.192	3	**.000***
Perceived helpfulness	22	18.269	3	**.000***

Table 5.11: Omnibus test for perceived adaptivity and helpfulness of the system (Friedman Test)

As reported in **Section 4.2.3**, in Experiment 1, participants exhibited significantly stronger entrainment behaviours in the ***Usr-Sys*** condition when the visual targets mirrored the rhythm participants set earlier. In each round of the ***Usr-Sys*** condition in Experiment 2, the four prompt beeps were triggered by participants' clicking actions, and the succeeding target beeps were generated by the system mirroring the intervals of the prompt beeps, just as in Experiment 1. If hypothesis $\boldsymbol{H_{MII}-2}$ still held true in Experiment 2, participants should have rated the system in ***Usr-Sys*** condition as more adaptive and more helpful than ***Sys-ii*** or ***Usr-r***. The ***Sys-pr*** condition should also have been less challenging than at least the ***Sys-ii*** condition according to hypothesis $\boldsymbol{H_{MII}-2}$. Therefore, a quadratic trend ("↗↘") should exist in the data of the two ratings following the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. For each rating, one set of contrasts were assigned for contrast analysis, as shown in Table 5.12.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weight for adaptivity rating, "↗↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 1$	$\lambda_4 = -1$
Contrast weight for helpfulness rating, "↗↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 1$	$\lambda_4 = -1$

Table 5.12: Weights in contrast analysis for perceived adaptivity and helpfulness of the system (Hypothesis $\boldsymbol{H_{MII}-2}$: "↗↘" trend)

The results of contrast analysis on participants' rating for their perceived system adaptivity and perceived system helpfulness are jointly presented in Table 5.12. A significant quadratic trend was confirmed in both the adaptivity rating (p=0.036) and the helpfulness rating (p=0.003). This means that participants did rate the system as "*adapted to me*" and "*intended to help me*" in the ***Usr-Sys*** condition, while rating more towards "*I adapted to the system*" and "*the system intended to challenge me*" in ***Sys-ii***, as shown in Figure 5.6 [3].

The results above suggest that participants were able to detect that in the ***Usr-Sys*** condition, during the *Attention* phase of each round, the system was mirroring the pace they set during the *Prompt* phase just before the targets, and they appreciated the system's alignment to their pace and perceived that as being adaptive and helpful.

[3] Again, the slider bar for the rating for adaptivity stated *"The software adapted to me"* on its left end and *"I adapted to the software"* on the right. Hence during the analysis, the value of the ratings were calculated in an inverse manner, hence in Figure 5.6, a higher the rating value indicates that participants perceived the system as more adaptive, and a lower value as less adaptive. Similarly, the slider bar for the rating for helpfulness said *"The software intended to help me"* on the left and *"The software intended to challenge me"* on the right. As a result of inverse calculation, a higher value in Figure 5.6 indicates that participants perceived the system as more helpful, and a lower value as more challenging.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
Adaptivity						
"↗↘"	Rhythm	11546.182	1	11546.182	4.996	**.036***
	Error (Rhythm)	48535.818	21	2311.229		
Helpfulness						
"↗↘"	Rhythm	15290.909	1	15290.909	11.572	**.003***
	Error (Rhythm)	27749.091	21	1321.385		

Table 5.13: Contrast analysis for perceived system adaptivity and helpfulness (Hypothesis $\boldsymbol{H_{MII} - 2}$: "↗↘" trend)

In order to further examine the effect of rhythm on participants' ratings for the system's adaptivity and helpfulness, pairwise comparisons were carried out among the four conditions. The data failed the normality test, hence the Wilcoxon Signed Ranks Test was adopted, with the alpha level set as $0.05/6 = 0.0083$ under the *Bonferroni* correction. The results are presented in Table 5.14. Participants reported that the system was significantly more adaptive in the ***Usr-Sys*** condition than other conditions, and the system challenged them significantly more in the ***Sys-ii*** condition than in other conditions. Other pairs were found to be different but not significantly so after the *Bonferroni* correction.

Pair	***Sys-ii<*** ***Sys-pr***	***Sys-ii<*** ***Usr-Sys***	***Sys-ii<*** ***Usr-r***	***Sys-pr<*** ***Usr-Sys***	***Sys-pr<*** ***Usr-r***	***Usr-Sys<*** ***Usr-r***
Adaptivity						
Z	-0.574	-3.129	-2.576	-3.529	-2.096	-2.334
Sig.	.566	**.002***	.010	**.000***	.036	.020
Helpfulness						
Z	-3.527	-3.463	-2.638	-0.947	-0.472	-1.791
Sig.	**.000***	**.001***	**.008***	.344	.637	.073

Table 5.14: Pairwise comparisons for the perceived adaptivity and helpfulness of the system (Wilcoxon Signed Ranks Test)

In short, the results of two omnibus tests showed that rhythm had a significant overall effect on how participants perceived the system - being adaptive or not, and

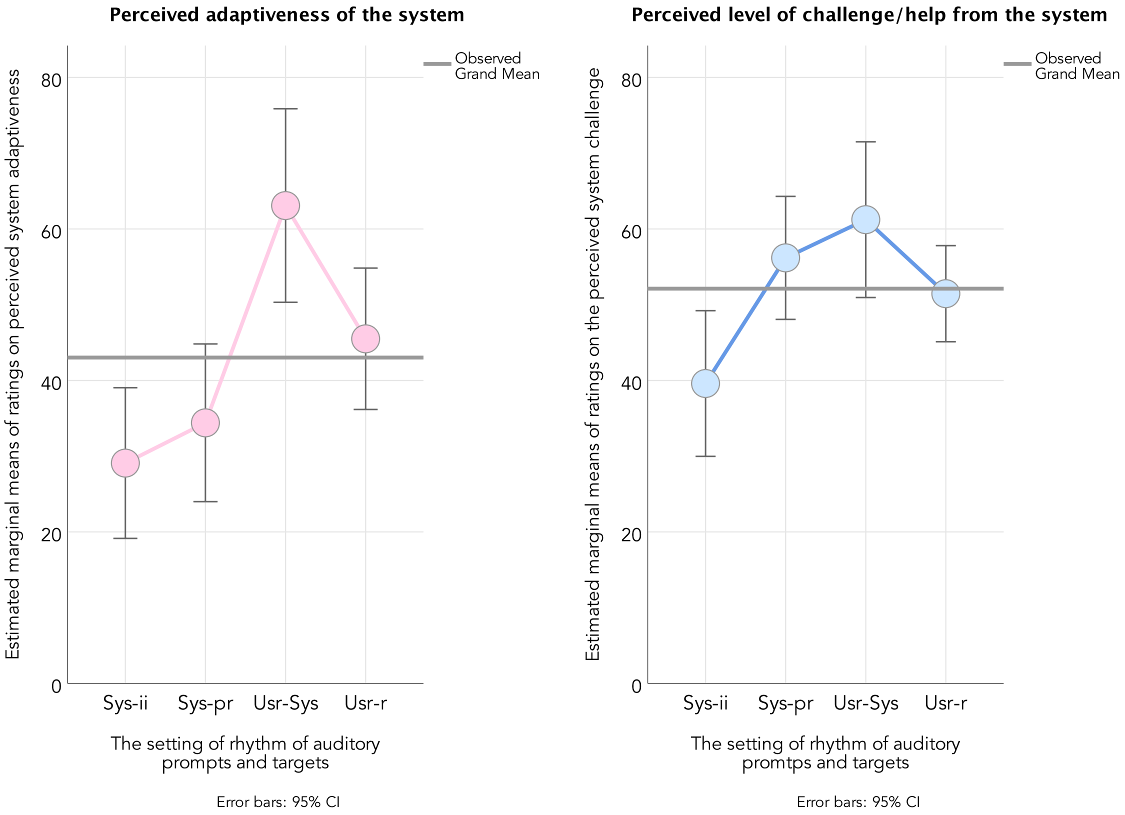


Figure 5.6: Participants' ratings for the perceived adaptivity and helpfulness of the system in different tasks in Experiment 2

being challenging or helpful. The results of contrast analysis confirmed that a more predictable rhythm (e.g. ***Sys-pr*** and ***Usr-Sys***) led the participants to perceive the system as more adaptive and helpful, whereas irregular timing (e.g. ***Sys-ii***) was perceived as challenging and unadaptive. The results above provide hypothesis $\boldsymbol{H_{MII} - 2}$ with strengthened support. However, not all condition pairs were proved to be significantly different in pairwise comparisons with *Bonferroni* correction, therefore the findings should be accepted with caution and limited confidence. The limitations will be discussed in **Section 5.3.2**.

5.2.4 Perceived stress and success

Hypotheses $\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$ are tested by analysing participants' rating for how relaxed/stressed they felt and how much effort they thought they had devoted during different tasks. The data was not normally distributed according to the Shapiro-Wilk test (p<0.05), therefore the non-parametric Friedman Test was used to test the overall main effect of the rhythm setting across four conditions. The rhythm caused a significant overall effect (p=0.008) on participants' ratings for their sense of relaxation/stress, as shown in Table 5.15.

Measurement	*N*	χ^2	*df*	*Sig.*
Sense of relaxation	22	11.816	3	**.008***

Table 5.15: Omnibus test for sense of relaxation (Friedman Test)

Hypothesis $\boldsymbol{H_{MII}} - \mathbf{3.2}$ predicts that a predictable rhythm in mixed-initiative interaction would reduce the user's perceived level of stress, while irregular timing would cause the opposite effect. If this were true, participants in Experiment 2 would report a higher sense of relaxation in tasks where the rhythm setting was more predictable, but a higher sense of stress in tasks where the intervals were more unpredictable and irregular. Therefore an upward linear trend ("↗") should exist in participants' rating for how relaxed/stressed they were during the tasks[4], in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. The linear weights for contrast analysis were assigned accordingly in Table 5.16.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 5.16: Weights in contrast analysis for ratings for sense of relaxation (Hypothesis $\boldsymbol{H_{MII}} - \mathbf{3.2}$: "↗" trend)

As shown in Table 5.17, the predicted upward linear trend in the data was proved to be very significant (p=0.004). As shown in Figure 5.7, participants reported that their level of relaxation increased as the rhythm setting of auditory signals became more predictable and under their own control across four task conditions in the direction of

[4] The slider bar for the rating for relaxation was labelled *"I felt relaxed during this task"* on the left and *"I felt stressed during this task"* on the right, and the confidence bar was labelled *"I felt confident in my answers"* and *"I felt unconfident in my answers"* on the left and right respectively. These ratings were calculated inversely, and a higher value in Figure 5.7 suggests either more relaxation or more confidence.

Sys-ii → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** as expected. Hence hypothesis $\boldsymbol{H_{MII} - 3.2}$ is supported by Experiment 2.

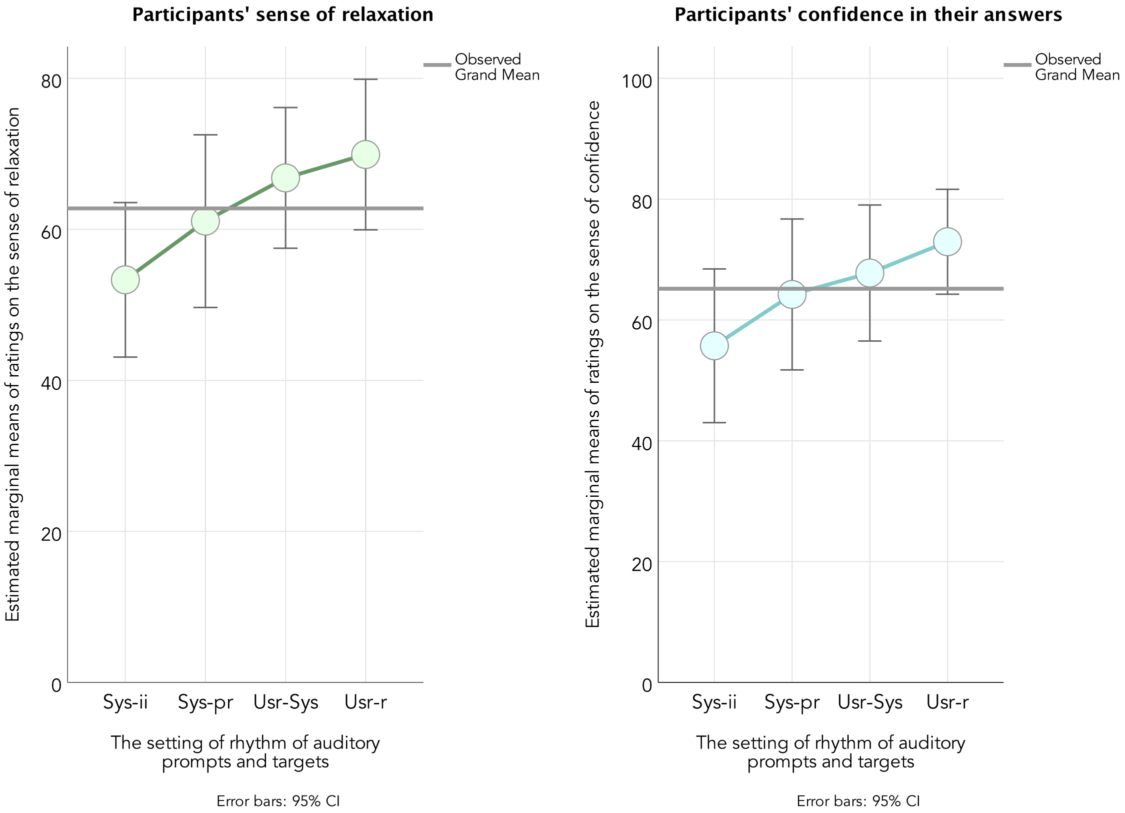


Figure 5.7: Participants' ratings for the sense of relaxation and confidence in different tasks in Experiment 2

Just as in Experiment 1, hypothesis $\boldsymbol{H_{MII} - 4.2}$ is tested by analysing participants' rating for how much confidence they had in their own answers during different tasks in Experiment 2. The rating data did not pass the Shapiro-Wilk normality test (p<0.05), hence the non-parametric Friedman Test was used to test the overall main effect of rhythm setting across conditions. As shown in Table 5.18, the rhythm caused a very significant overall effect (p<0.001) on participants' ratings for their sense of confidence during each task.

As hypothesised in $\boldsymbol{H_{MII} - 4.2}$, a more predictable rhythm in mixed-initiative interaction would make the user feel more confident in their own performance, while irregular interaction intervals would do the opposite. The hypotheses were supported in Experiment 1 in which the interaction was taking place in the visual modality.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"$\nearrow$"	Rhythm	67765.500	1	67765.500	10.141	**.004***
	Error (Rhythm)	140323.500	21	6682.071		

Table 5.17: Contrast analysis for ratings for sense of relaxation (Hypothesis $\boldsymbol{H_{MII}} - \mathbf{3.2}$: "$\nearrow$" trend)

Measurement	*N*	χ^2	*df*	*Sig.*
Sense of confidence	22	18.722	3	**.000***

Table 5.18: Omnibus test for participants' confidence in their answers (Friedman Test)

If $\boldsymbol{H_{MII}} - \mathbf{4.2}$ were still true in Experiment 2 in which the interaction was in the auditory modality, then an upward linear trend ("$\nearrow$") should be expected to exist in the rating data, in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. The linear weights for contrast analysis were assigned accordingly in Table 5.19.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "$\nearrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 5.19: Weights in contrast analysis for participants' confidence in their answers (Hypothesis $\boldsymbol{H_{MII}} - \mathbf{4.2}$: "$\nearrow$" trend)

The predicted linear trend was confirmed to be very significant (p=0.002) in participants' rating for how confident they were in their answers. As pictured in Figure 5.7, across the four task conditions, participants reported a higher level of confidence in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, where the rhythm of the mixed-initiative interaction became increasingly more predictable. Therefore $\boldsymbol{H_{MII}} - \mathbf{4.2}$ is also supported by Experiment 2.

Pairwise comparisons were carried out to further investigate the effect of rhythm on participants' reported level of relaxation and confidence. The Wilcoxon Signed Ranks Test was adopted given that the data did not pass the normality test, and again the alpha level was reduced to 0.05/6 = 0.0083 using the *Bonferroni* correction. The results are shown in Table 5.21, participants reported that they felt significantly more

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
“↗”	Rhythm	67101.136	1	67101.136	11.871	**.002***
	Error (Rhythm)	118699.864	21	5652.374		

Table 5.20: Contrast analysis for participants' confidence in their answers (Hypothesis **$H_{MII} - 4.2$**: “↗” trend)

Pair	***Sys-ii*<** ***Sys-pr***	***Sys-ii*<** ***Usr-Sys***	***Sys-ii*<** ***Usr-r***	***Sys-pr*<** ***Usr-Sys***	***Sys-pr*<** ***Usr-r***	***Usr-Sys*<** ***Usr-r***
Relaxation						
Z	-1.895	-2.781	-2.820	-1.250	-1.756	-1.457
Sig.	.058	**.005***	**.005***	.211	.079	.145
Confidence						
Z	-2.539	-2.550	-3.297	-0.635	-1.582	-1.612
Sig.	.011	.011	**.001***	.525	.114	.107

Table 5.21: Pairwise comparisons for relaxation and confidence (Wilcoxon Signed Ranks Test)

stressed (less relaxed) in the ***Sys-ii*** condition than ***Usr-Sys*** and ***Usr-r***, and had much lower confidence in their answers in ***Sys-ii*** than in ***Usr-r***. Difference was found in other pairs like ***Sys-ii*** vs. ***Sys-pr*** and ***Sys-ii*** vs. ***Usr-Sys***, but not recognised as significant after the *Bonferroni* correction.

Further evidence for hypotheses **$H_{MII} - 3.1$**, **$H_{MII} - 3.2$**, and **$H_{MII} - 4.2$** was found from participants' ratings for the TLX sub-scales.

As with Experiment 1, an omnibus test was carried out on participants' rating on each of the TLX sub-scales to reveal the main overall effect of rhythm across the four task conditions in Experiment 2. Because the data was not normally distributed, the non-parametric Friedman Test was employed. The results in Table 5.22 showed that the rhythm setting had caused significant effect across four conditions, especially in terms of participants' perceived mental demand of a task (p=0.021), successfulness in task performance [5] (p=0.005) and the amount of effort devoted (p=0.001).

[5] According to the original TLX questionnaire design, the scale for task successfulness was labelled *“Perfect”* on its left end, and *“Failure”* on its right end, hence the lower the original value of the

Measurement	*N*	χ^2	*df*	*Sig.*
TLX mental demand	22	9.690	3	**.021***
TLX physical demand	22	5.275	3	.153
TLX temporal demand	22	5.833	3	.120
TLX success	22	12.672	3	**.005***
TLX effort	22	15.426	3	**.001***
TLX frustration	22	5.524	3	.137

Table 5.22: Omnibus test for TLX sub-scales (Friedman Test)

According to $\boldsymbol{H_{MII}-3.1}$ and $\boldsymbol{H_{MII}-3.2}$, a user is more likely to perceive a lower level of stress and effort if the rhythm in the mixed-initiative interaction is more predictable, and to perceive it as more stressful and effort demanding if the interaction intervals are irregular and unpredictable. If they were true, a downward linear trend ("$\searrow$") should be expected in the data of participants' ratings on both the "mental demand" sub-scale and the "effort" sub-scale of TLX. Similarly, an upward linear trend ("$\nearrow$") should exist in the ratings on the TLX "success" sub-scale. In other words, it was predicted that participants' ratings on "mental demand" and "effort" would decrease, while the ratings on "success" would increase, following the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. Three sets of weights for contrast analysis were assigned accordingly in Table 5.23.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for TLX mental demand, "$\searrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 0$	$\lambda_3 = -1$	$\lambda_4 = -2$
Contrast weights for TLX effort, "$\searrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 0$	$\lambda_3 = -1$	$\lambda_4 = -2$
Contrast weights for TLX success, "$\nearrow$" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = 0$	$\lambda_3 = 1$	$\lambda_4 = 2$

Table 5.23: Weights in contrast analysis for ratings on TLX mental demand, effort and success (Hypotheses $\boldsymbol{H_{MII}-3.1}$, $\boldsymbol{H_{MII}-3.2}$, and $\boldsymbol{H_{MII}-4.2}$,: "$\searrow$" or "$\nearrow$" trend)

rating, the more successful participants perceived their task performance to be, or vice versa. During the analysis, the rating value on this item was calculated inversely, hence in Figure 5.8, a higher the rating value indicates that participants perceived their task performance as more perfect, and a lower value as a failure.

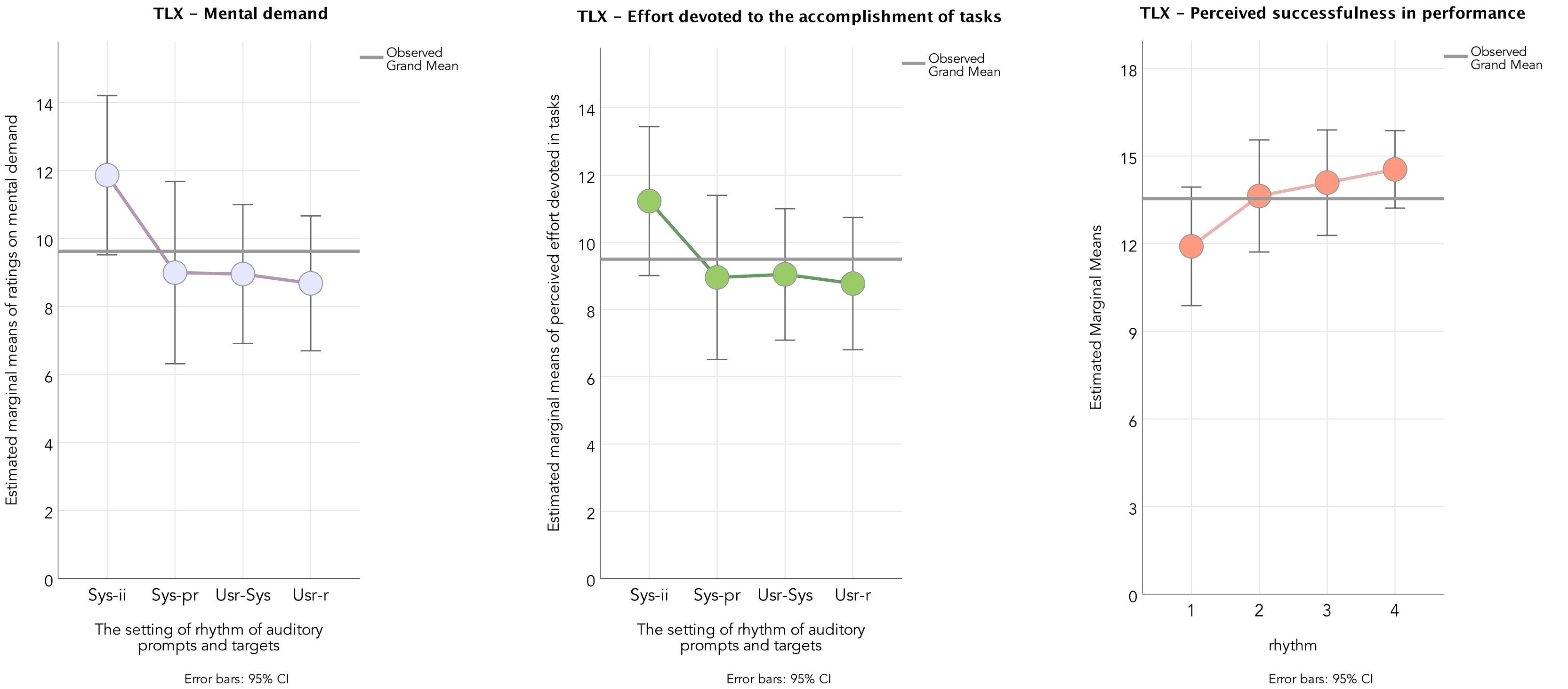


Figure 5.8: Participants' ratings for the TLX scale (mental demand, perceived effort devoted to the task, and perceived success/failure in task performance) in different tasks in Experiment 2

As shown in Table 5.24 and Table 5.25, the predicted downward linear trend ("↘") was confirmed to be significant in the rating data both on the "mental demand" sub-scale (p=0.001) and the "effort" sub-scale (p=0.001). As shown in Figure 5.8, participants rated the tasks as less mentally demanding when the rhythm of the mixed-initiative interaction became more predictable across the conditions in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, they also reported to have devoted less amount of effort when the rhythm was more predictable. Hypotheses $\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$ are therefore further supported in Experiment 2.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↘"	Rhythm	1891.636	1	1891.636	13.927	**.001***
	Error (Rhythm)	2852.364	21	135.827		

Table 5.24: Contrast analysis for ratings on TLX mental demand sub-scale (Hypothesis $\boldsymbol{H_{MII} - 3.2}$: "↘" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↘"	Rhythm	1106.182	1	1106.182	13.812	**.001***
	Error (Rhythm)	1681.818	21	80.087		

Table 5.25: Contrast analysis for ratings on TLX effort sub-scale (Hypothesis $\boldsymbol{H_{MII} - 3.1}$: "↘" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig.*
"↗"	Rhythm	1222.545	1	1222.545	13.516	**.001***
	Error (Rhythm)	1899.455	21	90.450		

Table 5.26: Contrast analysis for ratings on TLX success sub-scale (Hypothesis $\boldsymbol{H_{MII} - 4.2}$: "↗" trend)

A significant upward linear trend was found in participants' rating on the TLX "success" sub-scale in Experiment 1 in **Section 4.2.2**, hence a similar upward linear trend ("↗") should be expected to exist in the rating data in Experiment 2 if

$\boldsymbol{H_{MII}} - \mathbf{4.2}$ were still true. As presented in Table 5.26, the predicted upward linear trend was confirmed to be significant (p=0.001) in participants' rating on the TLX "success" sub-scale in Experiment 2. This means that participants perceived their task performance as increasingly more perfect as the rhythm of the mixed-initiative interaction became increasingly more predictable in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, as pictured in Figure 5.8. Therefore $\boldsymbol{H_{MII}} - \mathbf{4.2}$ is further supported by Experiment 2, meaning that the hypotheses had held true in both the visual and the auditory modalities.

Following contrast analysis, pairwise comparisons were conducted on the ratings on the TLX "mental demand", "success" and "effort" sub-scales. Because the data failed the normality test, the Wilcoxon Signed Ranks Test was used. Again the alpha level was adjusted to 0.05/6 = 0.0083 using the *Bonferroni* correction. The results are shown in Table 5.27. Participants reported that they experienced the highest mental demand in the ***Sys-ii*** condition, in which they also reported that they had devoted the most effort but perceived their performance as the poorest, compared with the other three conditions. The ratings in the other three conditions were not found significantly different under pairwise comparisons. Those test results indicate that when the system took the initiative in a temporally unpredictable and irregular manner, the task would become more mentally demanding, and participants would find themselves performing less successfully despite devoting more effort to the task. When the interaction actions happened in a relatively more predictable and rhythmic flow, even if those were all initiated by the system, the perceived task load might have been comparable to that when participants were taking the initiative and controlling the pace themselves.

In summary, all omnibus tests confirmed that the rhythm setting had caused a significant overall effect on participants' subjective ratings, including their sense of relaxation, confidence in their performance, perceived mental demand during the tasks, and how much effort they had devoted. The results of contrast analysis confirmed that as the participants took more initiative in setting the rhythm, their sense of relaxation and confidence significantly increased, while the mental demand and required effort of the tasks significantly decreased. Therefore hypotheses $\boldsymbol{H_{MII}} - \mathbf{3.1}$, $\boldsymbol{H_{MII}} - \mathbf{3.2}$, and $\boldsymbol{H_{MII}} - \mathbf{4.2}$ are supported by the results in Experiment 2. However, it is important to note that the results of pairwise comparisons after *Bonferroni* correction were not as significant as those of the omnibus tests and the planned contrast analysis, hence the findings should be accepted and interpreted with due caution. The limitations will

Pair	***Sys-ii<***	***Sys-ii<***	***Sys-ii<***	***Sys-pr<***	***Sys-pr<***	***Usr-Sys<***
	Sys-pr	***Usr-Sys***	***Usr-r***	***Usr-Sys***	***Usr-r***	***Usr-r***
TLX mental demand						
Z	-3.281	-2.774	-2.918	-0.024	-0.590	-0.364
Sig.	**.001***	**.006***	**.004***	.981	.555	.716
TLX sucess						
Z	-3.171	-2.892	-2.990	-0.387	-1.053	-0.699
Sig.	**.002***	**.004***	**.003***	.698	.292	.485
TLX effort						
Z	-3.348	-2.957	-2.926	-0.324	-0.285	-0.413
Sig.	**.001***	**.003***	**.003***	.746	.775	.680

Table 5.27: Pairwise comparisons for TLX sub-scales (Wilcoxon Signed Ranks Test)

be discussed in **Section 5.3.2**.

5.3 Further analysis and discussion

5.3.1 Design implications

Noticeably, when comparing participants' subjective ratings for their sense of control, perceived level of stress and how confident they felt about their performance across different tasks in both Experiment 1 and Experiment 2, the results shared a very similar "↗" trend in Figure 4.9 vs. Figure 5.5 (a), and a similar "↘" trend in Figure 4.10 vs. Figure 5.8 in the last chapter and in this chapter. The statistical significance of the effect of rhythm setting on participants' subjective report listed above was also consistent in two experiments, as confirmed by the results of contrast analysis in **Section 4.2.1** vs. **Section 5.2.1**, and **Section 4.2.2** vs. **Section 5.2.4**, which support hypotheses $\boldsymbol{H_{MII} - 1}$, $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 3.2}$ and $\boldsymbol{H_{MII} - 4.2}$ in a consistent manner. Therefore, the results in both experiments can not only confirm the causal relationship between the setting of rhythm during mixed-initiative interaction and the user's experience of agency, but also provide us with the following design implication:

> ***Design implication* 2.1**: During mixed-initiative interaction, when the information is presented in either visual or auditory modality, the effects of timing on the user's sense of control, their perceived level of stress and effort, and their confidence in task performance are congruent.

As reported in **Section 5.2.3**, in the data of participants' subjective ratings on how adaptive and how helpful they perceived the system to be in different task conditions, a significant "↗↘" trend in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** was confirmed by contrast analysis on both ratings. Particularly, participants perceived the system as being more adaptive to their pace in the ***Usr-Sys*** condition, in which the system would trigger beeps in the *Attention* phase following the same intervals as those between the four beeps triggered by participants in the *Prompt* phase in every round. They also perceived the system as being more helpful in the ***Usr-Sys*** condition than, for example, when the system took the initiative irregularly in ***Sys-ii***. The results above indicate that participants did notice that the system was emulating their rhythm in the ***Usr-Sys*** condition and interpreted it as a favourable gesture. Hence the design implication is:

> ***Design implication* 2.2**: During mixed-initiative interaction, the user is capable of recognising the temporal alignment when the system initiates events following their rhythm. The user also appreciates the system's temporal alignment as being helpful and adaptive.

According to the contrast analysis in **Sections 5.2.1** and **5.2.2**, a significant "↗" trend was confirmed to exist in both the subjective rating data and the outcome binding data, in which the reported sense of control was the lowest and the binding effect was the strongest in the ***Sys-ii*** condition when the system presented auditory stimuli at random intervals, but the reported sense of control became stronger and the binding effect became milder as participants had more control over the pace of those beeps, just as hypothesis $\boldsymbol{H_{MII} - 1}$ predicts. The two measures also exhibited a significant and positive correlation (see Table 5.10, r=0.249, p=0.019). The findings above suggest that the temporal structure of an interaction can operate as an external cue that affects both the judgement of agency and the feeling of agency, as discussed in **Section 5.2.2**.

Interestingly, the binding effect in the ***Sys-pr*** condition was in-between: weaker than that in the ***Sys-ii*** condition, but much stronger than that in ***Usr-r***, but not significantly different from that in ***Usr-Sys*** (see Table 5.9, p=0.614). Noticeably, as shown in **Section 5.2.1**, participants explicitly reported a stronger sense of control in the ***Usr-Sys*** condition than in the ***Sys-pr*** condition (see Table 5.5, p<0.001), despite the fact that the outcome binding effect in both conditions was not significantly different.

Similar phenomena were also observed in previous studies (Ebert & Wegner, 2010), that when measuring the sense of agency explicitly using self report (i.e. judgement of agency) or implicitly based on the intentional binding effect (i.e. feeling of agency), there can be incongruity between the two measures, due to the "partially dissociable mechanisms" that underlie the two aspects of the sense of agency (Synofzik et al., 2008; Moore & Obhi, 2012), as reviewed in **Section 2.1.3** and discussed in **Section 5.2.2**. In this experiment, while the results of two measures were significantly and positively correlated overall, local incongruity did emerge as analysed above. This has provided evidence that when participants were neither taking the initiative nor in control of the interaction pace, predictable and rhythmic intervals had mitigated the intentional binding effect. It also echoes the discussion in **Section 4.3.1**, that the user is happy with taking up more manual control during mixed-initiative interaction, compared with simply letting the system set the pace all along. Therefore the third design implication drawn from this experiment is:

> ***Design implication* 2.3**: The temporal structure of mixed-initiative interaction can operate as an external cue that affects both the judgement of agency and the feeling of agency. The user may experience a more explicit sense of control (i.e. judgement of agency) when the system initiates events following the "local" rhythm set by the user on immediate events compared with when it maintains a "global" rhythm that is static throughout the interaction, though the outcome binding effect on the user's perception of time (i.e. feeling of agency) is comparable either way.

Another interesting observation in this experiment was that the effect of interaction rhythm on the outcome binding effect on participants' temporal perception appeared to be less different across four conditions when there were eight beeps in one round, but became more salient and differentiating when comparing the rounds with seven, nine or ten beeps respectively, hence there may be an interaction between the rhythm setting and the number of beeps. *Post-hoc* analysis was carried out by testing the outcome binding effect across four conditions after grouping the rounds with seven, eight, nine and ten beeps separately. The results of the non-parametric Wilcoxon Signed Ranks pairwise comparison are shown in Table 5.28 and illustrated in Figure 5.9. The alpha level was adjusted to 0.05/6 = 0.0083 using the *Bonferroni* method.

Number of beeps	**Sys-ii** vs. **Sys-pr**	**Sys-ii** vs. **Usr-Sys**	**Sys-ii** vs. **Usr-r**	**Sys-pr** vs. **Usr-Sys**	**Sys-pr** vs. **Usr-r**	**Usr-Sys** vs. **Usr-r**
7	Z=−3.425 p=0.001*	–	Z=−3.782 p<0.001*	Z=−1.964 p=0.050	–	Z=−3.425 p=0.001*
8	–	–	Z=−2.549 p=0.011	–	–	–
9	Z=−2.091 p=0.037	Z=−2.581 p=0.010	Z=−3.230 p=0.001*	–	–	Z=−2.159 p=0.031
10	Z=−2.646 p=0.008*	Z=−3.425 p=0.001*	Z=−2.841 p=0.005*	–	–	–

Table 5.28: Average outcome binding effect given different number of beeps in one around in Experiment 2

In the rounds with eight beeps, no significant difference in the outcome binding effect was observed between any pairs of conditions. In those rounds that had seven, nine or ten beeps, by contrast, such binding effect was statistically different between other pairs of conditions, such as ***Sys-ii*** and ***Sys-pr***, ***Sys-ii*** and ***Usr-Sys***, and ***Usr-Sys*** and ***Usr-r***. It was also observed that the binding effect was not significantly different between the ***Sys-pr*** condition and the ***Usr-r*** condition, no matter how many beeps there were in one round, even though overall significant difference did appear when counting in and comparing all rounds in ***Sys-pr*** and ***Usr-r***.

As reviewed in **Section 2.3.1**, one possible explanation for the interaction effect observed as above may be people's ability to subjectively differentiate identical auditory signals and automatically group them into two, three or more (London, 2012b, 2012a). In this experiment, when participants were listening to an uncertain number of identical beeps, they might have grouped those beeps to make it easier to attend to. In each

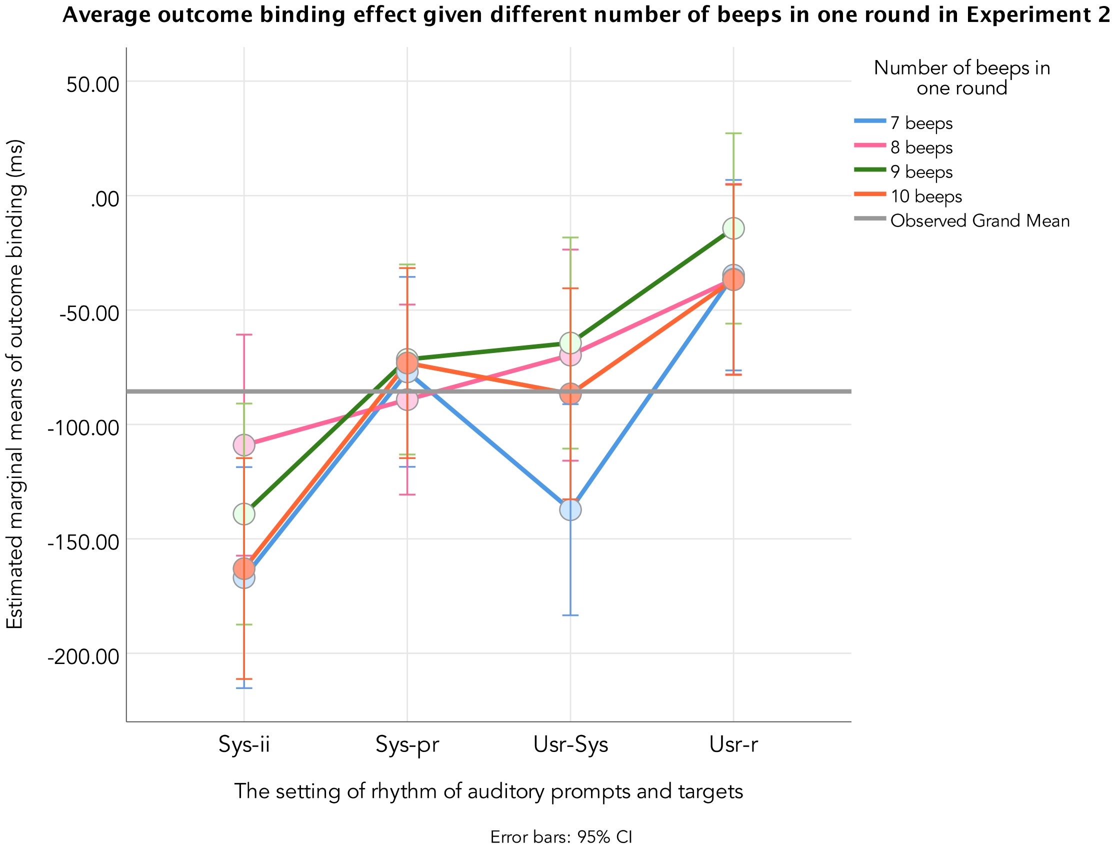


Figure 5.9: Average outcome binding effect on participants' perception of time given different number of beeps in one around in Experiment 2

round, there were four beeps in the *Prompt* phase, hence it is likely that participants attempted to group successive beeps into four. Should anticipatory grouping of this kind exist, it would have been easily and frequently violated, because the number of beeps in the following *Attention* phase was randomised (three, four, five or six). Hence when there happened to be four beeps in the *Attention* phase, participants' expectation would have been confirmed, which might have helped them to make a more accurate estimation of the time of the last beep. In this way, the outcome binding effect might have been mitigated, compared with when there were not exactly four beeps following the first four prompt beeps.

Given that the above analysis was carried out *post-hoc* and not all of the results of pairwise tests met the adjusted alpha level using the *Bonferroni* correction method $(0.05/6 = 0.0083)$, even though existing theories as introduced above can support the findings, it would still be an overstatement to say that they could amount to a design implication. I therefore put forward a prediction that is worth exploring in future studies:

> ***Observation—based prediction* 1**: In mixed-initiative system design, if it is not possible to control the intervals of system-initiated events in a rhythmic manner, we can consider and explore grouping them in twos, threes, fours, or other regular patterns. This has the potential to mitigate the violation of the user's temporal expectation, and alleviate the reduced sense of control that results from the temporal irregularity of individual events.

5.3.2 Limitations

The findings and design implications drawn from this experiment need to be generalised and applied with caution because of the following aspects of limitations:

- First, because Experiment 2 was a controlled experiment, the auditory stimuli that participants were attending to were identical beeps (frequency: 3600Hz, duration: 50ms). However, in real mixed-initiative interaction that employs the auditory modality, the length of auditory signals would often be longer than 50ms, and the amount of information carried by the signals and their acoustic or semantic complexity will be higher than individual beeps. For instance, when the user is using speech interaction, simply annunciating a word command like "Go" will take approximately 300ms (Limerick et al., 2015). Issuing commands to the system and receiving its response in natural language can greatly complicate both the content and its rhythmic structure, and more dynamic rhythm models developed in music and linguistic studies, such as those based on speech prosody, stressed syllables and other acoustic features, should be considered when manipulating the timing of complex auditory signals (Shneiderman, 2000; Inden et al., 2013; Hawkins et al., 2013).
- Second, the average interval length in each participant's experiment session ranged between 550ms and 1200ms, and the shortest interval occurred in a session was 231ms. Hence the manipulation of interaction timing in this experiment was

also in a low granularity like Experiment 1. The beeps were generated using the *Console.Beep(3600,50)* command in C# that can be executed within milliseconds, whereas in real applications, when manipulating the timing of system-initiated auditory signals during mixed-initiative interaction, both the runtime of back-end algorithms and the time it takes to generate and play sounds should be taken into account.

- Third, there were trade-offs in the design of experiment. In order to measure how interaction rhythm can influence participants' perception of when a consequence occurred (i.e. outcome binding), participants were asked to report their perceived time of the last beep in the *Attention* phase in every round by typing in the position of the hand of a Libet clock, which was essentially implementing an N-back paradigm where $N = 1$. This approach constrained the number of events that could be allowed in the *Recall* phase: if participants needed to attend to more (say k) events in the *Recall* phase before typing in their estimate of time when the last beep in the *Attention* phase occurred, then participants were in fact recalling a $(k + 1)$-back event rather than a 1-back event, hence the outcome binding protocol would have been violated. If each round had more phases, such as *Prompt(a)-Attention(a)-Prompt(b)-Attention(b)-Recall*, the calculation of auto-correlation/cross-correlation coefficients could have been viable, but it would have introduced complications such as how many beeps should happen in each phase, whether or not participants' "subjective metricization" (London, 2012a) across phases may mask the effect of outcome binding as discussed in **Section 5.3.1**, as well as the possibility of participants getting confused about what actions they should take in different phrases. Hence participants' entrainment behaviours as observed in Experiment 1 were not studied in this experiment. Given that in interpersonal interaction like a conversation, people tend to align to the rhythm set by the previous speaker (Couper-Kuhlen, 1993; Auer et al., 1999; Clayton et al., 2005), it would be interesting to see whether the user would align to system-initiated auditory events or maintain their own rhythm, how such alignment can affect their experience of control, and whether or not the tendency to entrainment is consistent between auditory and visual stimuli in future studies.

- The fourth aspect is again the ideality of the participants and the laboratory environment. None of the participants had an impaired hearing or an impaired ability to focus. They were also not exposed to any noise or distractions during the experiment. In future studies, it would be worthwhile exploring how human factors and environmental factors, such as whether or not the user has been primed with temporal expectations, loaded with secondary tasks, or exposed to noise, can mediate the effects of rhythm on the user's experience of agency during mixed-initiative interaction.

- The last aspect is the same as the fourth limitation discussed in **Section 4.3.2** in Experiment 1. While the results of contrast analysis in **Sections 5.2.1**, **5.2.2**, **5.2.3** and **5.2.4** confirmed the significance in the experiment data as predicted in $\boldsymbol{H_{MII} - 1}$, $\boldsymbol{H_{MII} - 2}$, $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 3.2}$, and $\boldsymbol{H_{MII} - 4.2}$, the results of independent pairwise tests were not always congruent with those of contrast analysis. Hence it is important to note the caveat that the findings and the resulting design implications in this chapter should also be interpreted with due caution and accepted with limited confidence. Nevertheless, as argued in **Section 4.3.2**, the results of contrast analysis should not be devalued for two reasons. Firstly, pairwise tests with the *Bonferroni* correction may have impaired the power of the statistical tests and led to more Type II errors, or even been "deleterious to statistical inferences" (Perneger, 1998; Nakagawa, 2004), especially when all the tests in Experiment 2 were *planned* to test hypotheses with solid theoretical grounds, and pairwise comparisons were *dependent* to each other, rather than *unplanned* without any predictions to explore potential effects in *independent* observations (Armstrong, 2014). Secondly, using pairwise tests under an omnibus test is not the single best method to test how effective an independent variable is in an experiment (Rosenthal et al., 1985; Furr & Rosenthal, 2003; Abdi & Williams, 2010), because it disregards how different levels of the independent variable are arranged. The independent variable of Experiment 2 was defined and manipulated in the same way as in Experiment 1 (i.e. the rhythmic character of the initiative-taking during the interaction were increasingly more predictable and under the user's control across the four conditions). Hence the results of contrast analysis are more focused and will match the hypotheses under investigation more precisely (Rosenthal et al., 1985).

5.4 Summary

In this chapter, I reported the design and results of Experiment 2 and provided consolidating evidence to the hypotheses in **Chapter 3**. The task design was reported in **Section 5.1**. I developed an experimental system that can manipulate the timing of a series of auditory stimuli. Participants' sense of agency was measured both explicitly and implicitly: they were asked to rate their sense of control on a numeric scale, and they needed to report their perceived time of a system-initiated or self-initiated event using the Libet clock paradigm.

I reported complementary evidence in **Sections 5.2.1** and **5.2.2** of hypothesis $\boldsymbol{H_{MII} - 1}$, that participants reported that they felt a stronger sense of control while the outcome binding on their perception of time was also the lowest when they initiated the auditory stimuli in their own pace or when the system initiated stimuli following the pace they set earlier, compared with when the system took all the initiative and determined the rhythm. The results therefore echo the findings in Experiment 1.

As analysed in **Section 5.2.4**, participants found it particularly stressful and demanding when the system was initiating auditory stimuli in an unpredictable and irregular manner. They also felt the least confident in their estimation of when a target stimuli occurred, despite the fact that they perceived that they devoted the most effort in the task. On the other hand, when participants were attending to auditory stimuli either initiated by themselves or by the system in a predictable manner, they perceived the tasks as less mentally demanding and more relaxing, and they were more confident of their time estimation. The results further support the hypotheses $\boldsymbol{H_{MII} - 3.1}$, $\boldsymbol{H_{MII} - 3.2}$ and $\boldsymbol{H_{MII} - 4.2}$ and the answers to the three research questions given in the last chapter.

Furthermore, I proposed three more design implications and one research prediction combining the findings in Experiments 1 and 2 in **Section 5.3.1**. The first implication is that the effects of timing on the user's sense of agency, perceived stress and confidence are congruent between the visual and auditory modality. The second implication is that when the system initiates events following the temporal pattern of the user's actions, the user can recognise it and would appreciate it as being adaptive and helpful. The third implication suggests that while predictable and rhythmic system-initiated events can shorten the intentional binding effect that is implicitly associated with a stronger agency experience for the user, letting the system emulate the rhythm set by the user themselves can give them a more explicit perception of being in control. Finally, based on the observations of temporal grouping discussed in **Section 5.3.1**, I predicted that when it is not possible to manipulate the timing of individual system-initiated events, grouping them in a regular pattern may mitigate the impairment of the user's sense of agency.

I then discussed five aspects of limitations that need to be considered when generalising or applying the findings and design implications drawn from this experiment in **Section 5.3.2**, including the lack of complexity in auditory signals, the relatively limited range of interaction timescale, the trade-off in the design of experiment, the sampling of participants, the distraction-free environment, and the lack of congruence between a) significant results in omnibus tests and contrast analysis and b) insignificant results in pairwise tests under the *Bonferroni* correction.

CHAPTER 6

Contextualising rhythmic agency in AI-assisted labelling - Experiment 3

6.1 Background

The results of Experiment 1 and Experiment 2 have supported the hypotheses proposed in **Chapter 3**: predictable rhythmic patterns in a mixed-initiative interaction can have a positive influence on participants' experience of agency and entrainment behaviours, and may reduce their cognitive load so that they can achieve better task performance and feel more relaxed. Conversely, arrhythmic intervals can have a negative influence.

However, as discussed in the previous two chapters, the question remains as to whether those findings are generalisable to actual mixed-initiative interaction tasks, which have looser constraints in timing, more complexity in decision making and a higher cognitive demand compared with the simple stimulus-response and shape-position recall tasks in Experiments 1 and 2. The question of whether those insights can be translated into HCI design practice also remains unanswered.

Motivated by the considerations above, Experiment 3 was designed and carried out to examine the findings in Experiments 1 and 2 in a more realistic setting. This experiment serves three purposes. Firstly, it tested hypotheses in **Chapter 3** and provided them with strengthened support. Secondly, it offers detailed insights into how users behave in a realistic mixed-initiative interaction scenario. Thirdly, by giving a concrete showcase, it is intended to attract more attention to and discussion on rhythmic agency among both the HCI and the machine learning community.

6.1.1 Interacting with assisted labelling tools

Labelling lays the foundation for the supervised training of a machine-learning based artificial intelligence (AI) algorithm (Brodley, Rebbapragada, Small, & Wallace, 2012). The primary purpose of labelling is to construct a training dataset that can demonstrate human's subjective behaviours (so called "ground truth"), and based on the user-built classifiers, AI can produce its own classifiers that emulate human intelligence and replicate human judgements (Ware, Frank, Holmes, Hall, & Witten, 2001; Blackwell, 2015). Well-established research resources have been constructed that way. For instance, the ImageNet database offers "millions of cleanly sorted images" to train computer vision and pattern recognition algorithms (Deng et al., 2009), and human experts are recruited to label databases of naturalistic facial expressions and non-verbal behaviours in order to train affective computing systems (Afzal & Robinson, 2014).

However, there are two challenges when commissioning labelling tasks: how to do it economically, and how to guarantee the labels' quality. Manually annotating sample datasets is a tedious, expensive and time-consuming job (Afzal & Robinson, 2014). Experts' knowledge is needed to establish the "ground truth", but experts may be reluctant to spend their precious time on such basic and repetitive tasks (Blackwell, 2015, 2017). Researchers then turn to online crowd-sourcing platforms such as Amazon Mechanical Turk, which specialises in such "Human Intelligence Tasks" (Irani & Silberman, 2013). The quality of the labels can be compromised by various factors too, such as human errors caused by the user's fatigue (Kamalian, Yeh, Zhang, Agogino, & Takagi, 2006; Brodley et al., 2012), the inconsistency in labels caused by the shifting criteria in the user's mental models (so called "concept evolution") (Kulesza, Amershi, Caruana, Fisher, & Charles, 2014), and the trade-off between the efficiency and consistency of the user's label judgements (Sarkar et al., 2016).

The pragmatic considerations above are leading to the design and development of interactive tools that can assist and improve labelling work. For example, existing labelling tools can present the user with cases that are organised under colour coded labels (Blackwell, 2017), collect the user's labels based on their choices in pairwise or setwise comparisons (Sarkar et al., 2016; Bennett, Chickering, & Mityagin, 2009), and allow the user to create and manipulate malleable structures under which they can review and re-organise the labels they have given (Kulesza et al., 2014).

Along with improved design solutions, the role of labelling tools has also evolved. Traditionally, the user would be giving labels case by case to a large and static dataset, which will be used for fully supervised training of machine learning classifiers later. This kind of labelling is performed "offline", where the user exerts a one-way influence on the to-be-trained classifiers through a labelling tool. With the development of interactive machine learning (IML) models (Fails & Olsen Jr, 2003), the user can now perform labelling "online": based on the feedback and recommendations from partially-trained classifiers, the user can see the performance and potential weaknesses of the current statistical model, then perform labelling to correct the model. In other words, the user is in fact interacting with a dynamic statistical model through labelling, hence the labelling tool needs to support a two-way interaction between the user and the classifiers being trained.

In short, desirable labelling tools are expected to provide the user with more decision support than merely presenting cases to label, so the user's labelling can be carried out, as Kulesza et al. (2014) put it, in an "assisted" manner. Hence I will refer to the labelling tools that employ techniques to present and manage labels (e.g. visualisation of label information and structure, pairwise/setwise comparisions) in order to facilitate labelling (e.g. to improve the user's labelling performance and experience) as "assisted labelling" tools. Particularly, if the assisting techniques have incorporated AI components such as machine learning algorithms (e.g. giving the user label predictions and recommendations), I will refer to such tools as "AI-assisted labelling" tools.

6.1.2 Training interactive machine learning (IML) algorithms with assisted labelling

An interface that is implemented to train an IML algorithm can often be characterised as an AI-assisted labelling tool. For instance, the practice of training sometimes takes the shape of giving and reviewing labels on an interactive spreadsheet (e.g. BrainCel, Microsoft Excel FlashFill and CODA), which is a familiar interface for non-expert end-users to access and manipulate data (Chang & Myers, 2014; Sarkar, 2015; Blackwell, 2017). A non-expert end-user here refers to a user who has little knowledge of the statistical models behind IML algorithms, while they still have the domain expertise for which they are recruited to do the labelling. At the same time, a usable IML training interface should be designed to support visual analytics (e.g. highlighting the performance and weakness of current models) (Sarkar, Jamnik, Blackwell, & Spott, 2015; Sarkar, 2015), which may be generated from the IML algorithm that is being trained.

Through AI-assisted labelling, the user can demonstrate desirable behaviours to the IML algorithms and manually correct the wrong behaviours, hence such a training process is a variation of the programming-by-example (PbE) paradigm, and the user is actually "debugging" the system by labelling (Kulesza et al., 2015; Sarkar, 2017). Furthermore, the rapid "train-feedback-correct" cycles (Kulesza et al., 2015) (e.g. less than 5 seconds as suggested by Fails and Olsen Jr (2003)) in which the user is engaged resemble the mixed-initiative characteristics of a conversation or a dialogue (Horvitz, 1999a; Sarkar, 2017). In summary, the training of an IML algorithm using an AI-assisted labelling tool can be performed on an easily controlled and user-familiar interface, the tasks are relatively rapid and often repetitive, and have dialogue-like mixed-initiative characteristics, therefore, it provides a natural context to study the effects of timing on the user's sense of agency during mixed-initiative interaction.

Previous studies raised questions such as *when* to pass control over to the user or to ask the user to provide advice during joint problem solving (Horvitz & Barry, 1995; Horvitz, 1999b; Wolber & Myers, 2001). However, they did not offer clear design guidelines, and the case studies they offered mainly adopted an event-driven stimulus-response approach (i.e. considering which response should be invoked *when* a certain stimulus appears), rather than driven by a rhythmic timing pattern of the kind this PhD dissertation is addressing.

As set out in **Chapter 1**, the user's sense of agency is a crucial design consideration for an interface with intelligent components for three reasons. Firstly, even when the system is automating the user's action, the user still wants to assert "fine-grained control" during the interaction (Kulesza et al., 2015). Secondly, the internal inferred model in an IML algorithm may lead to inscrutable system behaviours that the user cannot effectively link with their control actions (Blackwell, 2015), where failing to build such a link can impair their agency experience. Thirdly, the user should be able claim authorship of interaction outcomes because the program's behaviour is essentially a product of the user's coaching (Blackwell, 2015).

Combining the reasons above, this chapter will contextualise the findings introduced in previous chapters in an AI-assisted labelling setting, as well as investigating whether the hypotheses on timing and users' sense of agency will still hold.

6.1.3 Research questions and hypotheses

In the context of AI-assisted labelling (abbreviated as AIaL in hypotheses), four hypotheses were derived from **Chapter 3** as well as the results from the previous two experiments.

6.1.3.1 Rhythm setting and sense of agency

Since AI-assisted labelling tasks share the mixed-initiative characteristics of human conversation, theories of interpersonal communication offer potential insights. When evaluating the contributions of self and others in a collaborative activity, people typically evaluate the amount as well as the usefulness of the contribution differentially, taking into the account the relative ability of each member to participate and exert influence. This can produce a *power and prestige order* (Fişek et al., 1995). During AI-assisted labelling, the user is encouraged to consider him/herself as a teacher, responsible for training the machine learning algorithm. This perceived power dynamic can result in a negatively-valenced expectation violation when the system appears to assume control, or refuses to adapt to the user (Bonito et al., 1999; Scherer et al., 2004; Sanna & Turley, 1996). This experiment aims to investigate how the user's sense of control is affected when the labelling tool or the user respectively assumes control of the interaction rhythm. The first hypothesis is:
$\boldsymbol{H_{AIaL} - 1}$: During AI-assisted labelling, the rhythm imposed by the system can impair the user's sense of control, while the rhythm set by the user can preserve their sense of control.

6.1.3.2 Rhythm setting and stress level

The results of Experiment 1 and Experiment 2 have shown that when the rhythm was dictated by the system and appeared to be random, the user experienced a higher level of stress and devoted more effort to keep up with the pace, while such stress was significantly reduced when the user took the initiative and set the pace. Because AI-assisted labelling tasks are an instance of mixed-initiative interaction, similar effects can be expected. Again participants' stress level was measured using the NASA Task Load Index subjective ratings system (Hart & Staveland, 1988). The second hypothesis is:

$\boldsymbol{H_{AIaL} - 2}$: During AI-assisted labelling, the rhythm imposed by the system can cause the user to experience a higher level of stress, while the rhythm set by the user can reduce their level of stress.

6.1.3.3 Rhythm setting and accumulated task load

As reviewed in **Section 2.2.3**, the human brain routinely extracts timing patterns from external stimuli. Research in neuropsychology has found that people can better recognise and respond to random stimuli that occur in a regular and rhythmic manner, compared to stimuli that occur at random times (Fujioka et al., 2009; Rohenkohl et al., 2012; Arnal & Giraud, 2012). It has also been found that people are able to make accurate judgements faster, even given less information, if the information is presented rhythmically compared with it being presented randomly (Greatrex, 2018). This may be because when a series of events possesses two dimensions of uncertainty, both in its content and in its temporal attributes, a predictable rhythm can reduce the complexity and uncertainty down to only the content, allowing the user to devote more cognitive resources to dealing with the content and making a decision more promptly. In the context of AI-assisted labelling, every message carries both temporal and semantic uncertainty, hence presenting a series of messages in a more predictable rhythm may facilitate the user's decision making and help them process and label each message faster. Therefore there will be less unlabelled messages accumulated in the series over time, resulting in a lower objective task load. This leads to the third hypothesis:

$\boldsymbol{H_{AIaL} - 3}$: During AI-assisted labelling, the rhythm imposed by the system can increase the accumulation of task load, while the rhythm set by the user can reduce the accumulated task load.

6.1.3.4 Predictable rhythm and entrainment behaviours

Research in social psychology demonstrates that during interpersonal interaction, if people adapt to each other's rhythm via entrainment, they develop mutual trust based on temporal predictability, as well as a sense of "intersubjectivity" (i.e. a sense of *"being together"*) (Schegloff, 1992; Gill, 2012) and relaxation (Hawkins et al., 2013). Studies in rhythmic tapping have shown that people are less likely to synchronise tapping with an unresponsive and non-adaptive computer partner compared with an adaptive human partner (Himberg, 2006), and Experiments 1 and 2 have observed similar entrainment effects in mixed-initiative interaction. Therefore similar phenomena in Experiments 1 and 2 should be seen in AI-assisted labelling tasks. Hence the fourth hypothesis is:
$\boldsymbol{H_{AIaL} - 4}$: During AI-assisted labelling, predictable rhythm imposed by the system is more likely to induce the user's entrainment behaviour, while unpredictable rhythm imposed by the system is less likely to induce entrainment behaviour.

6.2 Method

Similar to Experiments 1 and 2, when designing the tasks in Experiment 3, the following set of characteristics should be considered and embedded in the tasks:

1. As with Experiments 1 and 2, the labelling tasks in Experiment 3 should be repetitive or have repetitive steps, upon which different temporal structures (i.e. rhythmic, random, entrained) can be imposed.

2. The flow of the labelling tasks should resemble the "turn-taking" dynamics in Experiments 1 and 2, with a mix of user-initiated actions and system-initiated actions.

3. The tasks in Experiment 3 should require a reasonable amount of cognitive resources. Participants should not be too cognitively occupied to perceive any control, or too idle to perform differently under different temporal structures. This will allow us to test hypotheses $\boldsymbol{H_{AIaL} - 1}$ and $\boldsymbol{H_{AIaL} - 4}$ validly.

4. As with in Experiments 1 and 2, participants should be experiencing an appropriate level of stress during different tasks in Experiment 3, so that participants' subjective ratings for their stress level will not be too high or too low simultaneously to be compared across conditions when testing $\boldsymbol{H_{AIaL} - 3}$.

5. The messages to be labelled should be written in a clear and simple manner with correct grammar and spelling. This is to minimise the confusions or hesitations caused by semantic ambiguity, so that the timing of participants' labelling actions can better reflect their entrainment behaviours predicted in $\boldsymbol{H_{AIaL} - 2}$ without involving confounding factors.

6. Participants' task performance should be measurable, so that their performance can be compared when testing $\boldsymbol{H_{AIaL} - 4}$.

In the following three subsections, I will report the task design for Experiment 3, and will point out how each characteristic was met by the design.

6.2.1 Design of task scenario

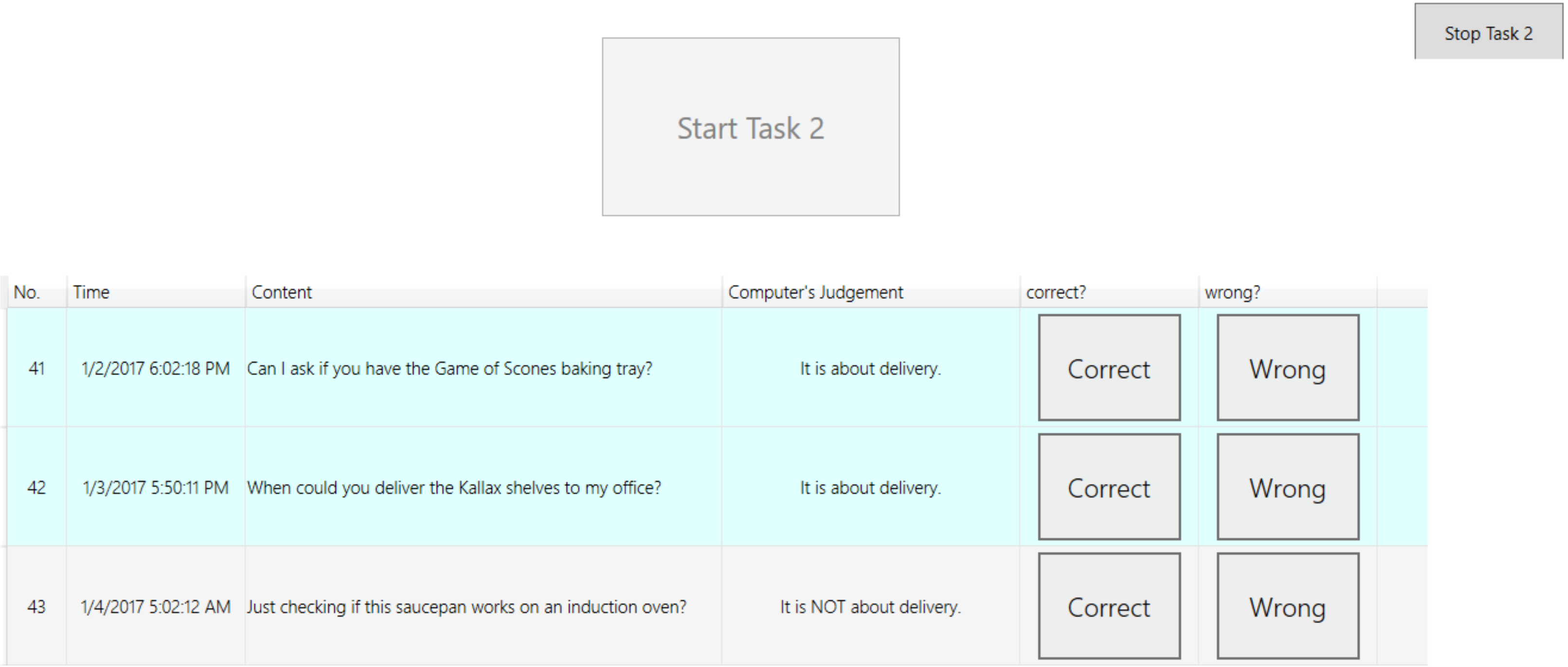
This experiment aimed to study how the user's interaction behaviour and their sense of control would be affected by the rhythmic aspects of interaction with an AI-assisted labelling system. The design of the experiment was inspired and motivated by the CODA system, which is an open source software created to support Africa's Voices Foundation (AVF) researchers to efficiently analyse a large amount of short texts ($>$ 250,000 text messages) in the Somali language and to categorise and review them thematically. When an AVF researcher starts to use CODA, each text message will be presented in a row in a white table, and as they code ("label") the messages one by one, each row will be filled with a colour that corresponds to the colour that is assigned to that category label. As the researcher goes through more messages, the table will be "progressively coloured in" (Blackwell, 2017). As CODA is bootstrapped with manual labels, its artificial intelligence and natural language processing components can offer more decision support. Based on every label decision made by the researcher, CODA can automatically infer the potential label for unlabelled messages and dynamically colour those rows into corresponding colours - though in different shades, of which the deepness corresponds to the level of the statistical confidence of those inferred labels, thus directing the researcher's focus to the rows with the lightest colour and facilitating the review process.

The main design purpose of AI-assisted labelling systems like CODA is to allow human experts to make the most efficient use of their valuable time, so as to "get the greatest benefit from their analytic decisions" (Blackwell, 2015, 2017), but the temporal aspects of the interaction have not been manipulated or evaluated in terms of the user's sense of control. Therefore in this experiment, an AI-assisted labelling interface prototype was developed drawing on the design of the CODA system, which can provide the results with external validity, while the labelling tasks were designed in a controlled manner in order to investigate the effects of timing on the user's agency perception. An imaginary task scenario was set as follows:

> "An online shopping mall has a data centre. Recently they developed a few machine learning algorithms, which can process customers' enquiry messages, and automatically label messages into several categories, such as 'delivery', 'exchange and return', 'membership' and so on.
>
> However the performance of those algorithms are quite poor at the moment, and the system often makes wrong judgements. Therefore they are now recruiting people to manually *train* the algorithms, to make them better.
>
> As one of the first steps, the data centre wants to let the algorithms judge whether an enquiry message is about 'product delivery' or not."

Participants were told that their job was to check the system's judgement, as demonstrated in a screenshot of the experimental system in Figure 6.1, by doing labelling tasks in the following manner:

- "If that message is about 'product delivery' and the system says so too, then you click the 'Correct' button, in this way you can reinforce the correct formula of the system."
- "If that message is about 'product delivery' but the system says it's not, then you click the 'Wrong' button, in this way you can rectify the wrong formula of the system."



No.	Time	Content	Computer's Judgement	correct?	wrong?
41	1/2/2017 6:02:18 PM	Can I ask if you have the Game of Scones baking tray?	It is about delivery.	Correct	Wrong
42	1/3/2017 5:50:11 PM	When could you deliver the Kallax shelves to my office?	It is about delivery.	Correct	Wrong
43	1/4/2017 5:02:12 AM	Just checking if this saucepan works on an induction oven?	It is NOT about delivery.	Correct	Wrong

Figure 6.1: Sample screenshot made during one of the tasks in Experiment 3

- “If that message is NOT about ‘product delivery’ and the system says it isn’t as well, then you click the ‘Correct’ button, in this way you can reinforce the correct formula.”
- “If that message is NOT about ‘product delivery’ but the system says it is, then you click the ‘Wrong’ button, in this way you can rectify the wrong formula.”

Participants were asked to complete four tasks during the formal stage of the experiment, every task involved making labelling judgements on thirty messages, either initiated by participants themselves or by the system depending on the task condition. Therefore the 1^{st} and 2^{nd} characteristics were met.

Furthermore, participants were informed that in this experiment they were only expected to distinguish whether a message was about *“product delivery”* or not, and they were given a definition of product delivery as well as a list of relevant keywords:

- “Any messages regarding how, when, to where the order is shipped to the customer is considered as in the ‘product delivery’ category. (Keywords: deliver, parcel, post, receive, shipping, address, courier, etc.)”
- “Complaints and enquiry about membership, product information, return and exchange, promotion, customisation, and other issues are not in the ‘product delivery’ category.”
- “When you find it hard to tell if it is about ‘product delivery’ or not, don’t stress, just make your best guess, then move on to the next message.”

The design described as above was to meet the 3^{rd} characteristic.

In order to minimise participants’ bias caused by experimental expectation, during both the recruitment and the introduction stage, participants were told that the goal of this experiment was to “study the efficiency and performance of different database algorithms developed for an online shopping mall data centre, which will be trained during their interaction with users in order to achieve better sentence processing and automatic labelling”. The term “timing” or “rhythm” was *not* mentioned in the briefing. The full introduction script of this experiment can be found in the Appendix B.3.

Before starting an experiment session, all participants agreed to sign an informed consent form, as included in Appendix B.1. After every session, participants were given a debriefing, which explained that in addition to their labelling results, this experiment also aimed to study how the timing pattern of the system's actions had influenced their interaction behaviours and subjective experience of agency. Each experiment session lasted for 25-30 minutes, and a £5 Amazon gift voucher was given to each participant as a reward. This experiment was reviewed and approved by the ethics committee of the Department of Computer Science and Technology, University of Cambridge.

6.2.2 Assignment of labels to experimental messages

The experimental system was a Wizard-of-Oz simulation (Dahlbäck, Jönsson, & Ahrenberg, 1993) of an AI-assisted labelling tool like CODA: participants were told that the label for each message was given by an intelligent algorithm behind the system, and the algorithm would be trained by their labelling decisions over time, while all the labels presented to them during the experiment were in fact randomly pre-assigned. Hence every message was designed to possess the following four attributes:

- *Designed truth*: Each message was designed to be unambiguous, and it should be in either the "product delivery" or "not product delivery" category, corresponding to a Boolean variable in the database behind the experimental interface.
- *Initial label*: Before an experiment session, each message was assigned with an "initial label" randomly, which may be in accord with or opposite to the designed truth. Participants were presented with messages together with their "initial label" during the experiment.
- *Expected label*: Participants were expected to confirm the messages whose "initial label" agreed with their designed truth, and to correct the messages whose "initial label" contradicted their designed truth. Consequently, the expected labels should be consistent with the designed truth.
- *User label*: By confirming or correcting the "initial labels", participants gave

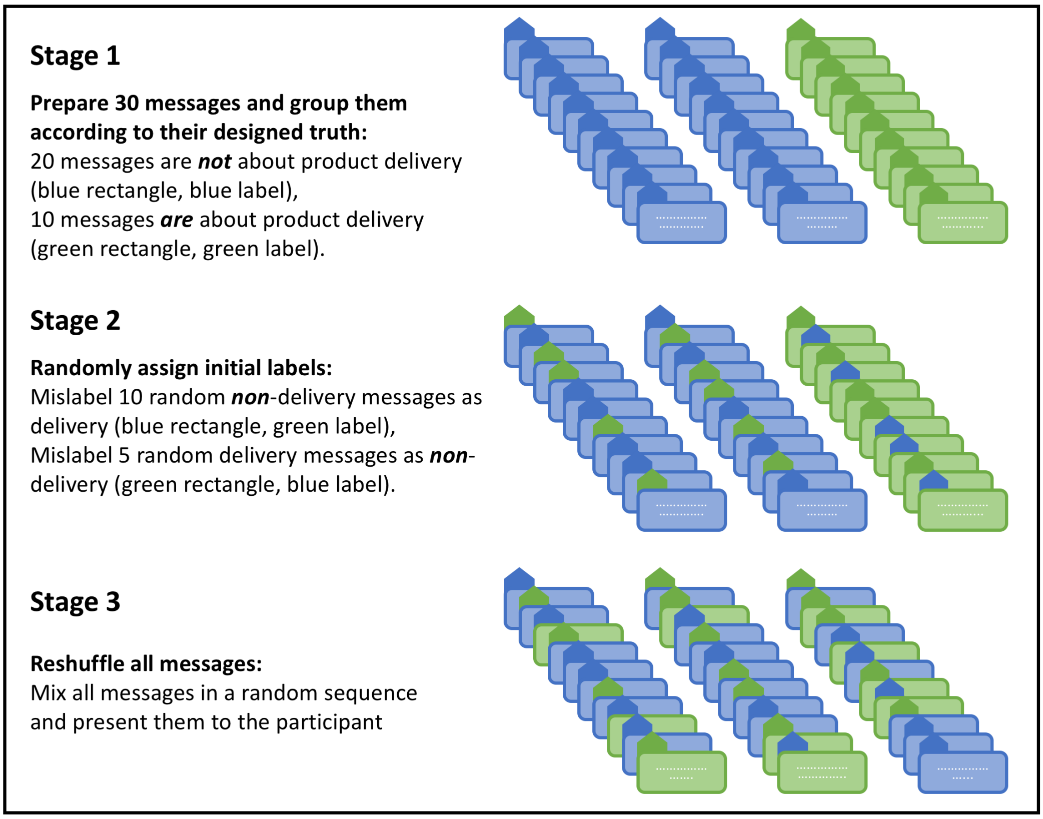
each message a "user label". The "user labels" may or may not be consistent with the expected labels/designed truth.

The user's performance can be measured by counting the number of user labels that are inconsistent with the corresponding expected labels. Hence the 6^{th} characteristic was met.

For the thirty messages in each task, the designed truth of twenty messages was *not* about product delivery, while the remaining ten messages' designed truth *was* about product delivery. Ten out of the twenty non-delivery messages were randomly selected and given an "initial label" as "product delivery", which contradicted their designed truth. Similarly, five out of the ten delivery messages were randomly selected and given an "initial label" as "not product delivery". Therefore, for each task, every participant would see fifteen messages with an "initial label" as "not product delivery", of which five messages were expected to be corrected to their designed truth, and see another fifteen messages be initially labeled as "product delivery", of which ten messages' label needed to be corrected. All 120 short messages, as presented in Appendix B.4, had been proofread by a native English speaker during a pilot study, hence the 5^{th} characteristic was met.

In order to mitigate learning effect, the sequence of all thirty messages within each task was randomised, and the sequence of four tasks was also randomised for every participant, as shown in Appendix B.6. The randomisation of "initial label" assignment and message sequences was done using Microsoft Office Excel for Mac (version 15.32). Each group of thirty messages labelled as their designed truth were organised into a static sequence, in which twenty non-delivery messages followed by ten delivery messages, as shown in Figure 6.2, Stage 1. The randomisation procedures are as follows:

Step 1: Using the $RAND()$ function, a random value was assigned to each of the first 20 non-delivery messages, producing a number between 0 to 1 following the uniform distribution. Then froze the 20 values in Excel so that they would not change as Excel refreshed the spreadsheet after each operation. Then calculated the median (M_1) of the 20 values, guaranteeing that 10 values would be greater than M_1 while the other 10 smaller. Each non-delivery message that bore a random value greater than M_1 was given an "initial label" as "not product delivery",



Stage 1

Prepare 30 messages and group them according to their designed truth:
20 messages are ***not*** about product delivery (blue rectangle, blue label),
10 messages ***are*** about product delivery (green rectangle, green label).

Stage 2

Randomly assign initial labels:
Mislabel 10 random ***non***-delivery messages as delivery (blue rectangle, green label),
Mislabel 5 random delivery messages as ***non***-delivery (green rectangle, blue label).

Stage 3

Reshuffle all messages:
Mix all messages in a random sequence and present them to the participant

Figure 6.2: Illustration of the process of randomly assigning message labels within one task for one participant in Experiment 3

otherwise as “product delivery”, see Figure 6.2, Stage 2.

Step 2: Again, using the $RAND()$ function, a random value was assigned to each of the latter 10 delivery messages. Then froze the 10 values and calculated the median (M_2) of them. Each delivery message that bore a random value greater than M_2 was given an “initial label” of “not product delivery”, otherwise “product delivery”, see Figure 6.2, Stage 2.

Step 3: Finally, using the $RAND()$ function again, a new random value was assigned to each of the 30 messages. Then froze those new values and sorted them in ascending order, see Figure 6.2, Stage 3.

Step 4: Thus 30 fully randomly-labeled-and-sequenced messages were produced for one task. The same procedures were repeated four times to prepare the four tasks in one experiment.

Step 5: Then repeated the whole process above for each session of the experiment.

6.2.3 Independent variable and manipulation

Within-subjects design was adopted in this experiment. In order to test whether the findings in Experiments 1 and 2 can still hold in the context of AI-assisted labelling, the independent variable in this experiment was the same as before, as shown in Table 6.1: the imposition of either predictable rhythmic intervals or randomised arrhythmic intervals. There were three sub-conditions under the rhythmic category, each of which had a different method of initiating an action and setting the pace. As shown in the table, from ***Sys-ii*** to ***Sys-pr***, then to ***Usr-Sys*** and ***Usr-r***, the rhythmic character of the interaction became increasingly predictable and under the user's control just as that in the first two experiments, while the method of timing manipulation had been specifically accommodated to AI-assisted labelling tasks. The rationale of using this set of temporal structures have been stated in **Section 4.1.2**, and manipulating the timing of the interaction under these structures has been proved effective in both Experiments 1 and 2.

Independent variable	*Description of treatment*	*Abbreviation*
Irregular intervals	***Sys***tem takes the initiative at ***i***rregular ***i***ntervals	***Sys-ii***
Predictable rhythm	***Sys***tem takes the initiative in a ***p***redictable ***r***hythm	***Sys-pr***
	Use***r*** takes the initiative, ***Sys***tem aligns	***Usr-Sys***
	Use***r*** takes the initiative in their own ***r***hythm	***Usr-r***

Table 6.1: Independent variable and its settings in Experiment 3 (the same as Experiments 1 and 2)

Before participants started to do formal tasks, they first needed to go through a practice stage with four small tasks, which gave them an overview of all the procedures and a chance to warm up. Each practice task required participants to label ten messages, and each formal tasks required thirty messages. These numbers were determined through pilot sessions. The settings of the interaction rhythm in the four practice tasks were the same as the four formal ones, each of which adopted one of the conditions in Table 6.1, and Figure 6.3 shows the temporal structures of the interaction.

In both Task 1 and Task 2, participants needed to click a "Start Task" button to trigger the task. The system would then start to automatically present ("push") messages together with their "initial label" one at a time at predetermined intervals. Participants needed to judge each "initial label" and give a "user label" by clicking either the "Correct" or the "Wrong" button, and the corresponding row would disappear after either button was clicked. The interval length in Task 2 (***Sys-pr*** condition) was a fixed value of 4.4 seconds. This value was determined based on previous literature, that the optimal line length for screen reading was 50-60 characters per line (cpl) (Dyson & Haselgrove, 2001), and the effective reading rate on screen was around 150 words per minute (Muter & Maurutto, 1991). All of the experiment messages fell into the 50-60 cpl range and the average length was around 11 words, which would take a native English speaker roughly 4.4 seconds to read and comprehend. The considerations above were made to meet the 4^{th} characteristic.

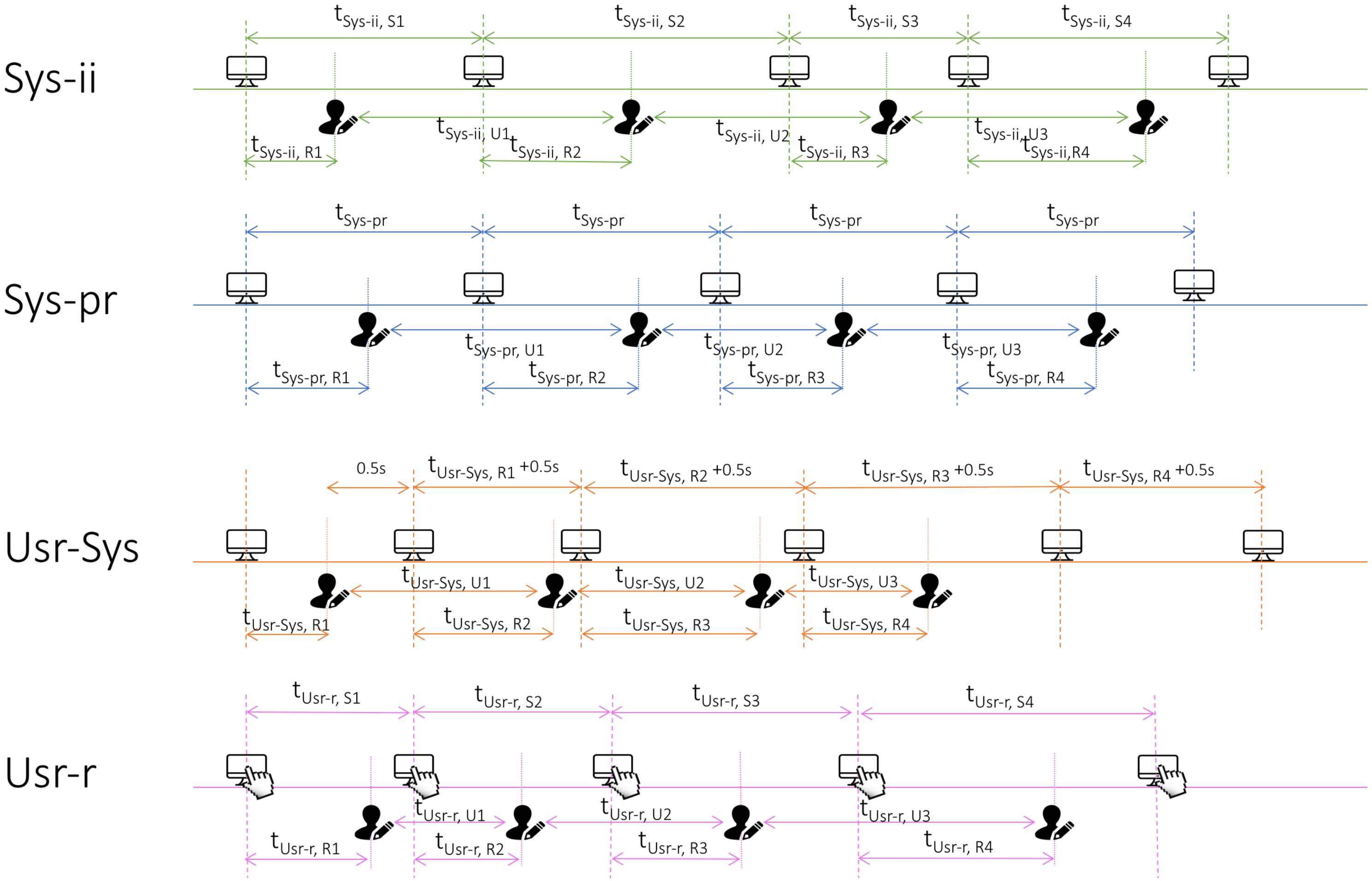


Figure 6.3: Illustration of the temporal structure of each task in Experiment 3

In the two conditions where the system set the pace (***Sys-ii*** and ***Sys-pr***), a small degree of time pressure was applied. This is because it is necessary for participants to see that the system was taking the initiative. Were there no time pressure, with participants labelling each message before another arrived, they might build a false causal link (i.e. "I confirmed/corrected the label of this message, then the system pushed more"). Such a false causal link may introduce a confounding effect on their agency experience. Therefore, the estimated reading time was reduced by 10% to 4 seconds. Previous research was also used to make an estimate of mouse selection time for large on-screen targets, which was roughly 0.4 seconds (Akamatsu & MacKenzie, 1996). Therefore altogether the rhythmic intervals in Task 2 were set as 4.4 seconds.

The random interval series in Task 1 (***Sys-ii*** condition) were generated in the same way as Experiments 1 and 2 using MathWorks MATLAB R2017a. The mean value was set as 4.4 seconds and the interval length ranged equally distributed between 2.2 seconds and 6.6 seconds. Every adjacent two intervals had at least a 0.5 second difference in length, in order to be long enough that participants could notice the variations. All random intervals in Task 1 were generated before the experiment and was imported into the experimental system before Task 1 started. A sample of intervals that were used in this experiment can be found in the Appendix B.5.

As illustrated in Figure 6.3, in Task 3 (***Usr-Sys*** condition), the system would initiate the interaction by first automatically presenting ("pushing") a message, then wait for the participant to judge the "initial label" and give a "user label" using the "Correct" or "Wrong" button. The first interval between the system's push and the participant's judgement click plus 0.5 seconds would then be the interval between the second and the third system's automatic message push. Then the second interval between the system's push and the participant's judgement click plus 0.5 seconds would be the interval between the third and the fourth system's push, and so on. Consequently, if the participant sped up when judging the "initial label" of the current message, their next message would be pushed by the system sooner, and if the participant slowed down for the current message, the system would correspondingly push the next message later. Therefore, the timing of the interaction was implicitly set by the participant, with whom the system aligned, though the message push actions were initiated by the system rather than the participant. Task 4 (***Usr-r*** condition) was considered as a baseline, where the participant could dictate the pace of the interaction and assume full control of the timing of all actions, including manually retrieving ("pulling") new messages by clicking a "Show Next" button to give them "user labels" one by one.

6.2.4 Dependent variables and measures

6.2.4.1 Subjective ratings

Immediately after each task, participants were asked to make subjective ratings for their sense of control and stress level on six NASA-TLX sub-scales together with another set of questions, see **Section 3.3** and **Section 4.1.3**. This experiment adopted the same rating interface as Experiments 1 and 2, as shown in Figure 4.3. During hypothesis testing, participants' ratings on each sub-scale were individually analysed and contrasted, as have been done in previous studies (Deaton & Parasuraman, 1993; Rowe et al., 1998; Yurko et al., 2010; Mehta & Agnew, 2011).

6.2.4.2 Behavioural data

Research studying people's rhythmic tapping behaviours has shown that people are less likely to synchronise tapping with an unresponsive and non-adaptive computer partner compared with an adaptive human partner (Himberg, 2006), and as reported in Experiments 1 and 2 in this dissertation, such entrainment effects also exist in mixed-initiative interaction. Therefore similar phenomena may also be expected to occur in AI-assisted labelling tasks.

Drawing on that work, two coefficients were used to measure entrainment. The first is the joint lag 1 autocorrelation of one series of intervals, which represents the "similarity between observations" of a signal itself. A positive value (0-1) suggests a greater tendency for temporal assimilation, whereas a negative value indicates compensation (Nowicki et al., 2013). The second measure is the windowed cross-correlation between two sets of intervals. This coefficient describes the "similarity of two interacting series as a function of the displacement of one relative to the other" with local stability assumed (Boker et al., 2002), and the greater the value, the stronger the similarity.

All timestamps of participants' mouse clicks, and all their labelling decisions, were recorded automatically by the experimental system. Based on the timestamps and participants' decisions, the following measurements as illustrated in Figure 6.3 were calculated as dependent variables for hypothesis testing:

1. Length of intervals between every two label judgement clicks by the participant. For instance, in the ***Sys-ii*** condition, the interval between the participant's the i^{th} and the $i+1^{th}$ clicks was t_{Sys-ii,U_i}. This notation rule also applied to the ***Sys-pr***, ***Usr-Sys*** and ***Usr-r*** condition, with intervals denoted as t_{Sys-pr,U_i}, $t_{Usr-Sys,U_i}$ and t_{Usr-r,U_i}.

2. Length of intervals between the display of a new message (either pushed by the system in ***Sys-ii***, ***Sys-pr*** and ***Usr-Sys***, or pulled by the participant themselves in ***Usr-r***) and the participant's corresponding judgement click, or "response interval" in short. For instance, in the ***Sys-ii*** condition, the interval between when the system pushed the i^{th} message and when the participant labelled the i^{th} message was denoted by t_{Sys-ii,R_i}. Similarly, in the ***Sys-pr***, ***Usr-Sys*** and ***Usr-r*** condition there were t_{Sys-pr,R_i}, $t_{Usr-Sys,R_i}$ and t_{Usr-r,R_i} respectively.

3. Queue length was recorded every time when the user had just finished labelling a message. It was the number of accumulated messages that were displayed in the table on the screen, including the message that had just been labelled by the participant. When the participant takes all of the control in the ***Usr-r*** condition, the queue length was not necessarily low, because participants could retrieve several messages by clicking the "Show Next" button several times to process messages in batches rather than one by one. By measuring queue length, the width of a time window was obtained, within which local stability could be assumed, so that the windowed auto-correlation/cross-correlation coefficients could be calculated to measure the entrainment effect.

4. Windowed auto-correlation of the participant's response intervals. If the width of the window is w, and the interval lag is 1, then the i^{th} auto-correlation coefficient in the ***Sys-ii*** condition was calculated using the following formula:

$$\frac{1}{w}\sum_{k=i}^{i+w-1}\frac{(t_{Sys-ii,U_k}-\overline{t_{Sys-ii,U_{k\sim k+w-1}}})\times(t_{Sys-ii,U_{k+1}}-\overline{t_{Sys-ii,U_{k+1\sim k+w}}})}{std(t_{Sys-ii,U_{k\sim k+w-1}})\times std(t_{Sys-ii,U_{k+1\sim k+w}})},$$

in which $\overline{t_{Sys-ii,U_{k\sim k+w-1}}}$ was the mean of the w intervals from t_{Sys-ii,U_k} to $t_{Sys-ii,U_{k+w-1}}$, and $\overline{t_{Sys-ii,U_{k+1\sim k+w}}}$ was the mean of the w intervals from $t_{Sys-ii,U_{k+1}}$

to $t_{Sys-ii,U_{k+w}}$. While $std(t_{Sys-ii,U_{k\sim k+w-1}})$ was the standard deviation of the w intervals from t_{Sys-ii,U_k} to $t_{Sys-ii,U_{k+w-1}}$, and $std(t_{Sys-ii,U_{k+1\sim k+w}})$ was the standard deviation of the w intervals from $t_{Sys-ii,U_{k+1}}$ to $t_{Sys-ii,U_{k+w}}$. The same formula was applied to the ***Sys-pr***, ***Usr-Sys*** and ***Usr-r*** condition too.

5. Windowed cross-correlation between message updating intervals and the participant's response intervals. If the width of the window is w, then the i^{th} cross-correlation coefficient in the ***Sys-ii*** condition was calculated using the following formula:

$$\frac{1}{w}\sum_{k=i}^{i+w-1}\frac{(t_{Sys-ii,S_k}-\overline{t_{Sys-ii,S_{k\sim k+w-1}}})\times(t_{Sys-ii,R_k}-\overline{t_{Sys-ii,R_{k\sim k+w-1}}})}{std(t_{Sys-ii,S_{k\sim k+w-1}})\times std(t_{Sys-ii,R_{k\sim k+w-1}})},$$

   in which $\overline{t_{Sys-ii,S_{k\sim k+w-1}}}$ was the mean of the w intervals from t_{Sys-ii,S_k} to $t_{Sys-ii,S_{k+w-1}}$, and $\overline{t_{Sys-ii,R_{k\sim k+w-1}}}$ was the mean of the w intervals from t_{Sys-ii,R_k} to $t_{Sys-ii,R_{k+w-1}}$. While $std(t_{Sys-ii,S_{k\sim k+w-1}})$ was the standard deviation of the w intervals from t_{Sys-ii,S_k} to $t_{Sys-ii,S_{k+w-1}}$, and $std(t_{Sys-ii,R_{k\sim k+w-1}})$ was the standard deviation of the w intervals from t_{Sys-ii,R_k} to $t_{Sys-ii,R_{k+w-1}}$. The same formula was also applied to the ***Sys-pr***, ***Usr-Sys*** and ***Usr-r*** condition.

6. The number of "initial labels" that were actually consistent with the designed truth but incorrectly rejected by the participant (i.e. "false positive"), and the number of initial labels that contradicted the designed truth but were incorrectly accepted by the participant (i.e. "false negative").

For each of the first three measurements, there were thirty data entries in every task completed by a given participant. Considering that participants usually had a short break between two tasks, they would need some time to warm up and get used to the new task setting when they started again. Their performance would be more stable after processing the first few messages, and therefore the first two entries were removed and the last twenty eight entries were averaged during the analysis. Similarly, for the fourth and fifth measurements, the first two intervals were removed and the rest of the intervals were calculated using the selected window width w. Therefore, in each task completed by a participant, $28 - w$ entries of auto-correlation/cross-correlation coefficients could be obtained.

6.2.5 Participants

Fifteen participants (age $M = 26.4$, $\sigma = 5.60$; 3 females) were recruited, all native English speakers. Three participants were left-handed. In each experimental session, participants were allowed to use the computer mouse on their preferred side. Seven participants have normal vision and eight have corrected-to-normal vision. One participant has "no green, minimal blue" colour blindness.

Participants' education level ranged from PhD to high school, the breakdown is as follows: five participants have obtained a PhD degree, five with a Masters degree, three with an undergraduate degree, and the other two had just completed high school and were studying as undergraduates. Eleven participants were studying STEM subjects (e.g. physics, biology, computer science, engineering), two studying music, one in development studies, and one didn't specify.

As shown in Table 6.2, thirteen participants reported that they had received music training for different periods of time. All thirteen have had experience in instrument playing, among them seven people had received training in singing, four had been composing, and two conducting.

6.2.6 Apparatus

All experiment sessions were carried out in the Usability Lab of the Computer Laboratory, University of Cambridge. All participants used the same desktop computer (System: Windows 10 Pro, 64-bit; CPU: 2.80GHz; RAM: 8.00GB) with the same computer monitor (Samsung, SM2443BW 24-inch Black Widescreen LCD, 1920×1200) and the same optical mouse (Microsoft IntelliMouse Optical 1.1A).

The experimental system was implemented using C# as a Windows Presentation Foundation (WPF) application. The software was written completely by the author of the dissertation, and they can be found via this link:

https://github.com/ChristineGuoYu/PhD_Experiment_3

During every experiment session, the programme ran in Visual Studio Community 2015 environment (Version 14.0.23107.0 D14REL). The front end interface was

Table 6.2: Participants' background information in Experiment 3

No.	*Age*	*Gender*	*Handedness*	*Vision*	*Colour-blindness*	*Education*	*Subject*	*Music Training*
1	21	M	R	Corrected to normal	No	Bachelor	Computer Science	11 years, 1hr/week instrument playing
2	21	M	R	Normal	No	Bachelor	Computer Science	14 years, 4hr/week instrument playing, singing, composing, conducting
3	31	M	R	Normal	No	PhD	Computer Science	25 years, 2hr/week instrument playing, singing
4	19	M	R	Corrected to normal	No	High school	Physics	3 years, 3hr/week instrument playing
5	19	M	R	Normal	No	High school	-	10 years, 1hr/week instrument playing
6	27	M	R	Normal	No	Bachelor	Microbiology	6 years, 20hr/week instrument playing
7	35	M	L	Corrected to normal	No green, minimal blue	Masters	Music	15 years, 5hr/week instrument playing, singing, composing
8	28	M	R	Corrected to normal	No	PhD	Biology	-
9	31	M	R	Normal	No	PhD	Computer Science	10 years, 0.5hr/week instrument playing, singing
10	24	M	L	Corrected to normal	No	PhD	Development Studies	9 years, 14hr/week instrument playing
11	38	M	R	Normal	No	PhD	Computer Science	5 years, 14hr/week instrument playing
12	27	F	L	Corrected to normal	No	Masters	Engineering	10 years, 0hr/week instrument playing, singing, composing
13	24	F	R	Normal	No	Masters	Music	17 years, 5hr/week instrument playing, singing, composing, conducting
14	27	F	R	Corrected to normal	No	Masters	Engineering	5 years, 2hr/week instrument playing, singing
15	24	M	R	Corrected to normal	No	Masters	Computer Science	-

connected to a backend MySQL database (MySQL Server, version 5.7.19; MySQL Connector/Net, version 6.9.9), from which the messages were stored, pulled and categorised in real time.

6.3 Result analysis

Using the same procedures described at the beginning of **Section 4.2** and **Section 5.2**, the data obtained from Experiment 3 were analysed in order to test hypotheses $\boldsymbol{H_{AIaL} - 1}$, $\boldsymbol{H_{AIaL} - 2}$, $\boldsymbol{H_{AIaL} - 3}$, and $\boldsymbol{H_{AIaL} - 4}$.

Step 1: The data under investigation was first examined by an omnibus test, either repeated-measure one-way ANOVA or non-parametric Friedman Test depending on the data distribution. This can reveal whether or not the independent variable had caused a significant overall effect across different task conditions.

Step 2: After confirming a significant overall effect, a *planned* contrast analysis was carried out in SPSS following the procedures introduced in Rosenthal et al. (1985)'s and Haans (2018)'s guides on contrast analysis. The hypotheses under investigation were translated into one or several sets of contrast weights.

Step 3: Each set of hypothesised contrasts defined in Step 2 were analysed under an F test ($F = \frac{MS_{contrast}}{MS_{error}}$) (Rosenthal et al., 1985; Furr & Rosenthal, 2003). When testing multiple sets of contrasts (e.g. k) against the same set of data, the alpha level for each F test was corrected using the *Bonferroni* method ($\alpha = 0.05/k$).

Step 4: Then *post-hoc* pairwise tests were used in order to reveal more insights from the results of the omnibus test in Step 1 and of the F test(s) in Step 3.

6.3.1 Balance of task load

In order to test whether the task load across four conditions was balanced, the average length of intervals between two adjacent labelling actions of the same participant in each condition was calculated. For twelve out of fifteen participants, the average interval length ranged between 4.012 seconds and 4.455 seconds, while three participants were removed as outliers based on the Q-Q plots in Figures B.3 and B.4 in Appendix B.7. The average interval length passed the Shapiro-Wilk Normality Test, but the data violated Mauchly's Test of Sphericity ($Mauchly's\ W$=0.041, χ^2=27.942, DoF=5, p<0.001), hence the non-parametric Friedman Test was used to analyse the effect of rhythm setting in the initiative taking across four conditions. No significant overall effect across four conditions was found in the result, as shown in Table 6.3.

Measurement	*N*	χ^2	*df*	***Sig***
Task load	11	3.109	3	.375

Table 6.3: Omnibus test for balanced task load (Friedman Test)

Therefore it could be confirmed that the task load across four conditions was balanced after randomising the labels, the sequence of messages and the sequence of tasks, and the initial estimate of single message reading time and labelling time as 4.4 seconds was valid.

6.3.2 Sense of control

The testing of hypothesis $\boldsymbol{H_{AIaL} - 1}$ is done by analysing participants' subjective ratings for their perceived sense of control. Two outliers were removed due to invalid responses (i.e. null response, or selecting "1" or "100" on all items). The ratings on "sense of control" did not pass the Shapiro-Wilk Normality Test, therefore the non-parametric Friedman Test was used to analyse the overall main effect of rhythm in initiative taking across four conditions. As shown in Table 6.4, the rhythm did cause a significant difference (p=0.002) in participants' reported sense of control in four tasks.

Measurement	*N*	χ^2	*df*	***Sig***
Task load	14	15.259	3	**.002***

Table 6.4: Omnibus test for participants' rating for sense of control (Friedman Test)

As reported in **Sections 4.2.1** and **5.2.1**, participants' rating for their sense of control increased as the rhythm of the mixed-initiative interaction became more predictable and under their control in both Experiment 1 and Experiment 2. Should hypothesis $\boldsymbol{H_{AIaL} - 1}$ be true, that a system-imposed rhythm during AI-assisted labelling can impair the user's perceived control, while a user-set rhythm can preserve their perceived control, participants' rating for sense of control in Experiment 3 should increase as they took more initiative. In other words, an upward linear trend ("↗") should exist in the rating data across the conditions in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, similar to the linear trends found in **Sections 4.2.1** and **5.2.1**. Hence the same set of weights were assigned in Table 6.5.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = -1$	$\lambda_3 = 1$	$\lambda_4 = 3$

Table 6.5: Weights in contrast analysis for participants' rating for sense of control (Hypothesis $\boldsymbol{H_{AIaL} - 1}$: "↗" trend)

The results of contrast analysis are shown in Table 6.6. The predicted upward linear trend was confirmed to be very significant (p<0.001). As can be seen in Figure 6.4, participants gave a higher rating for their sense of control when they took all the initiative in the ***Usr-r*** condition. Hence hypothesis $\boldsymbol{H_{AIaL} - 1}$ is supported.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↗"	Rhythm	191178.286	1	191178.286	33.169	**.000***
	Error (Rhythm)	74929.714	13	5763.824		

Table 6.6: Contrast analysis for participants' rating for sense of control (Hypothesis $\boldsymbol{H_{AIaL} - 1}$: "↗" trend)

In order to further examine the effect of rhythm setting on participants' reported sense of control during AI-assisted labelling, pairwise comparisons were conducted. The Wilcoxon Signed Ranks Test was adopted because the rating data did not pass the normality test earlier, and the alpha level was reduced to 0.05/6 = 0.0083 using the *Bonferroni* correction. The results are shown in Table 6.7. Participants reported a significantly stronger sense of control when they set all the rhythm in the ***Usr-r*** condition than in all other three conditions. No significant difference was found between ***Sys-ii*** vs. ***Sys-pr*** and ***Sys-pr*** vs. ***Usr-Sys***. This might be due to each message being "pushed" by the system in each of ***Sys-ii***, ***Sys-pr*** and ***Usr-Sys***, rather than “pulled” by participants in the ***Usr-r*** condition: hence the rhythm in the first three conditions might have all been perceived as imposed by the system. Therefore hypothesis $\boldsymbol{H_{AIaL} - 1}$ still stands.

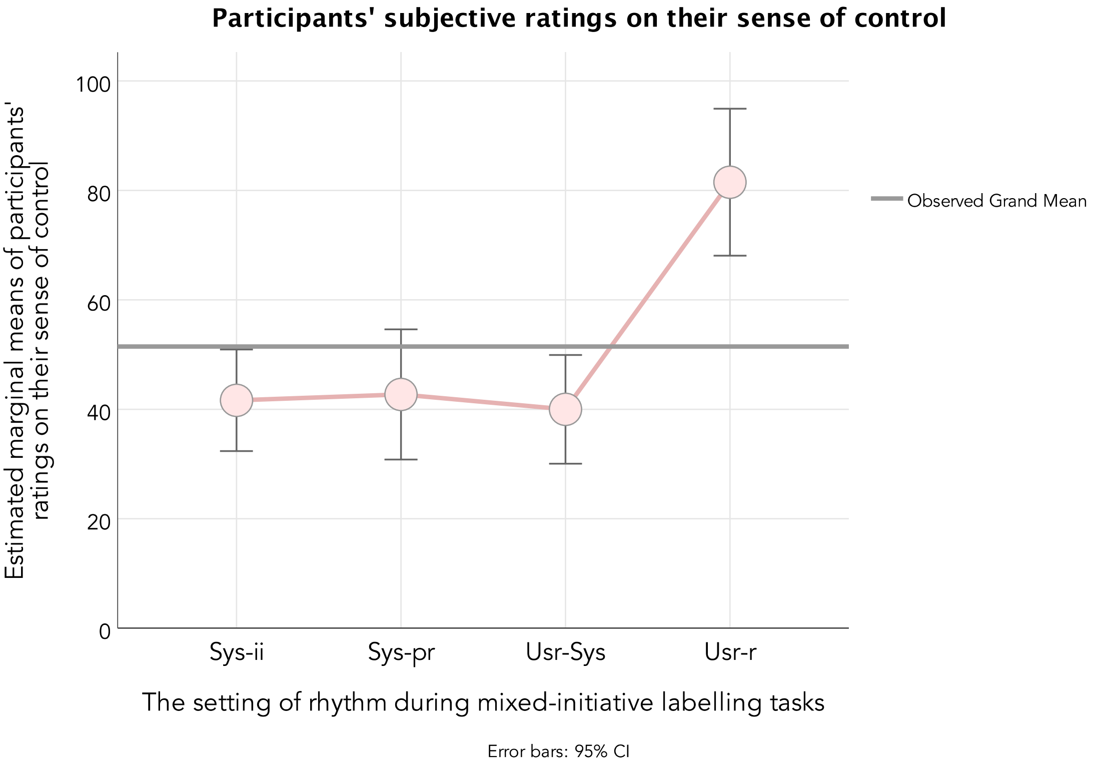


Figure 6.4: Participants’ subjective ratings on their sense of control in different tasks in Experiment 3

To sum up, the results of the omnibus test confirmed that the rhythm setting had a significant overall effect on participants’ reported sense of control, and the results of contrast analysis confirmed that participants felt significantly more in control when they set the rhythm during AI-assisted labelling tasks, less so when the rhythm was imposed by the system, as predicted in $\boldsymbol{H_{AIaL} - 1}$. These findings were further supported by the results of pairwise comparisons, and the limitations will be discussed in **Section 6.4.2**.

Pair	***Sys-ii*< *Sys-pr***	***Sys-ii*< *Usr-Sys***	***Sys-ii*< *Usr-r***	***Sys-pr*< *Usr-Sys***	***Sys-pr*< *Usr-r***	***Usr-Sys*< *Usr-r***
Z	-0.175	-0.377	-3.108	-0.663	-2.919	-2.983
Sig.	.861	.706	**.002***	.508	**.004***	**.003***

Table 6.7: Pairwise comparisons for participants' rating for sense of control (Wilcoxon Signed Ranks Test)

6.3.3 Perceived stress level

Participants' ratings on each of the six TLX sub-scales were analysed in order to test hypothesis $\boldsymbol{H_{AIaL} - 2}$. The ratings on the "mental demand", "temporal demand", and "effort" sub-scales passed the Shapiro-Wilk Normality Test, but those on "physical demand", "success", and "frustration" failed. Hence the overall main effect of rhythm setting on "mental demand", "temporal demand" and "effort" was tested using ANOVA with repeated measures, whilst the "physical demand", "success" and "frustration" tested with the non-parametric Friedman Test. As shown in Table 6.8 and Table 6.9, among the six sub-scales, a significant overall difference was found on participants' ratings on "mental demand" (p=0.020), "temporal demand" (p<0.001) and "effort" (p=0.019) sub-scales respectively.

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
TLX mental demand	35.786	3	11.929	3.657	**.020***
Error (Rhythm)	127.214	39	3.262		
TLX temporal demand	239.196	3	79.732	10.261	**.000***
Error (Rhythm)	303.054	39	7.771		
TLX effort	38.429	3	12.810	3.726	**.019***
Error (Rhythm)	134.071	39	3.438		

Table 6.8: Omnibus test for participants' ratings on TLX "mental demand", "temporal demand" and "effort" sub-scales (ANOVA with repeated measures)

Hypothesis $\boldsymbol{H_{AIaL} - 2}$ predicts that the user would feel more stressed when the rhythm was imposed by the system compared to the rhythm being set by themselves during AI-assisted labelling tasks. Should $\boldsymbol{H_{AIaL} - 2}$ be true, they would rate the

Measurement	*N*	χ^2	*df*	***Sig***
TLX physical demand	14	4.500	3	.212
TLX success	14	2.486	3	.478
TLX frustration	14	7.289	3	.063

Table 6.9: Omnibus test for participants' ratings on TLX "physical demand", "success" and "frustration" sub-scales (Friedman Test)

tasks as being less mentally and temporally demanding on the TLX sub-scales in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. Therefore, a downward linear trend ("↘") should be found in the rating data for both "mental demand" and "temporal demand". The same set of weights were assigned accordingly in Table 6.10.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for TLX mental demand, "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -3$
Contrast weights for TLX temporal demand, "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -3$

Table 6.10: Weights in contrast analysis for participants' ratings on TLX mental demand and temporal demand sub-scales (Hypothesis $\boldsymbol{H_{AIaL} - 2}$: "↘" trend)

The results of contrast analysis, as shown in Table 6.11, confirmed that the downward linear trend was significant in both the rating data of mental demand (p=0.009) and temporal demand (p<0.001). Participants rated the tasks as being less mentally and temporal demanding as the rhythm of AI-assisted labelling became more under their control, in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** as illustrated in Figure 6.5. Therefore, $\boldsymbol{H_{AIaL} - 2}$ is supported.

For the TLX "effort" sub-scale, participants' ratings should follow a downward linear trend ("↘") in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***. Alternatively, a cubic trend ("↘↗↘") similar to the one in Experiment 1 (as reported in **Section 4.2.2**) might also exist in the data in Experiment 3, given that $\boldsymbol{H_{AIaL} - 2}$ is derived from $\boldsymbol{H_{MII} - 3.1}$, and the rhythm settings in Experiment 3 are derived from Experiment 1. Therefore two sets of contrast weights were assigned in Table 6.12 to test the hypothesis $\boldsymbol{H_{AIaL} - 2}$.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
TLX mental demand						
"↘"	Rhythm	604.571	1	604.571	9.545	**.009***
	Error (Rhythm)	823.429	13	63.341		
TLX temporal demand						
"↘"	Rhythm	3363.500	1	3363.500	27.474	**.000***
	Error (Rhythm)	1591.500	13	122.423		

Table 6.11: Contrast analysis for participants' ratings on the TLX mental and temporal demand sub-scales (Hypothesis $\boldsymbol{H_{AIaL} - 2}$: "↘" trend)

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -3$
Contrast weights for "↘↗↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 1$	$\lambda_2 = -3$	$\lambda_3 = 3$	$\lambda_4 = -1$

Table 6.12: Weights in contrast analysis for participants' rating on the TLX effort sub-scale (Hypothesis $\boldsymbol{H_{AIaL} - 2}$: "↘" or "↘↗↘" trend)

The results of contrast analysis were shown in Table 6.13. With the alpha level corrected using the *Bonferroni* method ($0.05/2 = 0.025$), the predicted linear trend was not found to be significant in the rating data, whereas the predicted cubic trend was confirmed to be significant (p=0.023) in the ratings data on the TLX "effort" sub-scale, just as the results reported in **Section 4.2.2** in Experiment 1. Hence $\boldsymbol{H_{AIaL} - 2}$ is only partially supported in Experiment 3.

Following the contrast analysis above, pairwise comparisons were carried out on participants' ratings on the TLX "mental demand", "temporal demand", and "effort" sub-scales. The paired samples t test was used here because the data passed the normality test earlier. Again the alpha level was corrected as $0.05/6 = 0.0083$ using the *Bonferroni* method. The results are presented in Table 6.14. Participants reported the ***Sys-ii*** condition as being significantly more "mentally demanding" than ***Usr-Sys***, and they felt the pace of the task was significantly more "hurried"/"rushed" in ***Sys-ii***, ***Sys-pr*** and ***Usr-Sys*** than in the ***Usr-r*** condition. In addition, participants reported

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↘"	Rhythm	240.286	1	240.286	2.248	.158
	Error (Rhythm)	1389.714	13	106.901		
"↘↗↘"	Rhythm	528.286	1	528.286	6.593	**.023***
	Error (Rhythm)	1041.714	13	80.132		

Table 6.13: Contrast analysis for participants' rating on the TLX effort sub-scale (Hypothesis $\boldsymbol{H_{AIaL}} - \mathbf{2}$: "↘" or "↘↗↘" trend)

that they devoted significantly more effort in the ***Sys-ii*** condition than in the ***Sys-pr*** condition. Differences were also observed in other pairs, but not to a significant level under the *Bonferroni* correction.

Pair	***Sys-ii*** *vs.* ***Sys-pr***	***Sys-ii*** *vs.* ***Usr-Sys***	***Sys-ii*** *vs.* ***Usr-r***	***Sys-pr*** *vs.* ***Usr-Sys***	***Sys-pr*** *vs.* ***Usr-r***	***Usr-Sys*** *vs.* ***Usr-r***
TLX mental demand						
t	2.539	3.539	2.664	0.298	0.918	0.537
df	13	13	13	13	13	13
Sig.	.025	**.004***	.019	.770	.375	.600
TLX temporal demand						
t	0.773	0.552	5.709	0.058	5.114	3.520
df	13	13	13	13	13	13
Sig.	.453	.591	**.000***	.955	**.000***	**.004***
TLX effort						
t	3.226	0.295	2.253	-1.764	0.331	2.590
df	13	13	13	13	13	13
Sig.	**.007***	.773	.042	.101	.746	.022

Table 6.14: Pairwise comparisons for participants' ratings on TLX mental demand, temporal demand and effort sub-scales (paired samples *t* test)

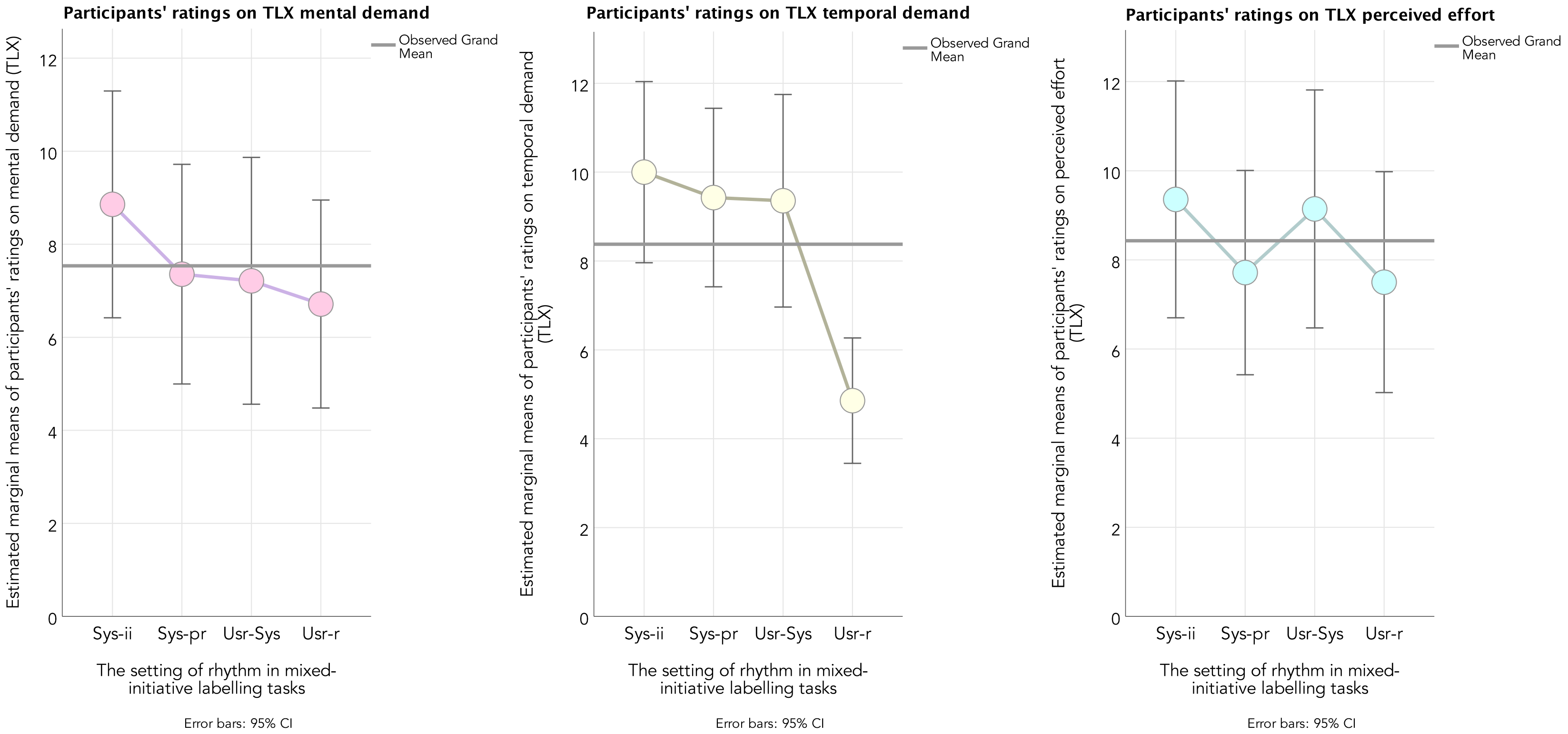


Figure 6.5: Participants' ratings on the TLX sub-scales (mental demand, temporal demand, and perceived effort devoted to the task) in different tasks in Experiment 3

Hypothesis $\boldsymbol{H_{AIaL}-2}$ is further supported by participants' post-task subjective ratings for how much they felt "the system was helping me vs. the system was challenging me".

The ratings on how much they felt "being helped/challenged by the system" passed the Shapiro-Wilk Normality Test after removing the outliers (as shown in the Q-Q plots in Figures B.5 and B.6 in Appendix B.7). Hence within-subjects repeated measures ANOVA was used as the omnibus test. As shown in Table 6.15, the rhythm setting caused a significant overall effect (p=0.029) on participants' rating across different task conditions.

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Perceived help	902.625	3	300.875	3.327	**.029***
Error (Rhythm)	3527.125	39	90.439		

Table 6.15: Omnibus test for the sense of being challenged/helped by the system (ANOVA with repeated measures)

As hypothesised in $\boldsymbol{H_{AIaL}-2}$, during AI-assisted labelling, the user will experience a higher level of stress if the rhythm is imposed by the system compared with it being set themselves. If this were true, participants would feel more challenged when the system imposed the labelling rhythm, and more helped when they set the rhythm. Therefore a downward linear trend ("↘") should exist in participants' rating for "the system was helping me vs. the system was challenging me" in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r***, and the corresponding weights were assigned in Table 6.16.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -3$

Table 6.16: Weights in contrast analysis for participants' rating for their sense of being challenged/helped by the system (Hypothesis $\boldsymbol{H_{AIaL}-2}$: "↘" trend)

The predicted downward linear trend was confirmed to be significant (p=0.018) by the contrast analysis, the results of which are shown in Table 6.17. As pictured in Figure 6.6, participants tended to rate the system as challenging them when it imposed the labelling rhythm in ***Sys-ii*** and ***Sys-pr*** conditions, but reported the system as being more helpful when they set the rhythm themselves, such as in ***Usr-Sys*** → ***Usr-r*** conditions. Therefore $\boldsymbol{H_{AIaL} - 2}$ is supported.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↘"	Rhythm	17643.500	1	17643.500	7.329	**.018***
	Error (Rhythm)	31297.500	13	2407.500		

Table 6.17: Contrast analysis for participants' rating for their sense of being challenged/helped by the system (Hypothesis $\boldsymbol{H_{AIaL} - 2}$: "↘" trend)

To further investigate the effect of rhythm settings on how the participants perceived the system (as being helpful or challenging), pairwise comparisons were conducted among the four task conditions. The data failed the normality test earlier, hence the Wilcoxon Signed Ranks Test was adopted. As before, the alpha level was $0.05/6 = 0.0083$ under the *Bonferroni* correction. The results are shown in Table 6.18. Participants felt significantly more challenged in the ***Sys-ii*** condition than in the ***Usr-r*** condition, while other pairs were not as significant under the adjusted alpha value.

Pair	***Sys-ii***< ***Sys-pr***	***Sys-ii***< ***Usr-Sys***	***Sys-ii***< ***Usr-r***	***Sys-pr***< ***Usr-Sys***	***Sys-pr***< ***Usr-r***	***Usr-Sys***< ***Usr-r***
Z	-0.664	-1.327	-2.852	-0.534	-2.081	-1.730
Sig.	.507	.185	**.004***	.593	.037	.084

Table 6.18: Pairwise comparisons for participants' rating for their sense of being challenged/helped by the system (Wilcoxon Signed Ranks Test)

In summary, the results of omnibus tests confirmed that the rhythm setting caused a significant overall effect on participants ratings for mental demand, temporal demand, effort devoted in tasks, and how challenging/helpful they perceived the system to be. The results of planned contrast analysis confirmed that, as predicted in $\boldsymbol{H_{AIaL} - 2}$, participants felt less stressed when the rhythm of AI-assisted labelling was set by themselves, given that their reported mental demand and temporal demand decreased significantly as participants took more initiative. The system was also perceived as significantly more helpful than challenging as participants had more control over the rhythm. However, not every pair of conditions were found to be different significantly during pairwise comparisons with the *Bonferroni* correction, hence the findings above should be accepted with due caution and limited confidence. The limitations will be discussed in **Section 6.4.2**.

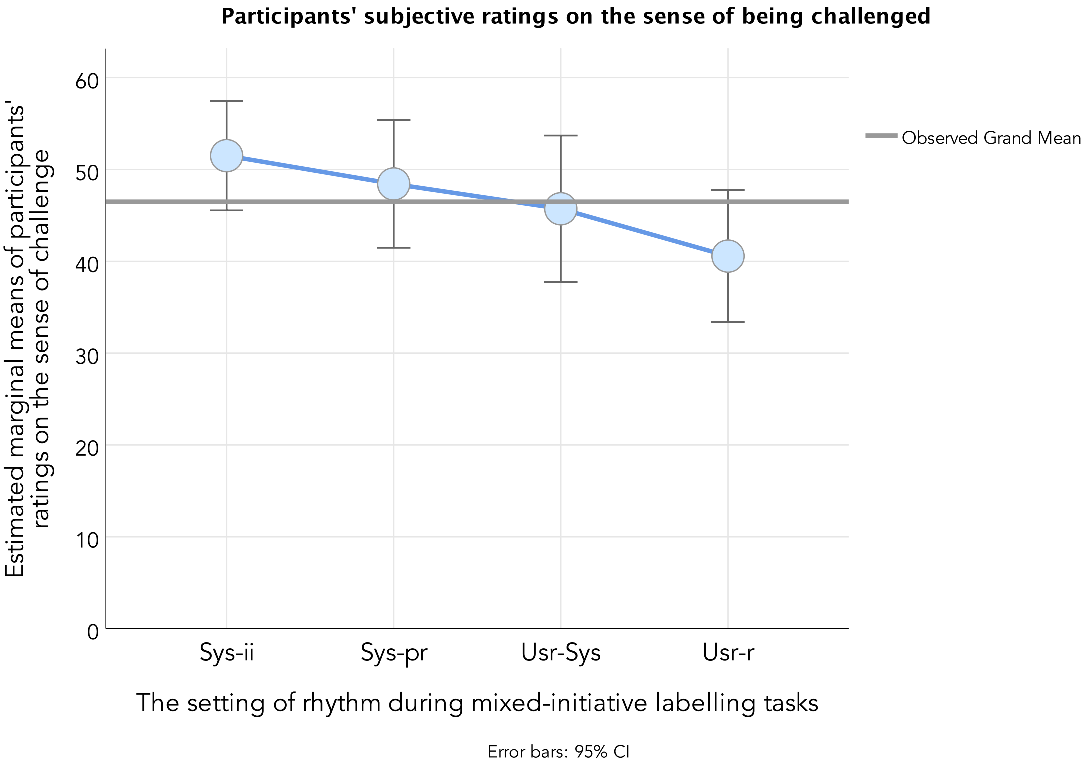


Figure 6.6: Participants' subjective ratings for their sense of being helped/challenged by the system in different tasks in Experiment 3

6.3.4 Accumulated task load

In order to test hypothesis $\boldsymbol{H_{AIaL}}\boldsymbol{-3}$, the average queue length (number of accumulated messages) across four task conditions is used as a measurement for objective task load. The average queue length did not pass the Shapiro-Wilk Normality Test, therefore the non-parametric Friedman Test was used to test whether or not the rhythm caused a significant overall effect on the average queue length. As shown in Table 6.19, the main overall effect was very significant (p<0.001).

According to $\boldsymbol{H_{AIaL} - 3}$, the accumulation of task load should be higher if the

Measurement	*N*	χ^2	*df*	***Sig***
Avg queue length	13	25.039	3	**.000***

Table 6.19: Omnibus test for average queue length (Friedman Test)

rhythm of AI-assisted labelling was imposed by the system, but lower if the rhythm was set by the user. Therefore the average queue length in either ***Sys-ii*** or ***Sys-pr*** condition should be longer than that in ***Usr-Sys*** or ***Usr-r***. The corresponding contrast weights were assigned in Table 6.20. However, as shown in Table 6.21, the predicted trend was not significant in the data.

Conditions (Tasks)	***Sys-ii***	***Sys-pr***	***Usr-Sys***	***Usr-r***
Contrast weights for "↘" trend ($\Sigma\lambda = 0$)	$\lambda_1 = 1$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = -1$

Table 6.20: Weights in contrast analysis for the average queue length (Hypothesis $\boldsymbol{H_{AIaL}}$**−3**: "↘" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↘"	Rhythm	0.555	1	0.555	3.648	.085
	Error (Rhythm)	1.520	10	0.152		

Table 6.21: Contrast analysis for the average queue length (Hypothesis $\boldsymbol{H_{AIaL}}$ **− 3**: "↘" trend)

Because the rhythm did cause a significant overall effect according to Table 6.19, a *post-hoc* analysis was carried out based on observation. As pictured in Figure 6.7, if the order of the four conditions were re-arranged as ***Usr-r*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** (by simply placing ***Usr-r*** at the beginning while keeping the order of the other three conditions), an upward linear trend ("↗") could be observed. To test whether or not this observed trend was significant, a new set of weights were assigned in Table 6.22. The results of contrast analysis, as shown in Table 6.23, proved that the observed upward linear trend was very significant (p<0.001). One potential explanation to this observation where the average queue length was increasing (rather than decreasing) in the order of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** will be discussed in **Section 6.4.1**. The reasons why the queue length in ***Usr-Sys*** was not as short expected will be discussed in **Section 6.4.2**.

Conditions (Tasks)	***Usr-r***	***Sys-ii***	***Sys-r***	***Usr-Sys***
New contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_4 = -3$	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 3$

Table 6.22: New weights in *post-hoc* contrast analysis for the average queue length (*post-hoc*: "↗" trend)

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↗"	Rhythm	33.005	1	33.005	66.831	**.000***
	Error (Rhythm)	4.939	10	0.494		

Table 6.23: *Post-hoc* contrast analysis for the average queue length (*post-hoc*: "↗" trend)

Following the contrast analysis reported as above, pairwise comparisons were carried out to further examine the effect of the rhythm setting on the average queue length during AI-assisted labelling. The Wilcoxon Signed Ranks Test was used because the data was not normally distributed, and the alpha level was reduced to $0.05/6 = 0.0083$ under the *Bonferroni* correction. The results are presented in Table 6.24. The average queue length in the ***Usr-r*** condition was significantly shorter than in all other three conditions. The average queue length of ***Usr-Sys*** was also longer than that of ***Sys-ii***. In other words, the average task load was the lowest when the user set the rhythm, and became higher when the system dictated the pace or mimicked participants' pace.

Pair	***Sys-ii< Sys-pr***	***Sys-ii> Usr-Sys***	***Sys-ii< Usr-r***	***Sys-pr< Usr-Sys***	***Sys-pr< Usr-r***	***Usr-Sys< Usr-r***
Z	-0.051	-2.805	-2.937	-1.531	-2.807	-2.936
Sig.	.959	**.005***	**.003***	.126	**.005***	**.003***

Table 6.24: Pairwise comparisons for average queue length (Wilcoxon Signed Ranks Test)

In short, the results of omnibus test showed changes to the rhythm setting caused a significant overall effect on the average queue length during AI-assisted labelling.

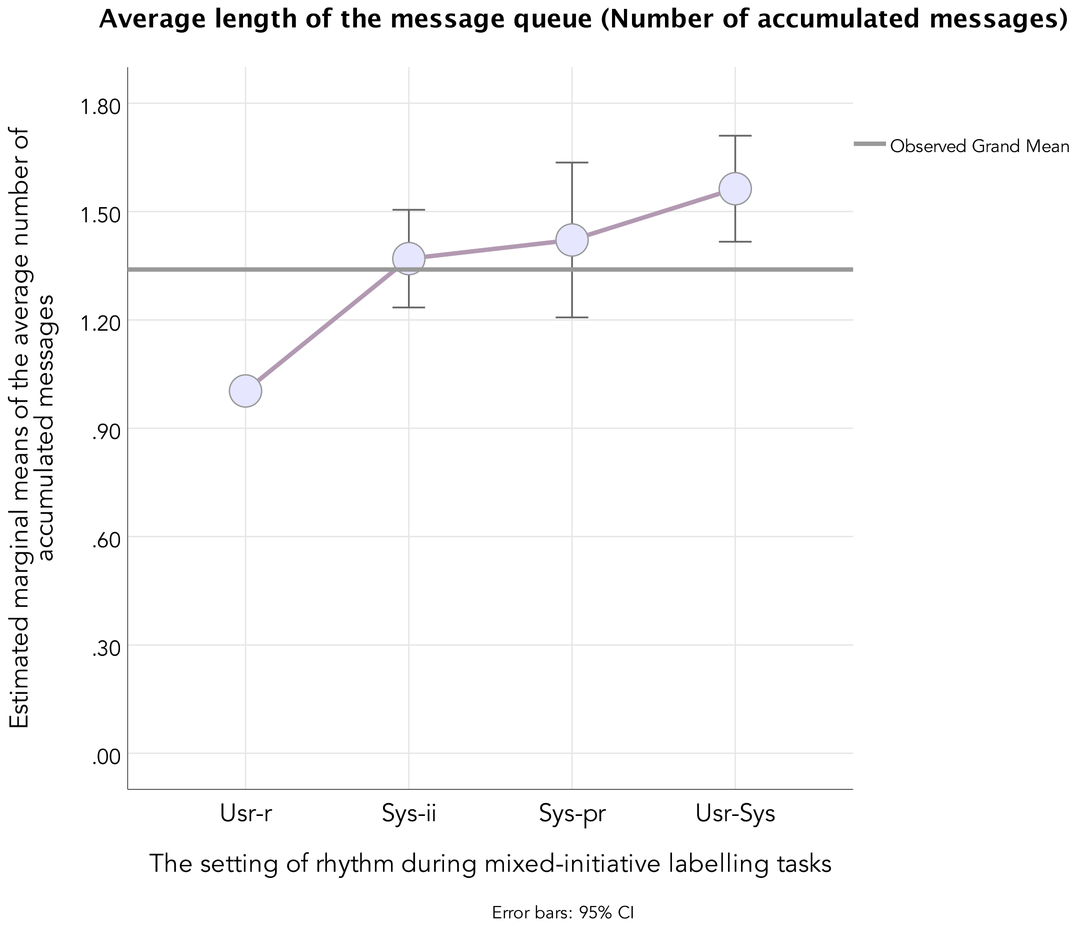


Figure 6.7: Average queue length (i.e. average number of accumulated messages, including the current message)

The results of *post-hoc* contrast analysis revealed that a significant trend existed in the data, that the average queue length significantly increased in the order of $\boldsymbol{Usr\text{-}r} \rightarrow \boldsymbol{Sys\text{-}ii} \rightarrow \boldsymbol{Sys\text{-}pr} \rightarrow \boldsymbol{Usr\text{-}Sys}$. Therefore, hypothesis $\boldsymbol{H_{AIaL} - 3}$ is only partially supported by ***Usr-r***'s queue length being the shortest among all conditions. Moreover, not all pairs of conditions were found to be significantly different under pairwise tests with the *Bonferroni* correction applied. Hence the findings above should be accepted with due caution and limited confidence.

6.3.5 Entrainment behaviours

Lastly hypothesis $\boldsymbol{H_{AIaL} - 4}$ is tested by comparing two coefficient series: the windowed auto-correlation of participants' response intervals (i.e. the interval between a message being displayed and its corresponding labelling click), and the windowed cross-correlation between message displaying intervals (i.e. the intervals between every two messages' appearance), and participants' response intervals. Among all fifteen participants and all four conditions, the maximum median queue length was 5, therefore the window width was set as 5 so that local stability can be assumed within the moving window.

Both coefficient series passed the Shapiro-Wilk Normality Test, and neither violated the assumption of sphericity according to the result of Mauchly's Test of Sphericity ($Mauchly's$ W=0.655, χ^2=5.390, DoF=5, $p_{auto-cor} = 0.371$; $Mauchly's$ W=0.616, χ^2=6.161, DoF=5, $p_{cross-cor} = 0.292$). Hence the main overall effect of the rhythm setting was tested by ANOVA with repeated measures in SPSS. As shown in Table 6.25, the rhythm setting caused a significant difference in both coefficients in different tasks (p<0.001, p=0.004).

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Auto-correlation	0.716	3	0.239	11.114	**.000***
Error (Rhythm)	0.901	42	0.021		
Cross-correlation	0.460	3	0.153	5.075	**.004***
Error (Rhythm)	1.270	42	0.030		

Table 6.25: Omnibus test for the auto-correlation coefficients of participants' response intervals, and the windowed cross-correlation coefficients between message displaying intervals and participants' response intervals (ANOVA with repeated measures)

As hypothesised in $\boldsymbol{H_{AIaL} - 4}$, a more predictable rhythm in AI-assisted labelling is more likely to induce the user's entrainment behaviours, while irregular timing is less likely so. If $\boldsymbol{H_{AIaL} - 4}$ held true in this experiment, a quadratic trend ("↗↘") similar to the ones reported in **Section 4.2.3** in Experiment 1 should exist in the coefficients.

However, given that the results in the last section (**Section 6.3.4**) do not fully

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for auto-correlation, “↗↘” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 1$	$\lambda_4 = -1$
1^{st} set of contrast weights for cross-correlation, “↗↘” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 1$	$\lambda_4 = -1$
2^{nd} set of contrast weights for cross-correlation, “↗↘↗” trend ($\Sigma\lambda = 0$)	$\lambda_1 = -1$	$\lambda_2 = 3$	$\lambda_3 = -3$	$\lambda_4 = 1$

Table 6.26: Weights in contrast analysis for the auto-correlation coefficients of participants' response intervals, and the windowed cross-correlation coefficients between message displaying intervals and participants' response intervals (Hypothesis $\boldsymbol{H_{AIaL} - 4}$: “↗↘” and “↗↘↗” trend)

confirm $\boldsymbol{H_{AIaL} - 3}$ due to unexpectedly long queue length in the ***Usr-Sys*** condition, the cross-correlation coefficient in ***Usr-Sys*** might also have a “dip” among other conditions, because such accumulated task load could be caused by unmatched timing between the system and the user, hence a cubic trend (“↗↘↗”) might exist in the cross-correlation coefficient data instead. The corresponding contrast weights for each coefficient were assigned as shown in Table 6.26, and the test results are presented in Table 6.27.

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Auto-correlation						
“↗↘”	Rhythm	2.001	1	2.001	19.536	**.001***
	Error (Rhythm)	1.434	14	0.102		
Cross-correlation						
“↗↘”	Rhythm	0.145	1	0.145	2.568	.131
	Error (Rhythm)	0.791	14	0.056		
“↗↘↗”	Rhythm	7.834	1	7.834	7.388	**.017***
	Error (Rhythm)	14.845	14	1.060		

Table 6.27: Contrast analysis for the auto-correlation coefficients of participants' response intervals, and the windowed cross-correlation coefficients between message displaying intervals and participants' response intervals (Hypothesis $\boldsymbol{H_{AIaL} - 4}$: “↗↘” or “↗↘↗” trend)

A significant quadratic trend was found in the auto-correlation coefficients

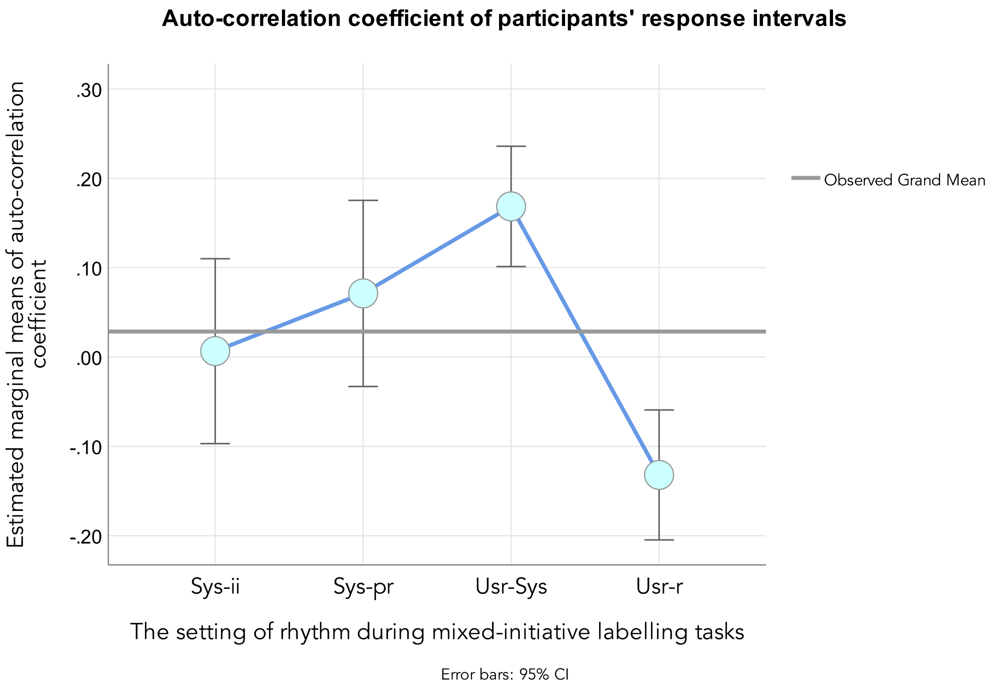
(p=0.001), but not in the cross-correlation coefficients. As shown in Figure 6.8 (a), the value of the auto-correlation of participants' response intervals rose and fell ("↗↘") in the direction of ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** → ***Usr-r*** as anticipated. As shown in Figure 6.8 (b), a significant cubic trend ("↗↘↗", p=0.017) rather than a quadratic trend ("↗↘") could be observed in the cross-correlation data, with the alpha level ($0.05/2 = 0.025$) adjusted using the *Bonferroni* method.

Following the contrast analysis above, pairwise comparisons were conducted in order to further examine the effect of the rhythm setting on the auto-correlation and cross-correlation coefficients. The data passed the normality test earlier, hence the paired samples t test was adopted, and again the alpha level was corrected as $0.05/6 = 0.0083$ using the *Bonferroni* method. The results are shown in Table 6.28. The auto-correlation of ***Usr-r*** was significantly weaker than that of ***Sys-pr*** and ***Usr-Sys***. Differences were observed among other pairs but not to a significant level. In terms of the cross-correlation, differences were also observed, but no pair reached the significant level under the *Bonferroni* correction.

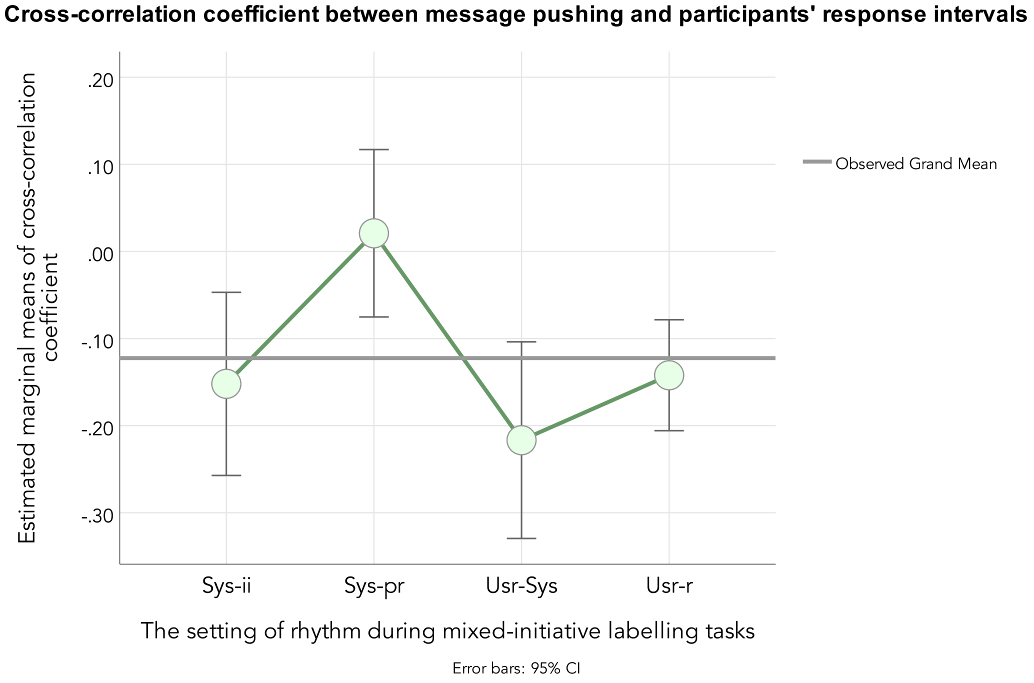
Pair	***Sys-ii–Sys-pr***	***Sys-ii–Usr-Sys***	***Sys-ii–Usr-r***	***Sys-pr–Usr-Sys***	***Sys-pr–Usr-r***	***Usr-Sys–Usr-r***
Auto-correlation						
t	-1.186	-2.945	2.326	-2.421	3.254	6.560
df	14	14	14	14	14	14
Sig.	.255	.011	.036	.030	**.006***	**.000***
Cross-correlation						
t	-2.574	1.197	-0.173	2.863	2.892	-1.285
df	14	14	14	14	14	14
Sig.	.022	.251	.865	.013	.012	.220

Table 6.28: Pairwise comparisons for contrast analysis for the auto-correlation coefficients of participants' response intervals, and the windowed cross-correlation coefficients between message displaying intervals and participants' response intervals (paired samples t test)

The results reported above echo with the findings in Experiment 1, meaning that participants entrained more with the system when it had rhythmic pace, did not entrain with it when it was arrhythmic, and slackened their own pace when they had full control. However, when the system mirrored participants' pace in the ***Usr-Sys*** condition, participants did not entrain with the system as much as expected. Therefore



(a)



(b)

Figure 6.8: (a) Average windowed auto-correlation coefficient of participants' response intervals (i.e. the interval between a message being displayed and its corresponding labelling click) and (b) average windowed cross-correlation coefficient between message displaying intervals (i.e. the intervals between every two messages' appearance) and participants' response intervals in different tasks in Experiment 3

hypothesis $\boldsymbol{H_{AIaL} - 4}$ is partially supported. The limitations will be discussed in **Section 6.4.2**.

6.4 Further analysis and discussion

6.4.1 Design implications

As stated in **Section 6.1.3**, the four hypotheses tested in this experiment were derived from the hypotheses tested in Experiments 1 and 2. The results of contrast analysis in **Section 6.3.2** support hypothesis $\boldsymbol{H_{AIaL} - 1}$ (derived from $\boldsymbol{H_{MII} - 1}$), further confirming that letting participants take the initiative and set the rhythm can preserved their sense of control. When the system took the initiative, predictable intervals made the tasks less mentally demanding and alleviated participants' perceived effort compared with irregular intervals, as shown in **Section 6.3.3**. Hence $\boldsymbol{H_{AIaL} - 2}$ (derived from $\boldsymbol{H_{MII} - 3.1}$ and $\boldsymbol{H_{MII} - 3.2}$) is partially supported. Participants also exhibited a stronger tendency to entrainment when the system took the initiative rhythmically or when the system was emulating the user's rhythm, as $\boldsymbol{H_{AIaL} - 4}$ (derived from $\boldsymbol{H_{MII} - 2}$) is tested and partially supported in **Section 6.3.5**. The first design implication is therefore:

> ***Design implication* 3.1**: The effects of timing on the user's sense of control and their perceived level of stress and effort observed in simplified and controlled stimulus-response experiments remain relatively congruent in a relatively more realistic mixed-initiative interaction context, such as interacting with AI-assisted labelling tools.

In the course of the results analysis, additional effects beyond the original hypotheses were noted, and *post-hoc* tests were made as follows. The average intervals between message display (i.e. push/pull) events and participants' corresponding labelling action were compared across four conditions. After four outliers were removed due to abnormally long average intervals (i.e. around 30s in one condition, but only around 4s in the other three), the rest of the data passed the Shapiro-Wilk Normality Test. The data did not violate Mauchly's Test of Sphericity ($Mauchly's\ W$=0.455,

χ^2=6.414, DoF=5, $p_{queue-length}$=0.271), so the overall main effect was tested using ANOVA with repeated measures in SPSS. As shown in Table 6.29, the rhythm setting caused a significant overall main effect (p=0.007).

Measurement	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
Average interval	7.682	3	2.621	4.892	**.007**
Error (Rhythm)	16.072	30	0.536		

Table 6.29: Omnibus test for average interval (ANOVA with repeated measures)

Recalling the analysis in **Section 6.3.4** regarding the average accumulated task load (i.e. queue length), a very significant upward linear trend ("↗") was found across four conditions in the order of ***Usr-r*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys***. In other words, more messages were accumulated in ***Usr-Sys*** than in ***Sys-pr*** and ***Sys-ii***, and ***Usr-r*** had the lowest queue length on average. This indicates that participants might have made labelling decisions faster in ***Usr-r***, slower in ***Sys-ii*** and ***Sys-pr***, and slowest in ***Usr-Sys***. The weights for this predicted trend were assigned in Table 6.30.

Conditions (Tasks)	***Usr-r***	***Sys-ii***	***Sys-r***	***Usr-Sys***
Contrast weights for "↗" trend ($\Sigma\lambda = 0$)	$\lambda_4 = -3$	$\lambda_1 = -1$	$\lambda_2 = 1$	$\lambda_3 = 3$

Table 6.30: Weights in contrast analysis for average labelling decision interval (hypothesis: "↗" trend)

As shown in Table 6.31, the upward linear trend ("↗") was found to be significant (p=0.006) in participants' average response interval, in the same direction of ***Usr-r*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** as in the analysis for average queue length in **Section 6.3.4**, as can be seen in Figure 6.9.

This observation further confirmed the finding in the previous two experiments in **Chapters 4** and **5**: that during mixed-initiative interaction, the user preferred to devote a bit more physical effort (e.g. more clicking) in exchange for less uncertainty and mental stress (when they were less stressed, they could process information even faster). This finding is also supported by the results in **Section 6.3.3**: participants reported the ***Usr-r*** condition as the *least* temporally demanding one (see Table 6.11

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
"↗"	Rhythm	152.036	1	152.036	12.314	**.006***
	Error (Rhythm)	123.471	10	12.347		

Table 6.31: Contrast analysis for average labelling decision interval (hypothesis: "↗" trend)

and Figure 6.5). Therefore, echoing the ***Design implication* 1.1** proposed in **Chapter 4**, the second design implication drawn from this experiment is:

> ***Design implication* 3.2**: In AI-assisted labelling tasks, if the user can take the initiative and have full control over the interaction rhythm, they experience a higher sense of control, a lower level of challenge, and feel less "rushed" in the tasks without actually slowing down.

Given that a "↗" trend was observed in both the average queue length and the average response interval in the order of ***Usr-r*** → ***Sys-ii*** → ***Sys-pr*** → ***Usr-Sys*** (see Figure 6.7 and Figure 6.9), it is worth noting that the ***Sys-ii*** condition ranked the second, after ***Usr-r***. In other words, when the system was pushing messages at random intervals, participants' average response interval was shortened. It suggests that participants might have been forced to respond *faster* to cope with external uncertainty and avoid accumulated task load, which was also why their average queue length during the ***Sys-ii*** condition was shorter too, as illustrated in Figure 6.7 in **Section 6.3.4**. This *post-hoc* observation is further supported by the findings in **Section 6.3.3**, where participants reported the ***Sys-ii*** condition as the *most* temporally demanding one, as shown in Table 6.11 and Figure 6.5. Hence the next design implication complements ***Design implication* 3.2** by presenting the opposite case:

> ***Design implication* 3.3**: In AI-assisted labelling tasks, if the system pushes messages at irregular and unpredictable intervals, the user is likely to shorten their processing time for each message to cope with the temporal irregularity. Consequently they perceive themselves as being "rushed", and perceive the task as more mentally demanding and challenging.

The number of wrongly labelled messages in different tasks was also analysed, including the occurrence of false positives (i.e. when a participant wrongly rejected

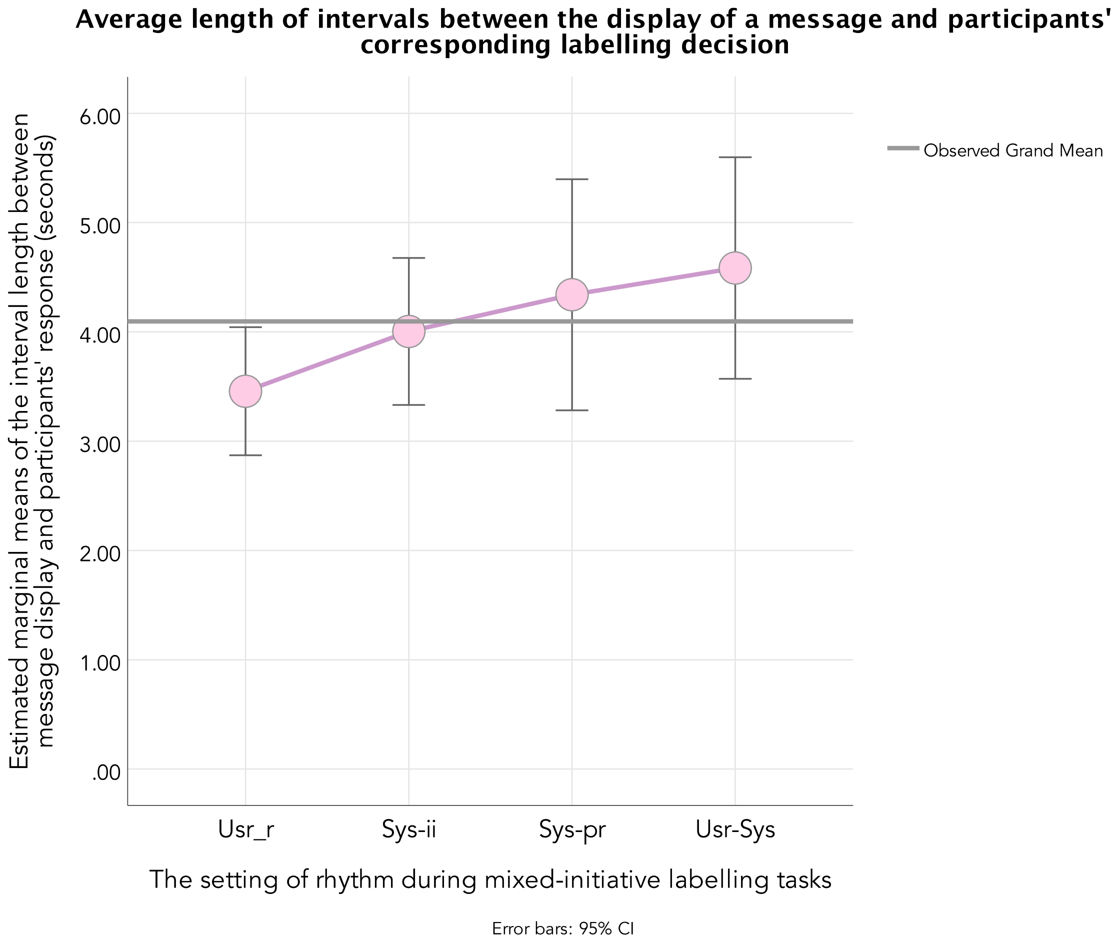


Figure 6.9: The average response interval (sec) between the display of a message and participants' corresponding labelling click on that message in different tasks in Experiment 3

the system's correct "initial label"), of false negatives (i.e. when a participant wrongly accepted the system's wrong "initial label"), and the sum of both. The data was not normally distributed, therefore the non-parametric Friedman Test was used to test the overall main effect of the setting of the interaction rhythm. Significant overall difference was only found in the false positive category (χ^2=8.205, p=0.042), as shown in Table 6.32.

Under the *expectation states theory* (EST), as reviewed in **Section 2.2.2**, when a person is at a higher position in the "power-and-prestige" order in a group, they are more likely to be more assertive, more critical of others' performance, give others fewer opportunities to speak, and attribute less credit to others' contribution. Given that

Measurement	*N*	χ^2	*df*	***Sig***
Wrongly rejected	15	8.205	3	**.042***

Table 6.32: Omnibus test for the number of wrongly rejected computer labels (Friedman Test)

mixed-initiative interaction such as AI-assisted labelling is essentially a co-operative task between the human user and the system where their contributions are interwoven (Horvitz, 1999b), participants might have considered themselves as a coach who was in a more powerful position that the system, hence they might have been more critical of the system's suggestions, and might have given less credit to the system's contribution (Bonito et al., 1999; Fişek et al., 1995). If the EST theory was still true in the AI-assisted labelling context, a cubic trend ("↗↘↗") should exist in the data of false positive errors (i.e. when a participant wrongly rejected the system's correct "initial label"). The weights for contrast analysis were assigned in Table 6.33.

Conditions (Tasks)	***Sys-ii***	***Sys-r***	***Usr-Sys***	***Usr-r***
Contrast weights for "↗↘↗" trend ($\Sigma\lambda = 0$)	$\lambda_1 = -3$	$\lambda_2 = 1$	$\lambda_3 = -1$	$\lambda_4 = 3$

Table 6.33: Weights in contrast analysis for wrongly rejected computer labels (hypothesis: "↗↘↗" trend)

The results of contrast analysis confirmed the prediction above, as shown in Table 6.34. There was a significant "↗↘↗" trend (p=0.014) in the number of participants' false positive errors among the four conditions, as pictured in Figure 6.10. In other words, participants were more likely to wrongly reject labels that were correctly given by the system when they had the full control or when the temporal structure was fully predictable.

However, when the data was put under the Wilcoxon Signed Ranks Test for pairwise comparison, the difference between the conditions was not as significant after the alpha level was reduced to $0.05/6 = 0.0083$ using the *Bonferroni* correction method, as shown in Table 6.35.

Given that the prediction made based on the EST was confirmed by the results of the omnibus test and the contrast analysis but not by the pairwise analysis, the

Direction	*Source*	*Sum of Square* (SS)	*df*	*Mean square* (MS)	*F*	*Sig*
“↗↘↗”	Rhythm	45.067	1	45.067	7.796	**.014***
	Error (Rhythm)	80.933	14	5.781		

Table 6.34: Contrast analysis for the number of wrongly rejected computer labels (hypothesis: “↗↘↗” trend)

Pair	***Sys-ii< Sys-pr***	***Sys-ii> Usr-Sys***	***Sys-ii< Usr-r***	***Sys-pr< Usr-Sys***	***Sys-pr< Usr-r***	***Usr-Sys< Usr-r***
Z	-1.667	-0.000	-2.111	-1.667	-.632	-2.111
Sig.	.096	1.000	.035	.096	.527	.035

Table 6.35: Pairwise comparisons for the number of wrongly rejected computer labels (Wilcoxon Signed Ranks Test)

following is a research prediction that is worth exploring in the future:

> ***Observation—based prediction* 2**: In AI-assisted labelling tasks, if the user takes more initiative and has more control over the interaction rhythm, they may perceive themselves as being in a higher power status than the system, and wrongly reject more correct recommendations made by the system compared with when the system takes part of the initiative during labelling.

Another observation is that during the experiment, all participants were told that in Task 4 (the ***Usr-r*** condition), they could manually pull more messages by clicking the “Next” button more times. However, only one participant (Participant 10) did it occasionally: s/he sometimes pulled a few messages in a row and labelled them one by one, then moved on to another batch of messages. All other participants pulled and labelled one message at a time when they took the initiative and had full control of the rhythm. It might be a variation of the “grouping” effect discussed in **Section 5.3.1**, that for highly repetitive events, some users may prefer to process in batches. Just like other Programming-by-Example applications, AI-assisted labelling is essentially a kind of end-user programming and the interface should be able to afford highly customised experience and outcomes. Therefore in future research, it would

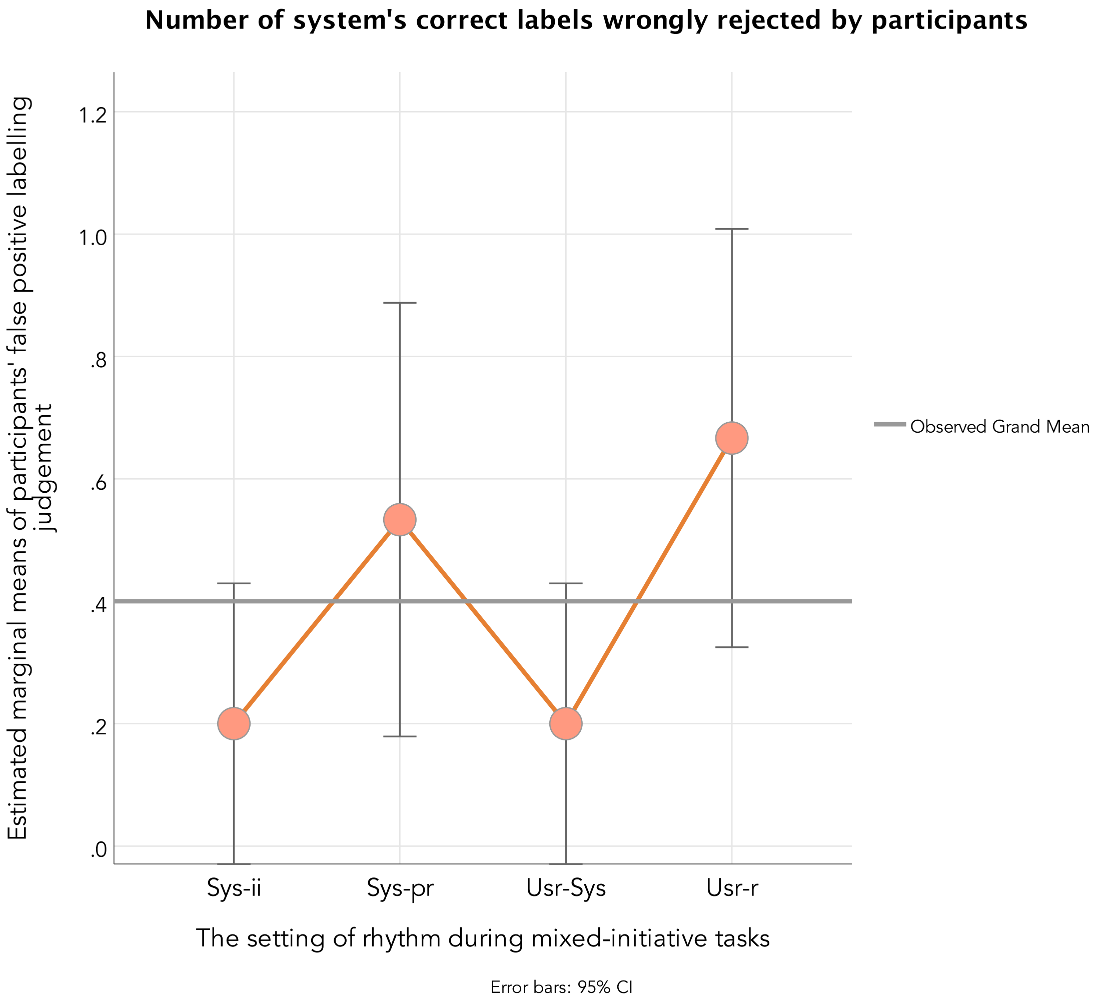


Figure 6.10: Number of the system's correct initial labels that were wrongly rejected by participants in different tasks in Experiment 3

be interesting to explore how to accommodate individuals' rhythm preference during labelling or demonstration, and whether or not their rhythm preference changes when the manual operations become more complicated.

6.4.2 Limitations

Although this experiment was designed to address some of the limitations in the first two experiments and has achieved that purpose, its findings and design implications were also limited in the following aspects:

- First, the messages presented to the participants were all predesigned carefully with controlled length and good grammar whilst containing distinctive keywords, as reported in **Section 6.2.1**. In realistic labelling tasks, the user may have to deal with messages that “vary in dialect, length, and legibility” (Blackwell, 2017). Consequently, the reading and processing time required by each message can vary greatly, and the user’s labelling decisions can also be different depending on how vague the meaning of messages is. Therefore, the findings and design implications should be applied and generalised with caution in a realistic interaction setting.

- Second, the initial labels of all the messages were randomly preassigned, so it is likely that participants did not see much improvement in the system’s performance over the course of each task. This might have contributed to participants’ overall frustration or confusion. It would be interesting to know whether or not the user would be more tolerant of random message pushing intervals and/or less reliant on manual control if the system appears to perform better over time. In existing AI-assisted labelling system like CODA (Blackwell, 2017), the system can direct the user’s attention to the initial labels (given by the machine learning algorithm in the system) that are presented with lower statistical confidence. Hence, it would also be worthwhile to explore how the user would distribute their time and invest their attention based on the statistical confidence of initial labels, and what kind of composition of the statistical confidence of initial labels can keep the user engaged and interested, rather than getting bored or frustrated over time.

- Third, after the experiment sessions, several participants reported the ***Usr-Sys*** condition (when the system aligns its next message pushing interval to the user’s last response interval) as *“a bit weird”* (Participant 8, 13) or *“dumb”* (Participant 6) but other conditions were *“fine”*. In addition, the ***Usr-Sys*** condition was also the primary reason that $\boldsymbol{H_{AIaL} - 3}$ and $\boldsymbol{H_{AIaL} - 4}$ are partially supported, because participants spent an unexpectedly long time labelling messages in the ***Usr-Sys*** condition, thus accumulating an unexpectedly long queue of unlabelled messages. There are two potential explanations to those impressions and observations: 1) Since giving a label took around 4 seconds on average, the timing pattern might have lagged too far behind for participants to recognise the causation or ascribe the authorship (Shanks et al., 1989; Wegner & Wheatley, 1999; Berthaut et al., 2015), especially when participants were not pre-informed with the temporal mimicry in the ***Usr-Sys*** condition, hence there was no "belief-like mental status" (Aarts et al., 2005) to support the formation of agency either. 2) Alternatively, participants might have noticed the temporal mimicry after all, but found it frustratingly rigid and awkward, as social psychology studies have found that strictly precise temporal entrainment during social interaction is not as likeable as the entrainment that is not-too-perfect (Clayton et al., 2005; Warner et al., 1987). In future studies, formal qualitative research methodologies, such as using video/audio-recording or think aloud protocol, may reveal more aspects of the user's experience, such as their impression, reaction and preference towards the setting of interaction rhythm. In addition, more methods of calculating intervals should be designed and evaluated, through which we may be able to find the ideal level of temporal entrainment that allows the system be perceived as aligning to the user *naturally* rather than *rigidly.*

- The fourth aspect is the same as the fourth limitation discussed in **Section 4.3.2** in Experiment 1 and the fifth limitation in **Section 5.3.2** in Experiment 2. Although all of the results of contrast analysis in **Sections 6.3.2**, **6.3.3**, **6.3.4** and **6.3.5** are significant and support hypotheses $\boldsymbol{H_{AIaL}-1}$, $\boldsymbol{H_{AIaL}-2}$, $\boldsymbol{H_{AIaL}-3}$ and $\boldsymbol{H_{AIaL}-4}$, the results of *independent* pairwise tests were not always in line with those of contrast analysis. Therefore, the findings and the resulting design implications in this chapter should be interpreted with due caution and accepted with limited confidence. However, it is also worthwhile reiterating the two arguments put forward in **Section 4.3.2**. Firstly, the *Bonferroni* correction may have been overly strict, and may have impaired the power of the pairwise tests (i.e. causing more Type II errors). Consequently statistically sound inferences may have been wrongly rejected (Perneger, 1998; Nakagawa, 2004; Armstrong, 2014), particularly when the tests in this experiment were *pre-planned* to test hypotheses based on existing theories, and the pairs being tested were *dependent* on each other. Secondly, contrast analysis can reveal the significant effect caused by the independent variable under investigation where an omnibus or pairwise test cannot (Rosenthal et al., 1985), because the latter disregards the arrangement of the multiple levels of the independent variable, for example, the four levels in this experiment and in Experiments 1 and 2 (i.e. the rhythmic character of the initiative-taking during the interaction was increasingly more predictable and under the user's control across the four conditions). Therefore, the results of contrast analysis should still be valued despite the fact that the results of pairwise tests were not always significant under the *Bonferroni* correction.

6.5 Summary

In this chapter, I contextualised the findings in previous chapters by designing and carrying out Experiment 3. In **Section 6.1.1**, I first introduced the significance of labelling in the training of artificial intelligence algorithms and the challenges faced during labelling. I then defined software tools that are designed to improve the efficiency and quality of labelling as "assisted labelling" tools, among which the ones that incorporate artificial intelligence components are "AI-assisted labelling" tools. I also recognised the mixed-initiative characteristics in AI-assisted labelling, as discussed in **Section 6.1.2**, following which I proposed four hypotheses in the context of AI-assisted labelling in **Section 6.1.3**.

I then reported the design of Experiment 3 that adopted the Wizard-of-Oz paradigm in **Section 6.2**, and analysed the results in **Section 6.3**. The four hypotheses were derived from the hypotheses tested in Experiments 1 and 2. The results of this experiment provided further evidence: when using an AI-assist labelling tool, if the user can take the initiative and set the rhythm of labelling, they will have a higher sense of control (Hypothesis $\boldsymbol{H_{AIaL} - 1}$), and a lower accumulated task load on average (Hypothesis $\boldsymbol{H_{AIaL} - 3}$) compared with letting the system take the initiative. When the system does take the initiative, rhythmic and predictable intervals can make the tasks appear less mentally demanding and effortful compared with arrhythmic ones (Hypothesis $\boldsymbol{H_{AIaL} - 2}$), and the user is also more likely to entrain with the system's rhythm (Hypothesis $\boldsymbol{H_{AIaL} - 4}$). It was also found that $\boldsymbol{H_{AIaL} - 3}$ and $\boldsymbol{H_{AIaL} - 4}$ are supported, with the exception of the ***Usr-Sys*** condition.

In addition, I proposed three design implications and one observation-based prediction based on the experiment results in **Section 6.4.1**. The first implication is that the effects of timing on the user's sense of control, level of stress and entrainment tendency observed in highly controlled experiments (i.e. Experiments 1 and 2) remain congruent in a more realistic task setting (i.e. using a simulation of an AI-assisted labelling tool). The second and the third implications came as a pair: the user feels the least stressed or rushed - even though they do not actually slow down - when they have full control of the rhythm, whereas if the system takes more initiative in an irregular manner, the user speeds up to cope with the temporal irregularity, and at the same time perceives a surge of stress and challenge in the task. Driven by the expectation states theory (EST), I put forward an observation-based prediction: that when the user takes the initiative and dictates the labelling pace, they may perceive themselves as in a more powerful position than the system (i.e. as a coach), hence they may be less likely to give credit to the system's contribution and more likely to reject correct predictions or recommendations made by the system.

The limitations of the experiment design and findings were discussed in **Section 6.4.2**. For example, the messages used in the experiment were written in perfect grammar, with clear meaning and controlled within a fixed length, during the Wizard-of-Oz experiment session, the system's labelling performance did not actually improve, the interval length was not calculated with flexibility when the system was designed to align with the user temporally, and incongruence exists between the results in contrast analysis and in pairwise tests. Therefore, the findings and design implications should be interpreted in their context and generalised with caution, and may require further investigations in future studies.

This experiment complemented this PhD research in three aspects: its results further support the main hypotheses proposed in **Chapter 3**, it showcases how interaction timing can be manipulated as a design resource in a more realistic and complex task scenario, and it offers practical design insights to interactive machine learning applications with mixed-initiative characteristics based on empirical evidence of the user's sense of control, experience of stress, and their entrainment behaviours. The insights from this experiment can also facilitate the end user's conversation-like interaction with more general decision support systems.

CHAPTER 7

Conclusion

Many intelligent systems are designed to complete or automate our actions, such as search boxes that can anticipate our questions and autocomplete our words (Ward, Hahn, & Feist, 2012), and semi-autonomous vehicles that tell us when to make a turn or even turn the steering wheel themselves (Casner, Hutchins, & Norman, 2016). While these systems relieve us from repetitive tasks (Cypher, 1995), we also risk losing our sense of control during the interaction (Blackwell, 2015). That is because we often interact with such systems in a *mixed-initiative* manner (Horvitz, 1999a), in which we and the system take turns as in a dialogue (Bauer et al., 2001; Sarkar, 2017), and the initiative is handed over in a back-and-forth manner.

As set out in **Chapter 1**, my PhD research was motivated by the challenge of how to preserve the user's sense of control during the transfer of initiative. My main goal was to address this challenge by identifying which design factors can influence the user's agency experience, and how these factors can be appropriately manipulated during interaction design. Given that mixed-initiative interaction resembles the turn-taking in human conversations (Horvitz, 1999b) where *timing* is a particularly important factor, I identified timing as a potential design resource.

7.1 Summary of findings

Having narrowed down the scope of my research to exploring the function of timing in the user's perceived control of an interaction, this research began with three research questions in **Chapter 1**. In order to answer my questions, I adopted an interdisciplinary approach and reviewed three bodies of literature in **Chapter 2**, which I used to formulate four sets of hypotheses in **Chapter 3**. The hypotheses were tested in three experiments as I reported in **Chapters 4**, **5** and **6**.

In this section, I will summarise my findings from the experiments. I will first answer the two closed questions, **Research Questions 2** and **3**. Then I will answer **Research Question 1**, which requires an open answer.

> **Research Question 2:** Can the timing of events become a design resource, which can be manipulated in a way that affects one's agency experience? If yes, then how can timing be manipulated to achieve this effect?

The answer is yes. In all three experiments, participants went through four kinds of task, each of which had one of the following temporal patterns: 1) the system took the initiative at irregular intervals (the ***Sys-ii*** condition), 2) the system took the initiative in a predictable rhythm (the ***Sys-pr*** condition), 3) participants took the initiative first, then the system aligned with their pace (the ***Usr-Sys*** condition), and 4) participants took the initiative at their own pace (the ***Usr-r*** condition). According to the results of contrast analysis in **Sections 4.2.1**, **5.2.1** and **6.3.2**, participants' reported sense of control increased significantly ("$\nearrow$") in the order of ***Sys-ii*** $\rightarrow$ ***Sys-pr*** $\rightarrow$ ***Usr-Sys*** $\rightarrow$ ***Usr-r*** as predicted in hypothesis $\boldsymbol{H_{MII} - 1}$, regardless of what they needed to do in the tasks - following random visual stimuli in Experiment 1, attending to random auditory stimuli in Experiment 2 or performing labelling tasks in Experiment 3. Therefore, it is confirmed that the timing of events can be a design resource, and manipulating timing can influence the user's perceived control.

I also noted that the three experiments in this dissertation were all controlled experiments, with Experiment 1 designed to test whether or not the timing manipulation was effective, Experiment 2 designed to consolidate the findings from Experiment 1 in a different modality, and Experiment 3 designed to contextualise and verify the findings from Experiments 1 and 2. Hence, only four basic kinds of temporal pattern were empirically evaluated. There are many more kinds of temporal pattern that can be applied to mixed-initiative interaction, and I will discuss them in **Section 7.4** as one direction for future research.

> **Research Question 3:** Can the rhythmic entrainment of a mixed-initiative interaction positively affect the user's experience, such as their sense of agency, perceived stress level, confidence and task performance? If yes, then what are the design guidelines?

The answer is yes. On the one hand, when the system aligned with (or, "entrained across turn boundaries") participants' pace in the ***Usr-Sys*** condition in Experiments 1 and 2, participants reported a stronger sense of control and higher confidence in their task accomplishment compared to when the system initiated events irregularly. In Experiment 2 in particular, participants appreciated such temporal alignment and rated the system as being more "helpful" and "adaptive" than other conditions. Further, as discussed in **Section 5.3.1**, participants explicitly reported a stronger sense of control when the system followed their pace compared to when the system maintained its own rhythmic pace, although there was no difference in the implicit measure for agency. This may imply that the system's rhythmic entrainment can give the user a more explicit experience of control. On the other hand, when the system followed participants' pace in Experiments 1 and 3[1], participants themselves also exhibited a stronger tendency to entrain to the local rhythm than when the system was initiating events irregularly or when participants had full control of the pace, as predicted in $\boldsymbol{H_{MII} - 2}$ and $\boldsymbol{H_{AIaL} - 4}$.

In summary, the system's rhythmic entrainment positively influenced participants' perceived control, confidence in task performance, as well as their perception of the system's helpfulness in Experiments 1 and 2.

However, as discussed in **Section 6.4.2**, when the system's temporal alignment was strict and rigid in a realistic and complex task setting - such as in Experiment 3 - where the interaction happened on a larger time scale, participants might not be able to

[1] As discussed in **Section 5.3.2**, entrainment was not studied in Experiment 2.

recognise the alignment, or might have simply found the temporal mimicry to be clumsy. This observation echoes the findings in social psychology that very precise rhythmic entrainment during social interaction may not appear as likeable as not-so-perfect entrainment (Clayton et al., 2005; Warner et al., 1987). If the observation above can be further investigated in qualitative studies or confirmed with more empirical evidence, it may bring good news to the mixed-initiative interaction community that there is leeway during the manipulation of the system's temporal alignment, and it needs not (or should not) be ultra-precise.

Finally, I will answer the first research question:

> **Research Question 1:** What timing characteristics are appropriate for mixed-initiative interaction?

According to the findings from all three experiments, letting the system take the initiative at irregular times in the ***Sys-ii*** condition was the *least* appropriate in mixed-initiative interaction, compared with other temporal settings that were tested with the same tasks. It had a negative influence on most aspects of the user's experience. For instance, participants reported the lowest sense of control, the highest amount of effort devoted to the tasks, and the least confidence in their own performance in Experiments 1, 2 and 3. In addition, as analysed in **Section 4.3.1**, participants had to take up a rhythm-keeping role to fight against the external randomness in time. They also reported that interacting with auditory stimuli or labelling messages that came irregularly were more mentally demanding, and they felt particularly rushed during the labelling task. As predicted in **Section 5.3.1**, if it is impossible to control the irregular timing of individual events, grouping them in a predictable pattern can potentially mitigate the negative influence on the user's experience.

It is unsurprising that when participants took the initiative and controlled the pace in the ***Usr-r*** condition, their perceived control was the strongest in all experiments: see **Sections 4.2.1**, **5.2.1** and **6.3.2**. While participants were happy with devoting more physical effort (e.g. more clicking) in exchange for a higher sense of control during a set amount of experimental tasks, letting a user take the initiative in realistic application may make the interaction appear tedious and boring over time, so that the user cannot enjoy the convenience brought by mixed-initiative interaction or end-user automation. Further, as noted in **Section 6.4.1**, when the users take all the initiative and set the pace, they may perceive themselves as having a higher power status than the system and become *too* confident, thereby giving the system less credit and wrongly rejecting correct suggestions made by the system.

Therefore, I suggest that the timing characteristics of the ***Sys-pr*** and the ***Usr-Sys*** conditions (falling between ***Sys-ii*** and ***Usr-r***) are more suitable and appropriate for mixed-initiative interaction. Firstly, in Experiments 1 and 3, participants reported that they devoted the least effort in both the ***Sys-pr*** and the ***Usr-r*** conditions. Participants' recall accuracy was also comparable between the two conditions in Experiment 1. Hence rhythmic system-initiated events in mixed-initiative interaction can make the interaction appear as effortless as dealing with user-initiated events while guaranteeing the task performance. Secondly, as summarised above in the answer to **Research Question 3**, participants appreciated it when the system temporally aligned with their pace, and explicitly reported a higher perceived control in the ***Usr-Sys*** condition than in ***Sys-pr*** in Experiment 2. Hence the system's rhythmic entrainment can preserve the user's experience of agency in mixed-initiative interaction. Lastly, as analysed in **Sections 4.2.2**, **5.2.3**, **5.2.4**, **6.3.3** and **6.3.4**, participants' reported level of confidence in their task performance, their sense of relaxation, their perceived level of challenge, mental demand and temporal demand of the tasks, as well as the average accumulated task load were mostly comparable (i.e. not significant in pairwise comparison) between the ***Sys-pr*** and the ***Usr-Sys*** conditions, while they both fall between the two extreme ends of the spectrum (***Sys-ii*** and ***Usr-r***) in contrast analysis. It is also important to note that the ***Usr-Sys*** condition in the three experiments was often, but not always, better than ***Sys-pr*** during contrast analysis (e.g. in **Sections 4.2.2**, **6.3.3** and **6.3.5**, the ***Usr-Sys*** was not as ideal as ***Sys-pr*** on some measures, twisting the predicted "↗" or "↘" trend into a cubic shape "↗↘↗" or "↘↗↘"). As discussed in **Sections 4.2.2** and **6.4.2**, it might be because the overly strict and rigid entrainment in the ***Usr-Sys*** condition was not perceived to be natural, hence the merits of entrainment were not as strong as expected.

In short, in mixed-initiative interaction, letting the system take the initiative in a predictable temporal pattern, either following a stable global rhythm or aligning with the local rhythm set by the user in immediate events, can provide the user with a satisfactory level of perceived control while preserving their confidence and guaranteeing a desirable task performance without increasing their level of stress. Moreover, when designing a system that actively entrains to the user's pace, more research and user testing are required to find the "right" temporal setting that is suitable for specific mixed-initiative task contexts.

7.2 Contributions

In seeking the answers to the three research questions above, this dissertation has made four contributions to the field of human-computer interaction, as summarised in **Section 1.4**. I will expound each contribution as follows:

> **Contribution 1:** This dissertation provides a cross-disciplinary review of the literature in the fields of human-computer interaction, cognitive neuroscience and social psychology, and establishes connections between the existing theories in the three fields to inform the design of mixed-initiative interaction that can preserve the user's perceived control (in **Chapters 2** and **3**).

As discussed in **Chapters 1** and **2**, although mixed-initiative interaction has been recognised as a common paradigm of interacting with intelligent user interfaces for two decades (Horvitz, 1999a), little research has been done regarding how to support the user's locus of control during the back-and-forth transfer of initiative. In order to address this issue, my first step was to lay a theoretical basis drawing from three bodies of literature, as presented thematically in **Chapter 2**.

I first reviewed the cognitive mechanisms underlying the production, the experience and the attribution of *agency* (i.e. sense of control) and the measures for the agency experience, then summarised the human factors and design solutions that the HCI community has already recognised as relevant to the user's sense of control in existing interaction paradigms.

I then highlighted the fact that the user's *expectations* for their computer counterpart can influence their interaction behaviours and subjective experience, including how they perceive the system's role and how they attribute credits to it during joint problem-solving, based on existing cognitive-behavioural models in social psychology. I also reviewed cognitive neuroscience studies that explain the formation and function of temporal expectation, which can throw light on the design of timing for mixed-initiative interaction.

Lastly, I introduced the *rhythmic entrainment* phenomenon in music and social psychology, and how it can facilitate interpersonal communication and coordination, and proposed that incorporating rhythmic entrainment in mixed-initiative interaction can be a solution to address the challenge of when the transfer of initiatives should happen so that it does not impair the user's locus of control.

Based on these interdisciplinary bodies of research, I established a theoretical research framework for this dissertation by proposing four sets of hypotheses in **Chapter 3**. My interdisciplinary review can also inform future research on the user's perceived control during mixed-initiative interaction.

> **Contribution 2:** This dissertation demonstrates the importance of timing during mixed-initiative interaction, proposes that the timing of an interaction, on both the visual and auditory modalities, can be manipulated as a design resource, and empirically tests the effect of timing on the user's perceived control (Experiments 1 and 2, in **Chapters 4** and **5**).

In **Section 7.1**, I answered **Research Questions 1**, **2** and **3** respectively based on the findings from the three experiments reported in this dissertation. All three experiments were a within-subject design, and the same four kinds of temporal patten of initiative taking were compared in terms of participants' perceived control, reported level of stress and confidence, as well as their entrainment behaviours and task performance.

As I hypothesised in **Chapter 3**, the results of Experiments 1 and 2 (in **Chapters 4** and **5**) showed that a predictable rhythm set either by participants themselves or by the system preserved participants' sense of agency, induced a stronger tendency to entrainment, reduced their perceived effort and level of stress, and helped them do better in the tasks and feel more confidence in their own performance, compared with irregular time intervals set by the system which had a negative impact in all aspects.

> **Contribution 3:** This dissertation provides quantitative measures for the user's entrainment behaviours during the handover of initiative on a relatively broad timescale, ranging from 250 milliseconds (Experiment 1 in **Chapter 4**) to 20 seconds (Experiment 3 in **Chapter 6**).

As reviewed in **Section 2.3.4**, previous studies on the entrainment phenomena in HCI viewed entrainment as a by-product of the interaction, and only recorded and discussed the observations from a qualitative perspective. In Experiments 1 and 3, I adapted the measures for entrainment from existing studies in social psychology and musicology studies (e.g. interpersonal synchrony) to the context of mixed-initiative interaction, thereby providing quantitative measures for the user's rhythmic entrainment behaviours.

Furthermore, my experiment findings remain congruent on a broad timescale. The average length of the intervals in Experiment 1 was between 600 and 1100 milliseconds, the shortest interval that was involved in the coefficient calculation was 257 milliseconds and the longest was 10707 milliseconds. Whilst the average interval length in Experiment 3 was between 2 and 15 seconds, with the shortest interval as 1.423 seconds, and the longest non-outlier interval was 22.652 seconds. Therefore, the findings in this dissertation exhibited a satisfactory level of robustness against temporal fluctuations in both simplified and realistic task settings.

> **Contribution 4:** This dissertation showcases how rhythmic entrainment principles can be applied to the design of mixed-initiative systems such as AI-assisted labelling tools (Experiment 3 in **Chapter 6**), offering insights that can inform the design of the temporal aspects of mixed-initiative systems that incorporate inference-based components (in **Chapter 7**).

As discussed in **Sections 4.3.2** and **5.3.2**, the tasks in Experiments 1 and 2 were controlled, using simple stimuli that did not carry much information. Hence, it was questionable whether or not the findings were generalisable to a more realistic context. In addition, during the course of this research, the theoretical framework and the findings from the first two experiments were published and presented to various audiences in the field of HCI, end-user programming, and machine learning. While the topic was generally received with great interest, the most frequently asked question was what kind of application would benefit from the design of timing.

Motivated by the two questions above, I designed Experiment 3 to contextualise my findings. At the beginning of **Chapter 6**, I defined existing design solutions that aim to improve the efficiency and quality of the labelling work as assisted-labelling tools. I further recognised that an interface that is used to train an interactive machine learning algorithm can be characterised as an AI-assisted labelling tool, which should afford a back-and-forth flow of interaction between the labeller and the statistical model being trained. I also identified that training an IML algorithm on an AI-assisted labelling tool is essentially a form of end-user programming. Combining the above reasons, I validated my findings of the effects of timing on the user's perceived control in the context of interacting with an AI-assisted labelling tool in Experiment 3, which successfully addressed both the limitations of previous experiments and the question raised by broader research communities.

I proposed ten design implications and two observation-based research predictions in total and discussed them in **Sections 4.3.1**, **5.3.1** and **6.4.1** respectively. Apart from the implications that were derived from the results of hypothesis testing, I offered additional insights gained from *post-hoc analysis*. For example, as suggested in ***Design implications*** **1.3** and **1.4**, the user may make an extra effort in maintaining their own rhythm if the system is taking more initiative, hence when the system detects a rhythm-keeping tendency from the user, it can provide the user with more opportunities to take the initiative, which can potentially alleviate their perceived effort and stress.

While ***Design implication*** **3.3** suggests that the user tends to accelerate their processing in order to cope with the temporal irregularity of system-initiated events, ***Observation—based prediction*** **1** highlights the potential that grouping a batch of individual events that occur at irregular times in a regular pattern may help the user anchor to a temporal expectation and preserve a sense of agency.

Further, while ***Design implication*** **1.1** points out that the user is happy to devote more physical effort in exchange for a higher sense of control during mixed-initiative interaction, ***Observation—based prediction*** **2** highlights the potential risk of letting the user take "too much" control, because the user may be more likely to reject the system's correct recommendations more often. This risk can limit the co-operative aspect of mixed-initiative interaction (Bauer et al., 2001) and the merits of end-user automation or intelligent decision support. Therefore, how the user's perceived control can influence their perceptions and expectations of the system's competence and contributions (Bonito et al., 1999; Pearson et al., 2006) should be taken into account during the design of mixed-initiative decision support systems.

7.3 Limitations

The limitations of the design and the findings of the three experiments were discussed in detail in **Sections 4.3.2**, **5.3.2** and **6.4.2** respectively. Here I will reflect on the limitations of the three experiments.

- First, all three experiments were designed and carried out in a controlled and simplified manner, in order to eliminate potential confounding factors introduced by the complexity of the task content. For example, the visual targets used in Experiment 1 were simple geometric shapes, the auditory signals used in Experiment 2 were identical beep sounds, and the messages presented to the participants in Experiment 3 were carefully designed with a fixed length, good grammar, and clear meaning. In realistic applications, however, the user will be interacting with visual and auditory representations that possess more dimensions of information, such as semantics, logic and analytics. For the labelling tasks, the user will need to deal with illegible messages, distinguish ambiguous categories, and process cases that have a mixed level of difficulty. Consequently, the user needs to exercise high-level cognitive skills, such as problem solving and decision making, and distribute their attention and time according to the task requirements. Therefore, the findings in this dissertation should be applied and generalised with caution.

- Second, because there was not a standard empirical paradigm of manipulating the timing in mixed-initiative interaction in the existing HCI literature, the four temporal patterns of initiative-taking that were compared in the three experiments, namely ***Sys-ii***, ***Sys-pr***, ***Usr-Sys*** and ***Usr-r***, were therefore designed to be basic and simple (as described in **Section 4.1.2**), and individual interaction events were triggered at rigid intervals. In realistic applications, there may be a variety of mixed-initiative patterns and temporal structures. For example, the user needs to attend to and interact with temporal structures that are more complicated and flexible, such as the prosody in natural speech and the dynamic runtime of a back-end machine learning algorithm. Therefore, more types of temporal pattern should be designed and evaluated to accommodate the specific timing characteristics of a given interaction scenario.

- Third, it is also not guaranteed or confirmed that the four conditions, ***Sys-ii***, ***Sys-pr***, ***Usr-Sys*** and ***Usr-r***, are equally spaced on the mixed-initiative spectrum. Indeed, the results of the contrast analysis confirmed significant "↗"/"↘", "↗↘" or "↗↘↗" trends in different sets of data, where the results in ***Sys-ii*** and ***Usr-r*** were always at the two far ends, with the results in ***Sys-pr*** and ***Usr-Sys*** interpolated in between, but not every pairs of conditions were confirmed to be significantly different after the alpha level was reduced to $0.05/6 = 0.0083$ using the *Bonferroni* correction, hence the research findings and design implications in this dissertation should be interpreted in their context with due caution, and be generalised with limited confidence. Nevertheless, as discussed in **Sections 4.3.2**, **5.3.2** and **6.4.2**, the *Bonferroni* correction might have led to more Type II errors and wrongly rejected valid statistical inferences (Perneger, 1998; Furr & Rosenthal, 2003; Nakagawa, 2004; Abdi & Williams, 2010; Armstrong, 2014), especially when the pairwise tests in this dissertation were *pre-planned* under sound theoretical *hypotheses* to investigate *dependent* observations, as opposed to running a large number of *unplanned independent* tests without *hypotheses* (Armstrong, 2014).

- Fourth, the number of participants recruited for the experiments was relatively small. Although the sample size was enough to reveal statistical significance, it might still cause a higher margin of error and thus affect the power of the results. In addition, the participants recruited for the experiments were cognitively competent and physically able. When users are from a population that comprises a wider range of capabilities, such as age, mobility and expertise, the effects of timing on the user's perceived control may be less than homogeneous.

7.4 Directions for future study

1. **Temporal irregularity vs. system performance**

As discussed in **Section 6.4.2**, all initial labels in Experiment 3 were randomly preassigned to the messages, therefore participants did not see much improvement in the system's performance over time. An easy step forward is to investigate whether or not the improvement of the system's performance can mitigate the negative effects of temporal irregularities on the user's perceived control.

An initial study in this direction can still be conducted in the context of AI-assisted labelling. We can either use a real machine learning algorithm behind the interface or use the Wizard of Oz paradigm again to create the impression of performance improvement. Either way, the level of improvement should be controlled and appear to happen at a comparable rate. In order to gain further insights into participants' experience and impression, qualitative methods such as post study interviews, the "think-aloud" protocol or coding participants' facial expression in videos should be considered.

The potential benefits and implications of such an investigation are twofold. Firstly, it can reveal whether or not the user can ascribe authorship to the system's improvement (i.e. "the system is doing better because of my coaching"), and whether or not the perceived authorship can make their perceived control more robust when faced with temporal irregularities. Secondly, it can inform designers how the user would perceive the system's performance improvement and when they would like to pass the initiative to the system given the perceived improvement.

2. **Rhythm-oriented design in end-user programming**

   In the three experiments, the tasks were adapted from a conventional stimulus-response paradigm where certain interaction events were designed to occur *before* the timing of the occurrences was manipulated. In other words, the rhythm was forced onto interaction events retrospectively. However, as discussed in **Section 4.3.2**, during the mixed-initiative interaction with an intelligent system, such as a Programming-by-Example application or an end-user programming tool that exhibits a certain level of "liveness" (Tanimoto, 2013) (as summarised in Table 2.3), the timing of one loop of interaction is often accidental, because it can be largely determined by how long it will take the machine learning based artificial intelligence algorithms behind the interface to run. Hence, it can be problematic to impose a pattern on merely the presentation of outcomes afterwards.

   One possible way to resolve this issue is to incorporate the design of timing as part of the design of the end-user programming tools. For example, for a given mixed-initiative interaction scenario, developers can first determine the appropriate timing characteristics and build a model for them, then design the machine learning algorithms the execution and output cycles of which can be mapped on to the temporal model. There should also be a timekeeping component that can monitor the fluctuation in time whilst supervising and informing the execution and the output of the system.

According to the definition of rhythm proposed in **Section 2.3.1**, rhythm can be a systematic patterning of *events* in terms of timing, accent, and grouping (adapted from Patel (2010)'s definition). Therefore, when building a temporal model, we can also take the accent and grouping into account.

The *grouping* effect discussed in **Section 5.3.1** is perhaps the easiest temporal model to be implemented and investigated. For example, in a Programming-by-Example application, we can compare the user's perceived control when: 1) the system emulates right after every new demonstration, or 2) the system starts to emulate after a group of two, three, or more user demonstrations, or 3) other dynamic grouping patterns.

Furthermore, considering that the amount of information conveyed by individual interaction events often varies in a realistic context, as discussed in **Sections 4.3.2** and **5.3.2**, we can characterise the amount of information or the level of task difficulty as the *accents* in mixed-initiative interaction, where a task with a large amount of information or high difficulty is more accentuated and requires more attention from the user at appropriate times. In addition, given the probabilistic nature of machine learning algorithms, every judgement made by the system has its confidence value, hence the ones with a low confidence value should be accentuated to the user following a designed temporal structure.

The resulting insights of the investigations suggested above can be beneficial in three ways. Firstly, they can give the design of interaction rhythm more flexibility, and we can try more combinations of rhythmic characteristics. Secondly, we can better accommodate the temporal structures for both the evocation and the content of events in a given interaction scenario. Third, they allow us to explore the appropriate interaction rhythm that can help the user better distribute their attention whilst keeping them engaged in the interaction.

3. **Rhythmic entrainment in cross-modal interaction**

As noted in the two design implications proposed in **Sections 5.3.1** and **6.4.1**, the effects of timing on the user's perceived control and task experience remain congruent in different modalities, hence there is the potential to manipulate the timing of cross-modal mixed-initiative interaction. For example, when the user is interacting with a semi-autonomous vehicle navigation system, the interaction may require the user to react to auditory prompts whilst visually attending to the environment and promptly activating their motor functions to turn a steering wheel or step on a brake pedal. Hence it is worthwhile exploring the effects of timing across different modalities and design the interaction surrounding the temporal structures.

A further suggestion is to employ novel interaction modalities. For instance, electromyography (EMG)-based human-computer interface (Barreto, Scargle, & Adjouadi, 2000; Cowley et al., 2016) can be considered, as EMG can be measured with unintrusive devices such as an armband (e.g. Myo [2])(Mulling & Sathiyanarayanan, 2015). Given that 1) EMG measures can detect a person's motor preparation and relaxation states (Rider, 1985; Chappell, Creighton, Giuliani, Yu, & Garrett, 2007), 2) a person's motor movements can entrain with external temporal structures (Repp, 2005; Repp & Su, 2013), which can help them develop strategies for anticipatory control (Knoblich & Jordan, 2003), and 3) interpersonal sensorimotor synchronisation can be beneficial in social interaction (Richardson et al., 2005; Clayton et al., 2005), EMG-based modality may work well with conventional modalities and enhance the user's experience of control if the temporal structures of the interaction are designed towards rhythmic entrainment.

An initial investigation along this direction could involve letting the user wear an EMG armband while driving a semi-automated vehicle. The system could take into account their muscle relaxation or tension states and adjust the timing of passing the manual control to the user more appropriately with well-timed auditory or visual prompts. We can expect three benefits from such an investigation. Firstly, it could reveal the user's general motion pattern when driving a semi-automated vehicle. Secondly, it could explore what kind of timing characteristics can facilitate the co-ordination between the user's motor function and their visual/auditory attention. And thirdly, it could potentially help designers

[2] The Myo armband (https://www.myo.com) is a commercialised wearable gesture control and motion control device, which operates based on the electromyography signal in the user's upper arm.

time the initiative transfer when the user's motor state is ready, thereby reducing the abruptness caused by a sudden handover of the control.

7.5 Closing remarks

Timing plays an important role in all kinds of social and human-computer interaction. It not only describes the arrangement of interaction events along a timeline, but also serves an active co-ordinating role that affects both the quality and outcome of an interaction. Therefore, in mixed-initiative interaction where the locus of control is repeatedly transferred between the user and the intelligent system, the timing of initiative-taking becomes a key issue in interaction design.

In this dissertation, I have demonstrated that the timing of mixed-initiate interaction *can* be and *should* be manipulated as a design resource. Based on the agency theories in cognitive psychology and the rhythmic entrainment theories in social psychology and cognitive neuroscience, I hypothesised that in mixed-initiative interaction, a predictable interaction rhythm can preserve the user's sense of agency, improve their confidence in task performance and facilitate their entrainment behaviours while lowering their perceived level of stress and effort, while irregular interaction timing can cause the opposite effects. I designed and carried out three controlled experiments, the results of which have provided empirical evidence to support my hypotheses. Furthermore, the experiments complemented each other, hence my findings were consolidated in different modalities (e.g. visual and auditory) and were validated in a realistic context (e.g. AI-assisted labelling).

While the delicacy of our agency experience and the complexity of realistic task requirements pose challenges to mixed-initiative interaction design, we should further leverage both the function and the flexibility of timing in future research, and find solutions that can not only allow the user to enjoy the merits of end-user automation, but also sustain their perceived control.

References

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and cognition*, *14*(3), 439–458. doi: 10.1016/j.concog.2004.11.001

Abdi, H., & Williams, L. J. (2010). Contrast analysis. *Encyclopedia of research design*, *1*, 243–251. doi: 10.4135/9781412961288.n75

Afzal, S., & Robinson, P. (2014). Emotion data collection and its implications for affective computing. In R. A. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), *The oxford handbook of affective computing* (pp. 359–369). Oxford University Press, UK.

Akamatsu, M., & MacKenzie, I. S. (1996). Movement characteristics using a mouse with tactile and force feedback. *International Journal of Human-Computer Studies*, *45*(4), 483–493. doi: 10.1006/ijhc.1996.0063

Akamatsu, M., MacKenzie, I. S., & Hasbroucq, T. (1995). A comparison of tactile, auditory, and visual feedback in a pointing task using a mouse-type device. *Ergonomics*, *38*(4), 816–827. doi: 10.1080/00140139508925152

André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Integrating models of personality and emotions into lifelike characters. *Lecture notes in computer science (Affective interactions: Towards a new generation of computer interfaces)*, 150–165. doi: 10.1007/10720296_11

Armstrong, R. A. (2014). When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508. doi: 10.1111/opo.12131

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in cognitive sciences*, *16*(7), 390–398. doi: 10.1016/j.tics.2012.05.003

Auer, P., Couper-Kuhlen, E., & Müller, F. (1999). *Language in time: The rhythm and* *tempo of spoken interaction.* Oxford University Press on Demand.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, *29*(7), 819–833. doi: 10.1177/0146167203029007002

Balkwell, J. W. (1991). From expectations to behavior: An improved postulate for expectation states theory. *American Sociological Review*, 355–369. doi: 10.2307/2096109

Banerjee, S. P., & Woodard, D. L. (2012). Biometric authentication and identification using keystroke dynamics: A survey. *Journal of Pattern Recognition Research*, *7*(1), 116–139. doi: 10.13176/11.427

Barnes, G., & Asselman, P. (1991). The mechanism of prediction in human smooth pursuit eye movements. *The Journal of Physiology*, *439*(1), 439–461. doi: 10.1113/jphysiol.1991.sp018675

Barnes, G., Collins, C., & Arnold, L. (2005). Predicting the duration of ocular pursuit in humans. *Experimental brain research*, *160*(1), 10–21. doi: 10.1007/s00221-004-1981-3

Baronas, A.-M. K., & Louis, M. R. (1988). Restoring a sense of control during implementation: how user involvement leads to system acceptance. *Mis Quarterly*, 111–124. doi: 10.2307/248811

Barreto, A. B., Scargle, S. D., & Adjouadi, M. (2000). A practical emg-based human-computer interface for users with motor disabilities. *Journal of rehabilitation research and development*, *37*(1), 53–63. Retrieved from `https://search.proquest.com/docview/215293187?accountid=9851`

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type. In *Proceedings of the 15th international congress of phonetic sciences (ICPhS'03)* (pp. 2693–2696). Retrieved from `https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2693.pdf`

Bauer, M., Dengler, D., & Paul, G. (2001). Trainable information agents for the web. In H. Lieberman (Ed.), *Your wish is my command: Giving users the power to instruct their software* (pp. 87–114). Morgan Kaufmann. doi: 10.1016/B978-155860688-3/50006-3

Baylor, A. (2000). Beyond butlers: Intelligent agents as mentors. *Journal* *of Educational Computing Research*, *22*(4), 373–382. Retrieved from https://www.researchgate.net/profile/Amy_Baylor/publication/245584164_Beyond_Butlers_Intelligent_Agents_as_Mentors/links/53f488060cf2fceacc6e895e.pdf

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, *75*(299), 493–517. doi: 10.2307/2228949

Bennett, P. N., Chickering, D. M., & Mityagin, A. (2009). Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th international conference on world wide web (WWW'09)* (pp. 121–130). ACM Press. doi: 10.1145/1526709.1526727

Berger, J., & Conner, T. L. (1969). Performance expectations and behavior in small groups. *Acta Sociologica*, *12*(4), 186–198. Retrieved from http://www.jstor.org/stable/4193723

Bernardi, L., Porta, C., Casucci, G., Balsamo, R., Bernardi, N. F., Fogari, R., & Sleight, P. (2009). Dynamic interactions between musical, cardiovascular, and cerebral rhythms in humans. *Circulation*, *119*(25), 3171–3180. doi: 10.1161/CIRCULATIONAHA.108.806174

Berthaut, F., Coyle, D., Moore, J. W., & Limerick, H. (2015). Liveness through the lens of agency and causality. In *Proceedings of the 15th international conference on new interfaces for musical expression (NIME'15)* (pp. 76–79). Retrieved from http://hdl.handle.net/10197/6635

Biocca, F., Inoue, Y., Lee, A., Polinsky, H., & Tang, A. (2002). Visual cues and virtual touch: Role of visual stimuli and intersensory integration in cross-modal haptic illusions and the sense of presence. *Proceedings of presence*, 410–428. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.482.9450&rep=rep1&type=pdf

Blackwell, A. F. (2015). Interacting with an inferred world: the challenge of machine learning for *humane* computer interaction. In *Proceedings of the 5th Decennial Aarhus conference on critical alternatives (AA'15)* (pp. 169–180). Aarhus, Denmark: Aarhus University Press. doi: 10.7146/aahcc.v1i1.21197

Blackwell, A. F. (2017). *Introducing CODA: A tool for data analysis.* Retrieved 2017-08-25, from http://www.africasvoices.org/ideas/newsblog/introducing-our-latest-analysis-tool-coda/

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in cognitive sciences*, *6*(6), 237–242. doi: 10.1016/S1364-6613(02)01907-1

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, *7*(3), 338-355. doi: 10.1037/1082-989X.7.3.338

Bonito, J. A., Burgoon, J. K., & Bengtsson, B. (1999). The role of expectations in human-computer interaction. In *Proceedings of the international ACM SIGGROUP conference on supporting group work (GROUP'99)* (pp. 229–238). ACM Press. doi: 10.1145/320297.320324

Boucugnani, L. L., & Jones, R. W. (1989). Behaviors analogous to frontal lobe dysfunction in children with attention deficit hyperactivity disorder. *Archives of Clinical Neuropsychology*, *4*(2), 161–173. doi: doi.org/10.1093/arclin/4.2.161

Bratman, M. (1999). *Faces of intention: Selected essays on intention and agency.* Cambridge, UK: Cambridge University Press.

Breazeal, C. (2002). Regulation and entrainment in human-robot interaction. *The International Journal of Robotics Research*, *21*(10-11), 883–902. doi: 10.1177/027836402321261814

Brodley, C. E., Rebbapragada, U., Small, K., & Wallace, B. (2012). Challenges and opportunities in applied machine learning. *AI Magazine*, *33*(1), 11–24. doi: 10.1609/aimag.v33i1.2367

Brusky, K. J., Frederick, J. W., & Lininger, J. T. (1999, May 11). *Method and apparatus for mapping remote control buttons onto keyboard stroke combinations.* Google Patents. (US Patent 5,903,259)

Buell, R. W., & Norton, M. I. (2011). The labor illusion: How operational transparency increases perceived value. *Management Science*, *57*(9), 1564–1579. doi: 10.1287/mnsc.1110.1376

Burgoon, J. K. (1978). A communication model of personal space violations: Explication and an initial test. *Human Communication Research*, *4*(2), 129–142. doi: 10.1111/j.1468-2958.1978.tb00603.x

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology*, *12*(1-2), 30–48. doi:

10.1177/0261927X93121003

Burgoon, J. K. (1995). Cross-cultural and intercultural applications of expectancy violations theory. In R. L. Wiseman (Ed.), *Intercultural communication theory (vol. 19)* (pp. 194–214). SAGE Publications Sage CA: Thousand Oaks, CA.

Burgoon, J. K., & Poire, L. (1993). Effects of communication expectancies, actual communication, and expectancy disconfirmation on evaluations of communicators and their communication behavior. *Human communication research*, *20*(1), 67–96. doi: 10.1111/j.1468-2958.1993.tb00316.x

Burgoon, J. K., White, C. H., & Greene, J. (1997). Researching nonverbal message production: A view from interaction adaptation theory. *Message production: Advances in communication theory*, 279–312.

Carruthers, P. (2007). The illusion of conscious will. *Synthese*, *159*(2), 197–213. doi: 10.1007/s11229-007-9204-7

Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. *Communications of the ACM*, *59*(5), 70–77. doi: 10.1145/2830565

Chang, K. S.-P., & Myers, B. A. (2014). Creating interactive web data applications with spreadsheets. In *Proceedings of the 27th annual ACM symposium on user interface software and technology (UIST'14)* (pp. 87–96). ACM Press. doi: 10.1145/2642918.2647371

Chappell, J. D., Creighton, R. A., Giuliani, C., Yu, B., & Garrett, W. E. (2007). Kinematics and electromyography of landing preparation in vertical stop-jump: risks for noncontact anterior cruciate ligament injury. *The American journal of sports medicine*, *35*(2), 235–241. doi: 10.1177/0363546506294077

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, *76*(6), 893–910. doi: 10.1037/0022-3514.76.6.893

Chertoff, D. B., Byers, R. W., & LaViola Jr, J. J. (2009). An exploration of menu techniques using a 3d game input device. In *Proceedings of the 4th international conference on foundations of digital games (FDG'09)* (pp. 256–262). ACM Press. doi: 10.1145/1536513.1536559

Choi, H., & Scholl, B. J. (2006). Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception. *Perception*, *35*(3), 385–399. doi:

10.1068/p5462

Church, L., Nash, C., & Blackwell, A. F. (2010). Liveness in notation use: From music to programming. In *Proceedings of the 22nd annual workshop of the psychology of programming interest group (PPIG'10)* (pp. 2–11). Retrieved from `http://www.ppig.org/papers/22nd-UX-1.pdf`

Clarke, A. A., & Smyth, M. G. G. (1993). A co-operative computer based on the principles of human co-operation. *International Journal of Man-machine studies*, *38*(1), 3–22. doi: 10.1006/imms.1993.1002

Clayton, M. (2012). What is entrainment? definition and applications in musical research. *Empirical Musicology Review*, *7*(1–2), 49–56. doi: 10.18061/1811/52979

Clayton, M., Sager, R., & Will, U. (2005). In time with the music: the concept of entrainment and its significance for ethnomusicology. In *European meetings in ethnomusicology* (Vol. 11, pp. 1–82). Romanian Society for Ethnomusicology. Retrieved from `http://oro.open.ac.uk/2661/1/InTimeWithTheMusic.pdf`

Collins, H., & Kusch, M. (1999). *The shape of actions: What humans and machines can do.* MIT Press.

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction* (Vol. 25). John Benjamins Publishing.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., ... Jacuccii, G. (2016). The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends® in Human–Computer Interaction*, *9*(3-4), 151–308. doi: 10.1561/1100000065

Coyle, D., Moore, J. W., Kristensson, P. O., Fletcher, P., & Blackwell, A. F. (2012). I did that! measuring users' experience of agency in their own actions. In *Proceedings of the 30th SIGCHI conference on human factors in computing systems (CHI'12)* (pp. 2025–2034). ACM Press. doi: 10.1145/2207676.2208350

Cross, I. (2013). "Does not compute"? Music as real-time communicative interaction. *AI & society*, *28*(4), 415–430. doi: 10.1007/s00146-013-0511-x

Cypher, A. (1995). Eager: Programming repetitive tasks by example. In R. M. Baecker (Ed.), *Readings in human-computer interaction (toward the year 2000)* (pp. 804–810). Elsevier. doi: 10.1016/B978-0-08-051574-8.50083-2

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies - why and how. *Knowledge-based systems*, *6*(4), 258–266.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

David, A. (1967). *Elements of general phonetics.* Edinburgh: Edinburgh University Press.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85–103. Retrieved from `https://www.uv.es/friasnav/Davis_1980.pdf`

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, *44*(1), 113–126. doi: 10.1037/0022-3514.44.1.113

Deaton, J. E., & Parasuraman, R. (1993). Sensory and cognitive vigilance: Effects of age on performance and subjective workload. *Human Performance*, *6*(1), 71–97. doi: 10.1207/s15327043hup0601_4

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009 (CVPR'09). The proceedings of the IEEE computer society conference on* (pp. 248–255). doi: 10.1109/CVPR.2009.5206848

Dennett, D. C. (1989). *The intentional stance.* MIT press.

Desmarais, M. C., Giroux, L., & Larochelle, S. (1993). An advice-giving interface based on plan-recognition and user-knowledge assessment. *International Journal of Man-Machine Studies*, *39*(6), 901–924. doi: 10.1006/imms.1993.1089

Drake, C., & Botte, M.-C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics*, *54*(3), 277–286. doi: /10.3758/BF03205262

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, *54*(4), 585–612. doi: 10.1006/ijhc.2001.0458

Ebert, J. P., & Wegner, D. M. (2010). Time warp: Authorship shapes the perceived timing of actions and events. *Consciousness and cognition*, *19*(1), 481–489. doi: 10.1016/j.concog.2009.10.002

Engbert, K., Wohlschläger, A., Thomas, R., & Haggard, P. (2007). Agency, subjective time, and other minds. *Journal of Experimental Psychology: Human Perception* *and Performance*, *33*(6), 1261–1268. doi: 10.1037/0096-1523.33.6.1261

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the 29th SIGCHI conference on human factors in computing systems (CHI'11)* (pp. 715–724). ACM Press. doi: 10.1145/1978942.1979046

Fails, J. A., & Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th international conference on intelligent user interfaces (IUI'03)* (pp. 39–45). ACM Press. doi: 10.1145/604045.604056

Faratin, P., Sierra, C., & Jennings, N. R. (1998). Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, *24*(3-4), 159–182. doi: 10.1016/S0921-8890(98)00029-3

Farrer, C., Bouchereau, M., Jeannerod, M., & Franck, N. (2008). Effect of distorted visual feedback on the sense of agency. *Behavioural neurology*, *19*(1, 2), 53–57. doi: 10.1155/2008/425267

Fişek, M. H., Berger, J., & Norman, R. Z. (1991). Participation in heterogeneous and homogeneous groups: A theoretical integration. *American Journal of Sociology*, *97*(1), 114–142. doi: 10.1086/229742

Fişek, M. H., Berger, J., & Norman, R. Z. (1995). Evaluations and the formation of expectations. *American Journal of Sociology*, *101*(3), 721–746. doi: 10.1086/230758

Fogg, B. J. (1998). Persuasive computers: perspectives and research directions. In *Proceedings of the 16th SIGCHI conference on human factors in computing systems (CHI'98)* (pp. 225–232). ACM Press/Addison-Wesley Publishing Co. doi: 10.1145/274644.274677

Fogg, B. J. (2002). Persuasive technology: using computers to change what we think and do (computers as persuasive social actors). *Ubiquity*, *2002*(December), 89–120. doi: 10.1145/764008.763957

Francese, R., Passero, I., & Tortora, G. (2012). Wiimote and Kinect: gestural user interfaces add a natural third dimension to HCI. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 116–123). ACM Press. doi: 10.1145/2254556.2254580

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *International workshop on agent theories, architectures,* *and languages* (pp. 21–35). doi: 10.1007/BFb0013570

Fujioka, T., Trainor, L. J., Large, E. W., & Ross, B. (2009). Beta and gamma rhythms in human auditory cortex during musical beat processing. *Annals of the New York Academy of Sciences*, *1169*(1), 89–92. doi: 10.1111/j.1749-6632.2009.04779.x

Furr, R. M., & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1), 33–67. doi: 10.1207/S15328031US0201_03

Gainotti, G., Silveri, M. C., Daniel, A., & Giustolisi, L. (1995). Neuroanatomical correlates of category-specific semantic disorders: A critical survey. *Memory*, *3*(3-4), 247–263. doi: 10.1080/09658219508253153

Gallagher, S. (2007). Sense of agency and higher-order cognition: Levels of explanation for schizophrenia. *Cognitive Semiotics*, *1*, 33–48. Retrieved from https://s3.amazonaws.com/academia.edu.documents/3439182/What_Do_Weather_Watchers_See_Perceptual_Intentionality_and_Agency.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1520334500&Signature=GbNwi244g65BGigDGLzZubX7bas%3D&response-content-disposition=inline%3B%20filename%3DWhat_Do_Weather_Watchers_See_Perceptual.pdf#page=32

Gallese, V. (2001). The 'shared manifold' hypothesis: From mirror neurons to empathy. *Journal of consciousness studies*, *8*(5-6), 33–50. Retrieved from https://pdfs.semanticscholar.org/2b11/cbfdf73a3bb5b9fdf588365aafb2a4b4c875.pdf

Gallese, V. (2003). The manifold nature of interpersonal relations: the quest for a common mechanism. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1431), 517–528. doi: 10.1098/rstb.2002.1234

Gallotti, M., Fairhurst, M., & Frith, C. (2017). Alignment in social interactions. *Consciousness and cognition*, *48*, 253–261. doi: 10.1016/j.concog.2016.12.002

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181–193. Retrieved from http://www.jstor.org/stable/40212196

Gill, S. P. (2012). Rhythmic synchrony and mediated interaction: towards a framework of rhythm in embodied interaction. *AI & society*, *27*(1), 111–127. doi: 10.1007/s00146-011-0362-2

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary Journal*, *26*(5), 427–438. doi: 10.1525/mp.2009.26.5.427

Goebl, W., & Parncutt, R. (2002). The influence of relative intensity on the perception of onset asynchronies. In *Proceedings of the 7th international conference on music perception and cognition (ICMPC7)* (pp. 1–4). Adelaide, S. Australia: Causal Productions. Retrieved from `https://pdfs.semanticscholar.org/15db/09f4fba8ff503dd1d3605f47001d7e276235.pdf`

Gorn, G. J., Chattopadhyay, A., Sengupta, J., & Tripathi, S. (2004). Waiting for the web: how screen color affects time perception. *Journal of marketing research*, *41*(2), 215–225. doi: 10.1509/jmkr.41.2.215.28668

Greatrex, D. (2018). *Effects of temporal expectation on complex decision making* (Unpublished doctoral dissertation). University of Cambridge.

Grossman, T., & Balakrishnan, R. (2005). The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of the 23rd SIGCHI conference on human factors in computing systems (CHI'05)* (pp. 281–290). doi: 10.1145/1054972.1055012

Guadagno, R. E., Blascovich, J., Bailenson, J. N., & Mccall, C. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, *10*(1), 1–22. doi: 10.108/15213260701300865

Guadagno, R. E., & Cialdini, R. B. (2002). Online persuasion: An examination of gender differences in computer-mediated interpersonal influence. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 38–51. doi: 10.1037/1089-2699.6.1.38

Haans, A. (2018). Contrast analysis: A tutorial. *Practical Assessment Research & Evaluation*, *23*(9), 1–21. Retrieved from `http://pareonline.net/getvn.asp?v=23&n=9`

Haggard, P., Aschersleben, G., Gehrke, J., & Prinz, W. (2002). Action, binding, and awareness. Oxford University Press.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature neuroscience*, *5*(4), 382–385. doi: 10.1038/nn827

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental brain research*, *126*(1), 128–133. doi: 10.1007/s002210050722

Hardian, B. (2006). Middleware support for transparency and user control in context-aware systems. In *Proceedings of the 3rd international middleware doctoral symposium (MDS'06)* (pp. 4–9). ACM Press. doi: 10.1145/1169100.1169104

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 (UIST'07)* (pp. 115–118). ACM Press. doi: 10.1145/1294211.1294231

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hawkins, S., Cross, I., & Ogden, R. (2013). Communicative interaction in spontaneous music and speech. *Music, language and interaction*, 285–329.

Himberg, T. (2006). Co-operative tapping and collective time-keeping–differences of timing accuracy in duet performance with human or computer partner. In *Proceedings of the 9th international conference on music perception and cognition (ICMPC'06).*

Hoffman, G., & Vanunu, K. (2013). Effects of robotic companionship on music enjoyment and agent perception. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* (pp. 317–324). IEEE Press. doi: http://guyhoffman.com/publications/HoffmanHRI13.pdf

Hoffman, R. E. (1986). Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences*, *9*(3), 503–517. doi: 10.1017/S0140525X00046781

Hofstede, G. (1984). Cultural dimensions in management and planning. *Asia Pacific journal of management*, *1*(2), 81–99. doi: 10.1007/BF01733682

Hoggan, E., Crossan, A., Brewster, S. A., & Kaaresoja, T. (2009). Audio or tactile feedback: which modality when? In *Proceedings of the 27th SIGCHI conference on human factors in computing systems (CHI'09)* (pp. 2253–2256). ACM Press. doi: 10.1145/1518701.1519045

Hon, N., Poh, J.-H., & Soon, C.-S. (2013). Preoccupied minds feel less control: Sense of agency is modulated by cognitive load. *Consciousness and cognition*, *22*(2), 556–561. doi: 10.1016/j.concog.2013.03.004

Horvitz, E. (1999a). Principles of mixed-initiative user interfaces. In *Proceedings of the 17th SIGCHI conference on human factors in computing systems (CHI'99)* (pp. 159–166). ACM Press. doi: 10.1145/302979.303030

Horvitz, E. (1999b). Uncertainty, action, and interaction: In pursuit of mixed-initiative computing. *IEEE Intelligent Systems*, *14*(5), 17–20. Retrieved from https://pdfs.semanticscholar.org/3f45/208609a0ae942ae65a0c3b1100247ec92655.pdf

Horvitz, E., & Barry, M. (1995). Display of information for time-critical decision making. In *Proceedings of the 11th conference on uncertainty in artificial intelligence (UAI'95)* (pp. 296–305). Retrieved from https://arxiv.org/pdf/1302.4959.pdf

Hove, M. J., & Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, *27*(6), 949–960. doi: 10.1521/soco.2009.27.6.949

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological bulletin*, *139*(1), 133–151. doi: 10.1037/a0028566

Hutchinson, T. E., White, K. P., Martin, W. N., Reichert, K. C., & Frey, L. A. (1989). Human-computer interaction using eye-gaze input. *IEEE Transactions on systems, man, and cybernetics*, *19*(6), 1527–1534. doi: 10.1109/21.44068

Hyrskykari, A., Istance, H., & Vickers, S. (2012). Gaze gestures or dwell-based interaction? In *Proceedings of the 2012 symposium on eye tracking research and applications (ETRA 2012)* (pp. 229–232). ACM Press. doi: 10.1145/2168556.2168602

Iacovides, I., Cox, A., Kennedy, R., Cairns, P., & Jennett, C. (2015). Removing the HUD: the impact of non-diegetic game elements and expertise on player involvement. In *Proceedings of the 2015 annual symposium on computer-human interaction in play (CHI PLAY'15)* (pp. 13–22). ACM Press. doi: 10.1145/2793107.2793120

Iio, T., Shiomi, M., Shinozawa, K., Akimoto, T., Shimohara, K., & Hagita, N. (2011). Investigating entrainment of people's pointing gestures by robot's gestures using a woz method. *International Journal of Social Robotics*, *3*(4), 405–414. doi:

10.1007/s12369-011-0112-0

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2013). Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM international conference on multimodal interaction (ICMI'13)* (pp. 181–188). ACM Press. doi: 10.1145/2522848.2522890

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2014). Micro-timing of backchannels in human-robot interaction. *Timing in Interaction, Workshop in Conjunction with Human-Robot Interaction (HRI) 2014*. Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1007.2108&rep=rep1&type=pdf`

Irani, L. C., & Silberman, M. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the 31st SIGCHI conference on human factors in computing systems (CHI'13)* (pp. 611–620). ACM Press. doi: 10.1145/2470654.2470742

Ishii, H., & Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the 15th SIGCHI conference on human factors in computing systems (CHI'97)* (pp. 234–241). ACM Press. doi: 10.1145/258549.258715

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1995). A history of facilitated communication: Science, pseudoscience, and antiscience science working group on facilitated communication. *American Psychologist*, *50*(9), 750.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological review*, *96*(3), 459–491. doi: 10.1037/0033-295X.96.3.459

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on computer supported cooperative work (CSCW'13)* (pp. 1555–1566). ACM Press. doi: 10.1145/2441776.2441954

Kamalian, R., Yeh, E., Zhang, Y., Agogino, A. M., & Takagi, H. (2006). Reducing human fatigue in interactive evolutionary computation through fuzzy systems and machine learning systems. In *Fuzzy systems, 2006. The proceedings of IEEE international conference on* (pp. 678–684). doi: 10.1109/FUZZY.2006.1681784

Kang, P., Park, S., Hwang, S.-s., Lee, H.-j., & Cho, S. (2008). Improvement of keystroke data quality through artificial rhythms and cues. *Computers & Security*, *27*(1), 3–11. doi: 10.1016/j.cose.2008.02.001

Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, *11*(2), 1565–1573. doi: 10.1016/j.asoc.2010.08.003

Kätsyri, J., Hari, R., Ravaja, N., & Nummenmaa, L. (2013). The opponent matters: elevated fMRI reward responses to winning against a human versus a computer opponent during interactive video game playing. *Cerebral Cortex*, *23*(12), 2829–2839. doi: 10.1093/cercor/bhs259

Keller, P. E., Knoblich, G., & Repp, B. H. (2007). Pianists duet better when they play with themselves: on the possible role of action simulation in synchronization. *Consciousness and cognition*, *16*(1), 102–111. doi: 10.1016/j.concog.2005.12.004

Keller, P. E., Novembre, G., & Hove, M. J. (2014). Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130394. doi: 10.1098/rstb.2013.0394

Klingspor, V., Demiris, J., & Kaiser, M. (1997). Human-robot communication and machine learning. *Applied Artificial Intelligence*, *11*(7), 719–746. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.8753&rep=rep1&type=pdf

Kneutgen, J. (1970). The biological function of a category of music: On the effect of lullabies. *Zeitschrift für experimentelle und angewandte Psychologie*, *17*(2), 245–265.

Knoblich, G., & Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 1006–1016. doi: 10.1037/0278-7393.29.5.1006

Kołakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *Proceedings of the 6th international conference on human system interaction (HSI'13)* (pp. 548–555). doi: 10.1109/HSI.2013.6577879

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus-response compatibility–A model and taxonomy. *Psychological review*, *97*(2), 253–270. doi: 10.1037/0033-295X.97.2.253

Kulesza, T., Amershi, S., Caruana, R., Fisher, D., & Charles, D. (2014). Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the 32nd SIGCHI conference on human factors in computing systems (CHI'14)* (pp. 3075–3084). ACM Press. doi: 10.1145/2556288.2557238

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces (IUI'15)* (pp. 126–137). ACM Press. doi: 10.1145/2678025.2701399

Kwon, B. c., Javed, W., Elmqvist, N., & Yi, J. S. (2011). Direct manipulation through surrogate objects. In *Proceedings of the 29th SIGCHI conference on human factors in computing systems (CHI'11)* (pp. 627–636). ACM Press. doi: 10.1145/1978942.1979033

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–113. doi: 10.1126/science.1154735

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological review*, *106*(1), 119–159. doi: 10.1037/0033-295X.106.1.119

Large, E. W., & Kolen, J. F. (1994). Resonance and the perception of musical meter. *Connection science*, *6*(2–3), 177–208. doi: 10.1080/09540099408915723

Lee, Y., Chen, A. N., & Ilie, V. (2012). Can online wait be managed? the effect of filler interfaces and presentation modes on perceived waiting time online. *MIS Quarterly*, *36*(2), 365–394. Retrieved from http://www.jstor.org/stable/41703460

Leman, M. (2012). Musical entrainment subsumes bodily gestures: its definition needs a spatiotemporal dimension. *Empirical Musicology Review*, *7*(1–2), 63–67. doi: 10.18061/1811/52981

Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67. doi: 10.1080/21548331.1967.11707799

Levitan, R., Gravano, A., & Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers-volume 2 (ACL/HLT'11)* (pp. 113–117). Retrieved from http://www.aclweb.org/

anthology/P11-2020

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of the 12th annual conference of the international speech communication association (INTERSPEECH'11).* International Speech Communication Association (ISCA). Retrieved from https://core.ac.uk/download/pdf/27296882.pdf

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) the unconscious initiation of a freely voluntary act. *Brain*, *106*(3), 623–642. doi: 10.1093/brain/106.3.623

Lieberman, H. (2000). *Your wish is my command: Giving users the power to instruct their software.* Morgan Kaufmann.

Limerick, H., Moore, J. W., & Coyle, D. (2015). Empirical evidence for a diminished sense of agency in speech interfaces. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI'15)* (pp. 3967–3970). ACM Press. doi: 10.1145/2702123.2702379

London, J. (2012a). *Hearing in time: Psychological aspects of musical meter.* Oxford University Press.

London, J. (2012b). Three things linguists need to know about rhythm and time in music. *Empirical Musicology Review*, *7*(1–2), 5–11. doi: 10.18061/1811/52973

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999 (ICCV'99). the proceedings of the 7th IEEE international conference on* (Vol. 2, pp. 1150–1157). doi: 10.1109/ICCV.1999.790410

Macrae, C. N., Duffy, O. K., Miles, L. K., & Lawrence, J. (2008). A case of hand waving: Action synchrony and person perception. *Cognition*, *109*(1), 152–156. doi: 10.1016/j.cognition.2008.07.007

Madison, J. (2012). *Damn you, autocorrect!* Random House.

Marr, D., & Vision, A. (1982). A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, *1*(2). Retrieved from http://www.contrib.andrew.cmu.edu/~kk3n/80-300/marr2.pdf

Martínez-Miranda, J., Bresó, A., & García-Gómez, J. M. (2012). Modelling therapeutic empathy in a virtual agent to support the remote treatment of major depression.

In *Proceedings of the 4th international conference on agents and artificial intelligence (ICAART'12)* (pp. 264–269). Science and Technology Publications. doi: 10.5220/0003833302640269

McCann, H. (1998). *The works of agency: On human action, will, and freedom.* Ithaca, NY, USA: Cornell University Press.

McGrath, J. E., & Kelly, J. R. (1986). *Time and human interaction: Toward a social psychology of time.* Guilford Press.

Medina, J. F., Carey, M. R., & Lisberger, S. G. (2005). The representation of time for motor learning. *Neuron*, *45*(1), 157–167. doi: 10.1016/j.neuron.2004.12.017

Mehdi, E. J., Nico, P., Dugdale, J., & Pavard, B. (2004). Modelling character emotion in an interactive virtual environment. In *Proceedings of the aisb 2004 convention, symposium on language, speech and gesture for expressive characters* (pp. 20–28). The Society for the Study of Artificial Intelligence and the Simulation of Behaviour. Retrieved from `http://membres-lig.imag.fr/dugdale/papers/final-modelling-character.pdf`

Mehta, R. K., & Agnew, M. J. (2011). Effects of concurrent physical and mental demands for a short duration static task. *International Journal of Industrial Ergonomics*, *41*(5), 488–493. doi: 10.1016/j.ergon.2011.04.005

Mellor, C. S. (1970). First rank symptoms of schizophrenia: I. the frequency in schizophrenics on admission to hospital. II. differences between individual first rank symptoms. *The British Journal of Psychiatry*, *117*, 15–23. doi: 10.1192/S0007125000192116

Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1654–1664. doi: 10.1109/TASL.2007.899293

Mikesell, L. (2010). Repetitional responses in frontotemporal dementia discourse: Asserting agency or demonstrating confusion? *Discourse Studies*, *12*(4), 465–500. doi: 10.1177/1461445610370127

Miles, L. K., Nind, L. K., & Macrae, C. N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of experimental social psychology*, *45*(3), 585–589. doi: 10.1016/j.jesp.2009.02.002

Milgram, S., & Gudehus, C. (1978). *Obedience to authority.* Ziff-Davis Publishing Company.

Moon, Y. (1999). The effects of physical distance and response latency on persuasion in computer-mediated communication and human–computer communication. *Journal of Experimental Psychology: Applied*, *5*(4), 379–392. doi: 10.1037/1076-898X.5.4.379

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? psychological responses to personality types in human-computer interaction. *Communication research*, *23*(6), 651–674. doi: 10.1177/009365096023006002

Moore, J. W., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and cognition*, *17*(1), 136–144. doi: 10.1016/j.concog.2006.12.004

Moore, J. W., Lagnado, D., Deal, D. C., & Haggard, P. (2009). Feelings of control: contingency determines experience of action. *Cognition*, *110*(2), 279–283. doi: 10.1016/j.cognition.2008.11.006

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and cognition*, *21*(1), 546–561. doi: 10.1016/j.concog.2011.12.002

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and cognition*, *18*(4), 1056–1064. doi: 10.1016/j.concog.2009.05.004

Mulling, T., & Sathiyanarayanan, M. (2015). Characteristics of hand gesture navigation: a case study using a wearable device (myo). In *Proceedings of the 29th British Human Computer Interaction Conference (BHCI'15)* (pp. 283–284). doi: 10.1145/2783446.2783612

Murata, A. (2006). Eye-gaze input versus mouse: Cursor control as a function of age. *International Journal of Human-Computer Interaction*, *21*(1), 1–14. doi: 10.1080/10447310609526168

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited. *Behaviour & information technology*, *10*(4), 257–266. doi: 10.1080/01449299108924288

Nakagawa, S. (2004). A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology*, *15*(6), 1044–1045. doi: 10.1093/beheco/arh107

Nash, C. (2012). *Supporting virtuosity and flow in computer music* (Unpublished doctoral dissertation). University of Cambridge.

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the 12th SIGCHI conference on human factors in computing systems (CHI'94)* (pp. 72–78). ACM Press. doi: 10.1145/191666.191703

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), 756–766. Retrieved from `https://www.ncbi.nlm.nih.gov/pubmed/9627414`

Néda, Z., Ravasz, E., Vicsek, T., Brechet, Y., & Barabási, A.-L. (2000). Physics of the rhythmic applause. *Physical Review E*, *61*(6), 6987. doi: 10.1103/PhysRevE.61.6987

Nessler, J. A., & Gilliland, S. J. (2009). Interpersonal synchronization during side by side treadmill walking is influenced by leg length differential and altered sensory feedback. *Human movement science*, *28*(6), 772–785. doi: 10.1016/j.humov.2009.04.007

Newell, A. (1982). The knowledge level. *Artificial intelligence*, *18*(1), 87–127. doi: 10.1016/0004-3702(82)90012-1

Newton-Dunn, H., Nakano, H., & Gibson, J. (2003). Block jam: a tangible interface for interactive music. In *Proceedings of the 2003 conference on new interfaces for musical expression (NIME'03)* (pp. 170–177). National University of Singapore. Retrieved from `https://pdfs.semanticscholar.org/03bd/37513d0a68bd9fd79373021cbf46a3e4c2c3.pdf`

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer vision and pattern recognition, 2015 (CVPR'15). The proceedings of the IEEE computer society conference on* (pp. 427–436). Retrieved from `https://pdfs.semanticscholar.org/7951/bdef73e196947e8002a51e6283a85fea33e8.pdf`

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological bulletin*, *89*(1), 133. doi: 10.1037/0033-2909.89.1.133

Nobre, A. C., Correa, A., & Coull, J. T. (2007). The hazards of time. *Current opinion* *in neurobiology*, *17*(4), 465–470. doi: 10.1016/j.conb.2007.07.006

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *12*(5), 481–494. doi: 10.1162/105474603322761289

Nowicki, L., Prinz, W., Grosjean, M., Repp, B. H., & Keller, P. E. (2013). Mutual adaptive timing in interpersonal action coordination. *Psychomusicology: Music, Mind, and Brain*, *23*(1), 6–20. doi: 10.1037/a0032039

Obendorf, H. (2009). Minimalism, industrial design and HCI. In *Minimalism* (pp. 81–95). Springer. doi: 10.1007/978-1-84882-371-6_4

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, *25*(1), 46–59. doi: 10.1002/hbm.20131

Panksepp, J., & Bernatzky, G. (2002). Emotional sounds and the brain: the neuro-affective foundations of musical appreciation. *Behavioural processes*, *60*(2), 133–155. doi: 10.1016/S0376-6357(02)00080-3

Patel, A. D. (2010). *Music, language, and the brain.* Oxford University Press.

Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., & Nass, C. I. (2006). Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the 24th SIGCHI conference on human factors in computing systems (CHI'06)* (pp. 1177–1180). ACM Press. doi: 10.1145/1124772.1124948

Pecenka, N., & Keller, P. E. (2011). The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Experimental Brain Research*, *211*(3-4), 505–515. doi: 10.1007/s00221-011-2616-0

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *the BMJ*, *316*(7139), 1236–1238. doi: 10.1136/bmj.316.7139.1236

Picard, R. W., & Picard, R. (1997). *Affective computing* (Vol. 252). MIT Press Cambridge.

Pylyshyn, Z. W. (1988). *Computing in cognitive science.* University of Western Ontario, Centre for Cognitive Science.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. doi: 10.1016/S0010-0277(99)00058-X

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

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review*, *12*(6), 969–992. doi: 10.3758/BF03206433

Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic bulletin & review*, *20*(3), 403–452. doi: 10.3758/s13423-012-0371-2

Reuter-Lorenz, P., Oonk, H., Barnes, L., & Hughes, H. (1995). Effects of warning signals and fixation point offsets on the latencies of pro-versus antisaccades: implications for an interpretation of the gap effect. *Experimental Brain Research*, *103*(2), 287–293. doi: 10.1007/BF00231715

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science*, *26*(6), 867–891. doi: 10.1016/j.humov.2007.07.002

Richardson, M. J., Marsh, K. L., & Schmidt, R. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 62. doi: 10.1037/0096-1523.31.1.62

Rider, M. S. (1985). Entrainment mechanisms are involved in pain reduction, muscle relaxation, and music-mediated imagery. *Journal of Music Therapy*, *22*(4), 183–192. doi: 10.1093/jmt/22.4.183

Riehle, A., Grün, S., Diesmann, M., & Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, *278*(5345), 1950–1953. doi: 10.1126/science.278.5345.1950

Rodin, J. (1986). Aging and health: Effects of the sense of control. *Science*, *233*(4770), 1271–1276. doi: 10.1126/science.3749877

Rohenkohl, G., Cravo, A. M., Wyart, V., & Nobre, A. C. (2012). Temporal expectation improves the quality of sensory information. *Journal of Neuroscience*, *32*(24), 8424–8428. doi: 10.1523/JNEUROSCI.0804-12.2012

Rose, G. M., & Straub, D. W. (2001). The effect of download time on consumer attitude toward the e-service retailer. *E-service Journal*, *1*(1), 55–76. doi: 10.1353/esj.2001.0005

Rosenblum, M. G., Pikovsky, A. S., & Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Physical review letters*, *76*(11), 1804–1807. doi: 10.1103/PhysRevLett.76.1804

Rosenthal, R., Robert, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* CUP Archive.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge University Press.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the 16th SIGCHI conference on human factors in computing systems (CHI'98)* (pp. 480–487). doi: 10.1145/274644.274709

Sanna, L. J., & Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin*, *22*(9), 906–919. doi: 10.1177/0146167296229005

Sarkar, A. (2015). Spreadsheet interfaces for usable machine learning. In *Visual languages and human-centric computing (vl/hcc), 2015 IEEE symposium on* (pp. 283–284). doi: 10.1109/VLHCC.2015.7357228

Sarkar, A. (2017). *Interactive analytical modelling* (Unpublished doctoral dissertation). University of Cambridge.

Sarkar, A., Jamnik, M., Blackwell, A. F., & Spott, M. (2015). Interactive visual machine learning in spreadsheets. In *Visual languages and human-centric computing (vl/hcc), 2015 IEEE symposium on* (pp. 159–163). IEEE. doi: 10.1109/VLHCC.2015.7357211

Sarkar, A., Morrison, C., Dorn, J. F., Bedi, R., Steinheimer, S., Boisvert, J., ... Rota Bulò, S. (2016). Setwise comparison: Consistent, scalable, continuum labels for computer vision. In *Proceedings of the 34th SIGCHI conference on human factors in computing systems (CHI'16)* (pp. 261–271). ACM Press. doi: 10.1145/2858036.2858199

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, *97*(5), 1295–1345. doi: 10.1086/229903

Scherer, K. R., & Zentner, M. R. (2001). Emotional effects of music: Production rules. In P. Juslin & J. Sloboda (Eds.), *Music and emotion: Theory and research*

(Vol. 16, pp. 361–392). New York, USA: Oxford University Press. Retrieved from http://charris.ucsd.edu/SchererZentner.pdf

Scherer, K. R., Zentner, M. R., & Stern, D. (2004). Beyond surprise: the puzzle of infants' expressive reactions to expectancy violation. *Emotion*, *4*(4), 389–402. doi: 10.1037/1528-3542.4.4.389

Schieman, S., & Campbell, J. E. (2001). Age variations in personal agency and self-esteem: the context of physical disability. *Journal of Aging and Health*, *13*(2), 155–185. Retrieved from https://www.researchgate.net/profile/Scott_Schieman/publication/11570010_Age_Variations_in_Personal_Agency_and_Self-Esteem/links/5844627608ae8e63e6271bf6/Age-Variations-in-Personal-Agency-and-Self-Esteem.pdf

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in psychology*, *5*, 1475. doi: 10.3389/fpsyg.2014.01475

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences*, *32*(1), 9–18. doi: 10.1016/j.tins.2008.09.012

Shaffer, L. H. (1981). Performances of chopin, bach, and bartok: Studies in motor programming. *Cognitive psychology*, *13*(3), 326–376. doi: 10.1016/0010-0285(81)90013-X

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139–159. doi: 10.1080/14640748908401189

Shneiderman, B. (1981). Direct manipulation: A step beyond programming languages. In *ACM SIGSOC bulletin (CHI'81)* (Vol. 13, p. 143). ACM Press. doi: 10.1145/800276.810991

Shneiderman, B. (1982). The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology*, *1*(3), 237–256. doi: 10.1080/01449298208914450

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, *43*(9), 63–65. doi: 10.1145/348941.348990

Shneiderman, B. (2010). *Designing the user interface: strategies for effective human-computer interaction.* Pearson Education India.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326–332. doi: 10.1037/0096-1523.29.2.326

Shuler, M. G., & Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, *311*(5767), 1606–1609. doi: 10.1126/science.1123513

Simon, J. R., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics*, *6*(1), 99–105. doi: 10.1080/00140136308930679

Skalski, P., & Tamborini, R. (2007). The role of social presence in interactive agent-based persuasion. *Media psychology*, *10*(3), 385–413. doi: 10.1080/15213260701533102

Skvoretz, J. (1988). Models of participation in status-differentiated groups. *Social Psychology Quarterly*, 43–57. doi: 10.2307/2786983

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *The quarterly journal of experimental psychology*, *35*(2), 377–396. doi: 10.1080/14640748308402140

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3–22. doi: 10.1037/0033-2909.119.1.3

Smith, E. R., & DeCoster, J. (1999). Associative and rule-based processing. *Dual-process theories in social psychology*, 323–336.

Spiro, N., & Himberg, T. (2012). Musicians and non-musicians adapting to tempo differences in cooperative tapping tasks. In *Proceedings of the 12th international conference on music perception and cognition (ICMPC'12)* (Vol. 12, pp. 950–955). Retrieved from `http://icmpc-escom2012.web.auth.gr/files/papers/950_Proc.pdfs`

Spiro, N., Schofield, M., & Himberg, T. (2013). Empathy in musical interaction. In *Proceedings of the 3rd international conference on music & emotion (ICME'13).* Retrieved from `https://jyx.jyu.fi/dspace/bitstream/handle/123456789/41610/Neta%20Spiro%20-%20Empathy%20In%20Musical%20Interaction.pdf?sequence=1`

Spodick, D. H., Raju, P., Bishop, R. L., & Rifkin, R. D. (1992). Operational definition of normal sinus heart rate. *The American journal of cardiology*, *69*(14), 1245–1246.

doi: 10.1016/0002-9149(92)90947-W

Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, *51*(5), 651–659. doi: 10.1016/j.neuron.2006.08.006

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions.* Cambridge University Press.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in cognitive sciences*, *13*(9), 403–409. doi: 10.1016/j.tics.2009.06.003

Sundberg, J., Friberg, A., & Frydén, L. (1991). Threshold and preference quantities of rules for music performance. *Music Perception: An Interdisciplinary Journal*, *9*(1), 71–91. doi: 10.2307/40286159

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and cognition*, *17*(1), 219–239. doi: 10.1016/j.concog.2007.03.010

Tanimoto, S. L. (1990). VIVA: A visual language for image processing. *Journal of Visual Languages & Computing*, *1*(2), 127–139. doi: 10.1016/S1045-926X(05)80012-6

Tanimoto, S. L. (2013). A perspective on the evolution of live programming. In *Proceedings of the 1st international workshop on live programming (LIVE'13)* (pp. 31–34).

Thaut, M. H., McIntosh, G. C., Prassas, S. G., & Rice, R. R. (1993). Effect of rhythmic auditory cuing on temporal stride parameters and EMG patterns in hemiparetic gait of stroke patients. *Journal of Neurologic Rehabilitation*, *7*(1), 9–16. doi: 10.1177/136140969300700103

Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception: An Interdisciplinary Journal*, *3*(1), 33–57. doi: 10.2307/40285321

Tognoli, J. (1969). Response matching in interpersonal information exchange. *British Journal of Clinical Psychology*, *8*(2), 116–123. doi: 10.1111/j.2044-8260.1969.tb00596.x

Tsujimoto, S., & Sawaguchi, T. (2005). Neuronal activity representing temporal prediction of reward in the primate prefrontal cortex. *Journal of Neurophysiology*, *93*(6), 3687–3692. doi: 10.1152/jn.01149.2004

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293–302. doi: 10.1109/

TSA.2002.800560

Tzur, G., & Berger, A. (2009). Fast and slow brain rhythms in rule/expectation violation tasks: Focusing on evaluation processes by excluding motor action. *Behavioural brain research*, *198*(2), 420–428. doi: 10.1016/j.bbr.2008.11.041

Ward, D., Hahn, J., & Feist, K. (2012). Autocomplete as a research tool: a study on providing search suggestions. *Information Technology and Libraries (Online)*, *31*(4), 6-19. Retrieved from `https://search.proquest.com/docview/1356913024?accountid=9851`

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. (2001). Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, *55*(3), 281–292. doi: 10.1006/ijhc.2001.0499

Warner, R. M., Malloy, D., Schneider, K., Knoth, R., & Wilder, B. (1987). Rhythmic organization of social interaction and observer ratings of positive affect and involvement. *Journal of Nonverbal Behavior*, *11*(2), 57–74. doi: 10.1007/BF00990958

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853. doi: 10.1093/brain/107.3.829

Webster Jr, M., Hysom, S. J., & Fullmer, E. M. (1998). Sexual orientation and occupation as status. *Advances in group processes*, *15*, 1–21.

Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends in cognitive sciences*, *7*(2), 65–69. doi: 10.1016/S1364-6613(03)00002-0

Wegner, D. M., & Sparrow, B. (2004). Authorship processing. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1201–1209). Cambridge, MA, USA: Cambridge, MA: MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American psychologist*, *54*(7), 480–492. doi: 10.1037/0003-066X.54.7.480

Weinberg, B. D. (2000). Don't keep your internet customers waiting too long at the (virtual) front door. *Journal of Interactive Marketing*, *14*(1), 30–39. doi: 10.1002/(SICI)1520-6653(200024)14:1⟨30::AID-DIR3⟩3.0.CO;2-M

Williams, R. B., & Clippinger, C. A. (2002). Aggression, competition and computer games: computer and human opponents. *Computers in human behavior*, *18*(5), 495–506. doi: 10.1016/S0747-5632(02)00009-2

Wolber, D. W., & Myers, B. A. (2001). Stimulus-response PBD: Demonstrating "when" as well as "what". In H. Lieberman (Ed.), *Your wish is my command: Giving users the power to instruct their software* (pp. 321–344). Morgan Kaufmann. doi: 10.1016/B978-155860688-3/50017-8

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature neuroscience*, *3*, 1212–1217. doi: 10.1038/81497

Woodall, W. G., & Burgoon, J. K. (1981). The effects of nonverbal synchrony on message comprehension and persuasiveness. *Journal of Nonverbal Behavior*, *5*(4), 207–223. doi: 10.1007/BF00987460

Worringham, C. J., & Beringer, D. B. (1998). Directional stimulus-response compatibility: A test of three alternative principles. *Ergonomics*, *41*(6), 864–880. doi: 10.1080/001401398186694

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(9), 3593–3598. doi: 10.1073/pnas.1120118109

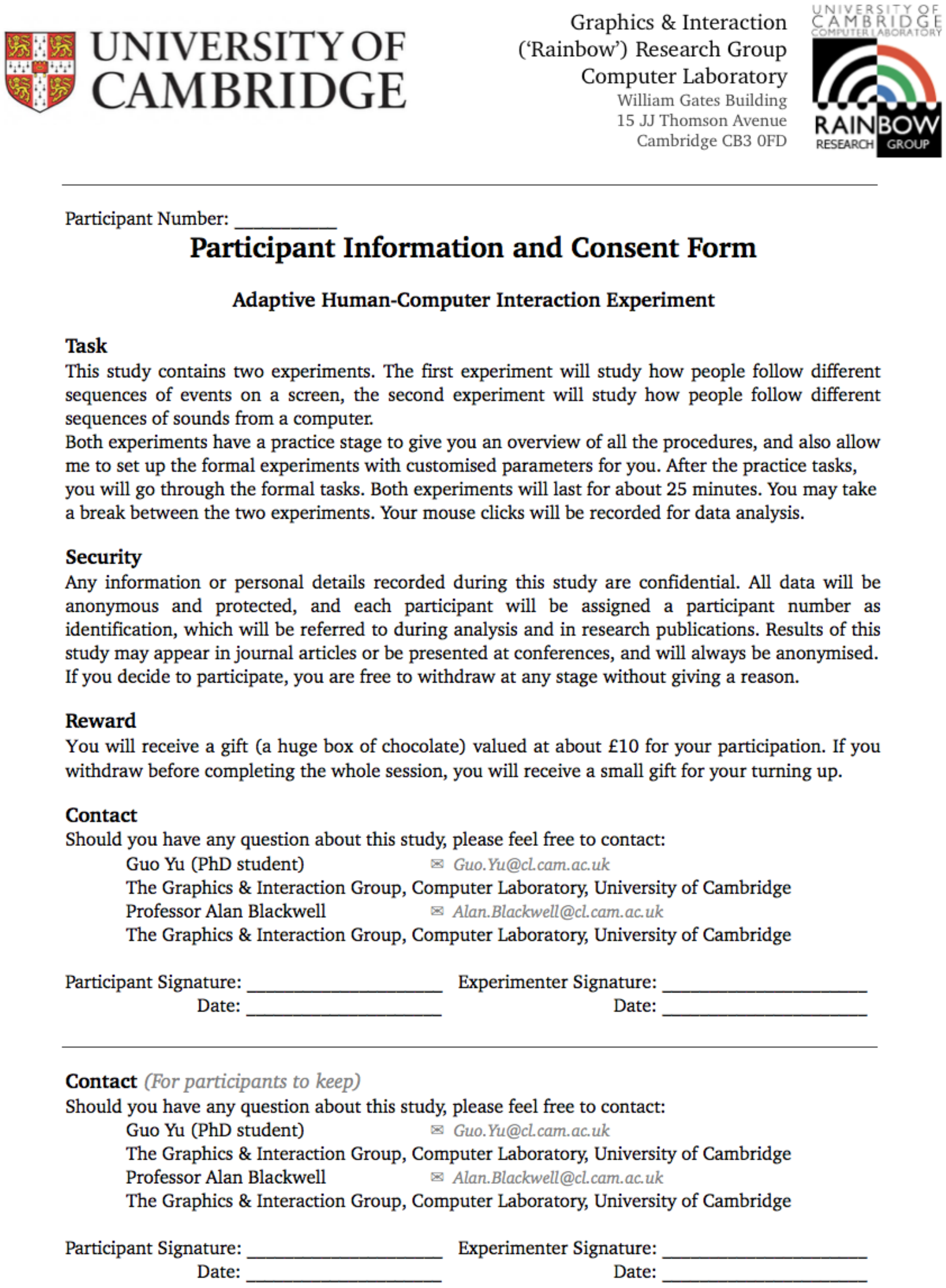
Yu, E., & Cho, S. (2004). Keystroke dynamics identity verification - its problems and practical solutions. *Computers & Security*, *23*(5), 428–440. doi: 10.1016/j.cose.2004.02.004

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, C. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the nasa-tlx tool. *Simulation in healthcare*, *5*(5), 267–271. doi: 10.1097/SIH.0b013e3181e3f329

appendix A

Experiment materials for Experiments 1 and 2 (Chapters 4 and 5)

A.1 Consent form for Experiments 1 and 2



UNIVERSITY OF CAMBRIDGE

Graphics & Interaction
('Rainbow') Research Group
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD

UNIVERSITY OF CAMBRIDGE
COMPUTER LABORATORY
RAINBOW
RESEARCH GROUP

Participant Number: __________

Participant Information and Consent Form

Adaptive Human-Computer Interaction Experiment

Task
This study contains two experiments. The first experiment will study how people follow different sequences of events on a screen, the second experiment will study how people follow different sequences of sounds from a computer.
Both experiments have a practice stage to give you an overview of all the procedures, and also allow me to set up the formal experiments with customised parameters for you. After the practice tasks, you will go through the formal tasks. Both experiments will last for about 25 minutes. You may take a break between the two experiments. Your mouse clicks will be recorded for data analysis.

Security
Any information or personal details recorded during this study are confidential. All data will be anonymous and protected, and each participant will be assigned a participant number as identification, which will be referred to during analysis and in research publications. Results of this study may appear in journal articles or be presented at conferences, and will always be anonymised. If you decide to participate, you are free to withdraw at any stage without giving a reason.

Reward
You will receive a gift (a huge box of chocolate) valued at about £10 for your participation. If you withdraw before completing the whole session, you will receive a small gift for your turning up.

Contact
Should you have any question about this study, please feel free to contact:
Guo Yu (PhD student) ✉ *Guo.Yu@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge
Professor Alan Blackwell ✉ *Alan.Blackwell@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge

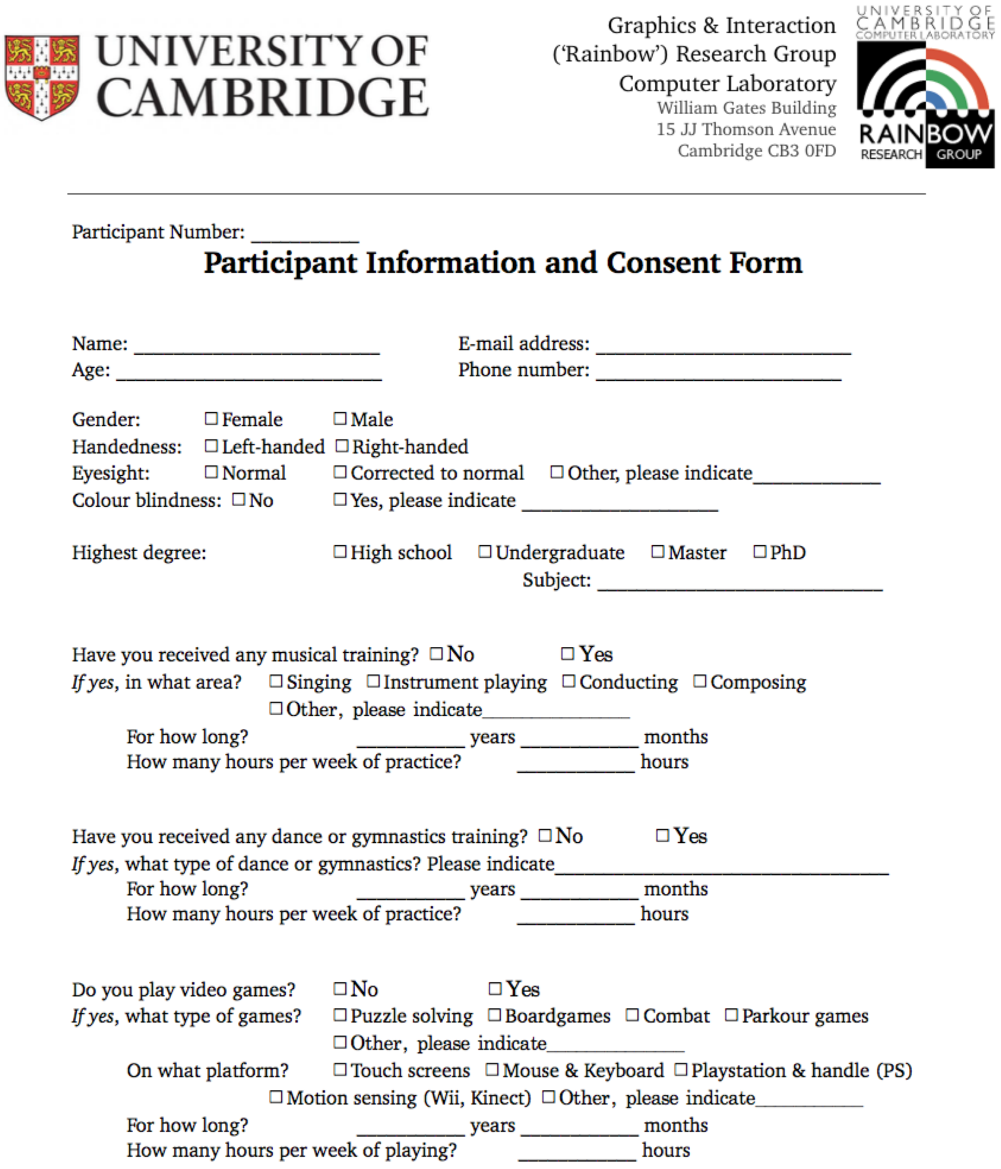
Participant Signature: ____________________ Experimenter Signature: ____________________
Date: ____________________ Date: ____________________

Contact *(For participants to keep)*
Should you have any question about this study, please feel free to contact:
Guo Yu (PhD student) ✉ *Guo.Yu@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge
Professor Alan Blackwell ✉ *Alan.Blackwell@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge

Participant Signature: ____________________ Experimenter Signature: ____________________
Date: ____________________ Date: ____________________

Figure A.1

A.2 Participants' information questionnaire for Experiments 1 and 2



UNIVERSITY OF CAMBRIDGE

Graphics & Interaction
('Rainbow') Research Group
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD

UNIVERSITY OF CAMBRIDGE COMPUTER LABORATORY
RAINBOW RESEARCH GROUP

Participant Number: __________

Participant Information and Consent Form

Name: ______________________ E-mail address: ______________________
Age: ________________________ Phone number: ______________________

Gender: □ Female □ Male
Handedness: □ Left-handed □ Right-handed
Eyesight: □ Normal □ Corrected to normal □ Other, please indicate___________
Colour blindness: □ No □ Yes, please indicate __________________

Highest degree: □ High school □ Undergraduate □ Master □ PhD
Subject: __________________________

Have you received any musical training? □ No □ Yes
If yes, in what area? □ Singing □ Instrument playing □ Conducting □ Composing
□ Other, please indicate_____________
For how long? __________ years __________ months
How many hours per week of practice? __________ hours

Have you received any dance or gymnastics training? □ No □ Yes
If yes, what type of dance or gymnastics? Please indicate______________________________
For how long? __________ years __________ months
How many hours per week of practice? __________ hours

Do you play video games? □ No □ Yes
If yes, what type of games? □ Puzzle solving □ Boardgames □ Combat □ Parkour games
□ Other, please indicate____________
On what platform? □ Touch screens □ Mouse & Keyboard □ Playstation & handle (PS)
□ Motion sensing (Wii, Kinect) □ Other, please indicate__________
For how long? __________ years __________ months
How many hours per week of playing? __________ hours

Figure A.2

A.3 Experiment task briefing scripts for Experiments 1 and 2

A.3.1 Recruitment / General Introduction

This study contains two experiments. The first experiment will study how people follow different sequences of events on a screen, the second experiment will study how people follow different sequences of sounds from a computer.

Both experiments have a practice stage to give you an overview of all the procedures, and also allow me to set up the formal experiments with customised parameters for you. After the practice tasks, you will go through the formal tasks.

Both experiments will last for about 25 minutes. You may take a break between the two experiments.

A.3.2 Experiment 1: Introduction

This experiment will study how people follow various sequences of events on a screen. There are 4 kinds of task. You will be reminded of each one before it starts, so you don't need to remember complicated instructions, but I will just go over now the 5 kinds of task so that you know what kind of action will be involved.

- In one kind of task, you will need to click 4 target crosses, which will appear in order at 4 locations on the screen. Please click at a speed that you find comfortable.
- In another kind of task the screen will first display 4 crosses (you don't need to click), then 4 simple shapes (randomly selected from triangle, square, pentagon and circle), also in order at the 4 locations on the screen. After they have been shown, you will need to recall which shape was displayed at each location.
- In another kind of task, you will click on 4 target crosses at the same locations, then 4 randomised shapes. Then you will need to recall the shapes as before.

- In the remaining kind of task, you will click on 4 target crosses, then wait and observe the display of 4 randomised shapes (you don't need to click). Then you will need to recall the shapes again.

Don't worry about remembering all of these - you will receive instructions before each task.

Do you have any questions?

A.3.3 Experiment 2: Introduction

This experiment will study how people follow various sequences of sounds from a computer.

There are 4 kinds of task. You will be reminded of each one before it starts, so you don't need to remember complicated instructions, but I will just go over now the 4 kinds of task so that you know what kind of action will be involved.

- In the first task, you will need to observe a rotating clock on the screen. During the observation you would be asked to either click on a button or listen to a beep. Then you will need to recall where the clock hand was when you clicked the button or heard the beep.
- In one kind of task, you will need to listen to a series of beeps while observing a rotating clock on the screen. The number of beeps could be 7, 8, 9 or 10 and it is completely random. After the beeps, you will need to recall where the clock hand was on the last beep.
- In another kind of task, you will need to click the 'Click!' button to make the computer beep while observing the clock. Keep clicking until the button completely disappears - there could be either 7, 8, 9 or 10 clicks. After the beeps, you will need to recall where the clock hand was on the last beep.
- In the remaining kind of task, you will need to click the 'Click!' button to make computer beep for 4 times, then the computer will continue to beep for another 3, 4, 5 or 6 times (still random). Keep observing the clock, because you will need to recall the clock hand position on the last beep again.

Don't worry about remembering all of these - you will receive instructions before each task.

Do you have any questions?

A.4 Sample intervals

A.4.1 Sample intervals used in Experiment 1

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Target(1)	Target(2)	Target(3)	Target(4)	Recall(1)	Recall(2)	Recall(3)	Recall(4)
3	856	1047	892	1383	1584	702	785	1280	935	840	761	856
4	821	775	874	1416	1140	837	1277	1381	1025	873	848	792
5	781	1172	652	1082	1425	1513	1063	828	1059	825	768	944
6	1540	1463	653	910	790	1129	1185	853	823	712	1737	792
7	1035	1220	830	1084	794	849	1225	1482	932	1112	960	801
8	1189	718	1226	1064	836	1417	1532	533	1003	785	792	744
9	1142	542	1297	1011	1235	1335	1390	564	984	777	752	872
10	998	1185	1318	855	578	1422	926	1237	790	760	689	736
11	1525	648	1227	688	1272	1199	579	1382	1105	761	904	768
12	1167	1512	845	1113	1365	881	634	1003	834	808	712	897
13	774	646	1238	1549	1400	1196	882	828	1485	945	928	680
14	793	999	1315	1127	999	1175	725	1394	929	896	664	680
15	555	1317	1421	568	1504	634	1439	1089	775	992	1232	625
16	879	1135	549	1121	1529	1046	882	1388	851	720	769	808
17	1409	613	862	1444	1038	878	1022	1250	932	1239	873	784
18	810	535	838	1407	1486	1593	541	1313	942	817	672	713
19	1396	794	1476	1426	911	868	697	946	807	824	873	808
20	1020	1297	890	1027	935	601	1579	1168	758	824	856	784
21	1543	1319	813	626	1263	1396	693	867	933	968	696	681
22	602	944	1070	1458	1330	1436	785	892	1411	832	752	769
23	965	681	1381	674	1052	1341	1551	875	1071	848	728	977
24	565	797	1292	1479	1106	737	1355	1191	1293	928	736	729
25	1396	1437	783	1074	1245	629	1111	850	1030	840	912	609
26	1404	1463	1195	544	769	589	1080	1483	1031	713	712	648
27	579	1041	1545	1482	762	1198	548	1367	1024	808	680	921
28	1245	1192	1507	753	587	1281	703	1247	863	840	841	808
29	685	1366	1471	1417	941	742	1078	828	1037	744	969	704
30	1486	1432	662	695	1334	721	1168	1022	732	744	696	801

Table A.1: Intervals (ms) calculated in Experiment 1 (Participant 1, Task 1 **Sys-ii**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Target(1)	Target(2)	Target(3)	Target(4)	Recall(1)	Recall(2)	Recall(3)	Recall(4)
3	1065	1065	1065	1065	1065	1065	1065	1065	902	888	855	880
4	1065	1065	1065	1065	1065	1065	1065	1065	1063	872	973	858
5	1065	1065	1065	1065	1065	1065	1065	1065	989	921	639	880
6	1065	1065	1065	1065	1065	1065	1065	1065	964	984	735	801
7	1065	1065	1065	1065	1065	1065	1065	1065	980	832	1016	809
8	1065	1065	1065	1065	1065	1065	1065	1065	976	664	816	567
9	1065	1065	1065	1065	1065	1065	1065	1065	1107	825	823	767
10	1065	1065	1065	1065	1065	1065	1065	1065	1422	929	808	872
11	1065	1065	1065	1065	1065	1065	1065	1065	948	1056	767	992
12	1065	1065	1065	1065	1065	1065	1065	1065	940	816	848	880
13	1065	1065	1065	1065	1065	1065	1065	1065	1147	1352	1016	800
14	1065	1065	1065	1065	1065	1065	1065	1065	961	904	1024	984
15	1065	1065	1065	1065	1065	1065	1065	1065	874	928	671	937
16	1065	1065	1065	1065	1065	1065	1065	1065	951	775	704	800
17	1065	1065	1065	1065	1065	1065	1065	1065	832	864	743	712
18	1065	1065	1065	1065	1065	1065	1065	1065	844	872	968	616
19	1065	1065	1065	1065	1065	1065	1065	1065	1470	912	792	840
20	1065	1065	1065	1065	1065	1065	1065	1065	979	1440	808	928
21	1065	1065	1065	1065	1065	1065	1065	1065	1033	873	848	760
22	1065	1065	1065	1065	1065	1065	1065	1065	903	768	768	888
23	1065	1065	1065	1065	1065	1065	1065	1065	1490	928	880	800
24	1065	1065	1065	1065	1065	1065	1065	1065	1063	896	904	760
25	1065	1065	1065	1065	1065	1065	1065	1065	1325	864	688	784
26	1065	1065	1065	1065	1065	1065	1065	1065	923	760	824	792
27	1065	1065	1065	1065	1065	1065	1065	1065	878	800	1072	761
28	1065	1065	1065	1065	1065	1065	1065	1065	1407	928	840	856
29	1065	1065	1065	1065	1065	1065	1065	1065	1051	840	1000	960
30	1065	1065	1065	1065	1065	1065	1065	1065	1475	896	1000	745

Table A.2: Intervals (ms) calculated in Experiment 1 (Participant 1, Task 2 **Sys-pr**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Target(1)	Target(2)	Target(3)	Target(4)	Recall(1)	Recall(2)	Recall(3)	Recall(4)
3	862	697	760	1176	862	697	760	1176	1723	992	793	792
4	1366	840	896	1168	1366	840	896	1168	1085	825	752	840
5	1102	824	881	928	1102	824	881	928	1158	729	816	736
6	807	672	688	825	807	672	688	825	1601	904	776	792
7	831	776	592	817	831	776	592	817	1009	689	696	872
8	767	809	768	736	767	809	768	736	1118	697	728	1216
9	998	696	737	1192	998	696	737	1192	1457	793	696	800
10	1134	744	849	888	1134	744	849	888	1382	705	800	768
11	695	728	776	769	695	728	776	769	967	728	745	712
12	822	880	713	904	822	880	713	904	738	809	600	848
13	831	696	736	888	831	696	736	888	686	648	657	688
14	607	559	705	624	607	559	705	624	1712	776	697	712
15	639	680	696	617	639	680	696	617	1417	721	1096	944
16	598	649	664	776	598	649	664	776	1023	824	761	768
17	615	593	696	808	615	593	696	808	736	817	920	808
18	702	641	832	760	702	641	832	760	1597	992	673	784
19	967	1024	896	937	967	1024	896	937	1214	888	817	712
20	734	649	640	744	734	649	640	744	1108	776	697	680
21	735	665	664	648	735	665	664	648	1349	792	681	840
22	735	632	641	656	735	632	641	656	1028	768	777	704
23	790	729	616	680	790	729	616	680	1842	1857	744	728
24	886	1624	784	936	886	1624	784	936	810	664	761	648
25	518	592	601	760	518	592	601	760	716	664	737	672
26	656	640	696	936	656	640	696	936	674	816	713	1000
27	759	633	744	640	759	633	744	640	704	697	696	944
28	678	625	599	609	678	625	599	609	1441	761	752	840
29	822	641	648	984	822	641	648	984	690	680	777	760
30	838	712	681	736	838	712	681	736	829	712	801	768

Table A.3: Intervals (ms) calculated in Experiment 1 (Participant 1, Task 3 **Usr-Sys**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Target(1)	Target(2)	Target(3)	Target(4)	Recall(1)	Recall(2)	Recall(3)	Recall(4)
3	879	848	672	745	920	904	712	785	1312	1000	2727	825
4	982	985	1136	1128	1032	992	864	856	1256	1000	913	856
5	823	760	657	608	808	801	768	1192	1416	984	936	920
6	1094	753	776	552	737	704	608	673	680	809	664	872
7	774	745	880	888	768	1488	760	865	912	1376	784	824
8	806	737	648	616	745	976	1112	896	1424	744	769	976
9	671	728	721	912	824	1440	976	768	1592	984	760	641
10	775	753	735	689	784	705	720	960	1216	832	856	753
11	718	665	632	673	720	824	840	833	1319	801	784	681
12	751	672	648	673	736	881	704	632	1000	809	576	792
13	743	656	736	673	816	1096	976	752	1632	776	889	704
14	711	721	624	697	864	696	664	745	1224	1136	808	784
15	774	929	880	615	793	936	784	752	1696	1304	808	744
16	1694	1184	752	768	801	952	984	696	864	745	832	784
17	671	656	609	736	928	840	857	752	856	880	784	721
18	855	1016	785	784	744	856	833	768	992	1184	1144	760
19	1014	761	776	928	792	881	848	897	895	2113	800	864
20	694	713	680	744	665	736	848	776	969	792	944	840
21	823	768	816	713	1024	1048	1072	800	1048	912	880	769
22	767	697	728	744	856	977	808	768	840	761	696	856
23	886	833	768	712	753	864	904	736	721	776	800	1024
24	751	752	664	633	816	960	793	728	936	928	1072	697
25	807	712	689	808	856	1448	1383	1001	1352	752	768	768
26	807	744	624	697	720	849	856	736	1440	1024	656	792
27	1095	616	872	744	737	904	792	713	800	784	720	777
28	655	712	673	776	904	1096	889	784	1152	800	745	768
29	703	784	697	712	1664	1320	1792	880	1288	792	784	817
30	647	736	640	744	968	969	1120	888	1535	801	832	840

Table A.4: Intervals (ms) calculated in Experiment 1 (Participant 1, Task 4 **Usr-r**).

A.4.2 Sample intervals used in Experiment 2

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Attention(1)	Attention(2)	Attention(3)	Attention(4)	Attention(5)	Attention(6)
3	855	692	779	525	448	495	882	445	533	314
4	471	690	819	724	336	763	308	507	697	646
5	323	869	587	348	675	481	740	635	724	
6	540	579	868	471	692	304	726			
7	713	659	532	580	471	880	347			
8	717	856	613	463	436	485	517	892	390	
9	396	300	815	880	729	627	658	362		
10	429	755	703	577	450	422	633	760	638	
11	637	377	778	833	472	636	310	481	841	
12	538	503	747	375	872	420	608	715		
13	688	619	830	331	472	640	598			
14	822	525	720	392	847	386	495			
15	354	883	472	654	841	515	464	597		
16	719	351	693	841	737	451	396			
17	674	625	418	701	408	682	560	699		
18	488	791	363	859	545	763	558	411		
19	404	748	302	745	484	879	655	566		
20	663	623	340	772	474	415	767	611	757	551
21	467	379	824	440	792	556	828	484		
22	687	341	783	341	885	649	493			
23	889	509	878	575	396	334	592			
24	511	887	587	399	871	480	696	391	555	
25	514	563	496	388	820	641	892	458	600	
26	312	480	372	650	771	723	861	586	624	

Table A.5: Intervals (ms) calculated in Experiment 2 (Participant 1, Task 1 **Sys-ii**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Attention(1)	Attention(2)	Attention(3)	Attention(4)	Attention(5)	Attention(6)
3	597	597	597	597	597	597	597	597	597	597
4	597	597	597	597	597	597	597	597	597	597
5	597	597	597	597	597	597	597	597	597	
6	597	597	597	597	597	597	597			
7	597	597	597	597	597	597	597			
8	597	597	597	597	597	597	597	597	597	
9	597	597	597	597	597	597	597	597		
10	597	597	597	597	597	597	597	597	597	
11	597	597	597	597	597	597	597	597	597	
12	597	597	597	597	597	597	597	597		
13	597	597	597	597	597	597	597			
14	597	597	597	597	597	597	597			
15	597	597	597	597	597	597	597	597		
16	597	597	597	597	597	597	597			
17	597	597	597	597	597	597	597	597		
18	597	597	597	597	597	597	597	597		
19	597	597	597	597	597	597	597	597		
20	597	597	597	597	597	597	597	597	597	597
21	597	597	597	597	597	597	597	597		
22	597	597	597	597	597	597	597			
23	597	597	597	597	597	597	597			
24	597	597	597	597	597	597	597	597	597	
25	597	597	597	597	597	597	597	597	597	
26	597	597	597	597	597	597	597	597	597	

Table A.6: Intervals (ms) calculated in Experiment 2 (Participant 1, Task 2 **Sys-pr**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Attention(1)	Attention(2)	Attention(3)	Attention(4)	Attention(5)	Attention(6)
3	765	570	640	680	765	570	640	680	663	663
4	559	497	495	472	559	497	495	472	505	505
5	567	680	424	479	567	680	424	479	537	
6	663	457	544	552	663	457	544			
7	438	384	415	448	438	384	415			
8	415	383	864	488	415	383	864	488	537	
9	542	761	463	480	542	761	463	480		
10	447	407	448	440	447	407	448	440	435	
11	367	393	815	480	367	393	815	480	513	
12	351	415	432	424	351	415	432	424		
13	391	608	1167	472	391	608	1167			
14	520	400	752	505	520	400	752			
15	648	423	912	657	648	423	912	657		
16	462	624	608	632	462	624	608			
17	422	512	472	528	422	512	472	528		
18	664	856	848	672	664	856	848	672		
19	391	416	856	408	391	416	856	408		
20	463	416	903	456	463	416	903	456	559	559
21	439	440	448	456	439	440	448	456		
22	375	424	728	456	375	424	728			
23	367	416	864	489	367	416	864			
24	510	992	448	1048	510	992	448	1048	749	
25	568	568	464	464	568	568	464	464	516	
26	431	496	1000	480	431	496	1000	480	601	

Table A.7: Intervals (ms) calculated in Experiment 2 (Participant 1, Task 3 **Usr-Sys**).

	Prompt(1)	Prompt(2)	Prompt(3)	Prompt(4)	Attention(1)	Attention(2)	Attention(3)	Attention(4)	Attention(5)	Attention(6)
3	1148	4913	2289	2600	2520	2496	2544			
4	879	2240	2489	2567	2456	2599	2632	2600		
5	1920	2616	591	1968	624	1824	775	1920	665	
6	1358	752	1728	552	2017	751	1720	848	1872	529
7	1382	673	1776	655	1968	648	584	680	752	
8	615	816	1144	632	585	888	1120	1192	856	
9	1391	1016	1551	1216	1376	553	528			
10	696	648	656	632	624	680	656			
11	704	424	856	648	559	688	657	704	560	720
12	703	592	696	512	687	640	792	632		
13	704	695	528	648	688	711	512	696		
14	712	544	584	663	576	857	600	416	816	719
15	632	687	704	528	656	697	703			
16	744	503	600	608	753	680	664	743		
17	631	704	624	640	552	688	647	657		
18	583	673	640	608	672	769	641	590		
19	623	568	679	648	616	687	721	616	695	
20	680	672	575	512	792	664	632	672	593	
21	847	840	952	551	768	568	697	488		
22	895	576	544	664	616	592	623			
23	672	663	616	639	553	696	704			
24	999	1511	600	824	584	624	593	672	689	
25	478	648	656	617	760	768	648			
26	703	592	592	608	680	672	736	649	584	

Table A.8: Intervals (ms) calculated in Experiment 2 (Participant 1, Task 4 **Usr-r**).

A.5 Randomised order of tasks

A.5.1 Randomised order of tasks in Experiment 1

Participant				
1	Task 1	Task 2	Task 3	Task 4
2	Task 1	Task 3	Task 2	Task 4
3	Task 1	Task 4	Task 2	Task 3
4	Task 1	Task 2	Task 4	Task 3
5	Task 1	Task 3	Task 4	Task 2
6	Task 1	Task 4	Task 3	Task 2
7	Task 2	Task 1	Task 3	Task 4
8	Task 2	Task 3	Task 4	Task 1
9	Task 2	Task 4	Task 3	Task 1
10	Task 2	Task 1	Task 4	Task 3
11	Task 2	Task 3	Task 1	Task 4
12	Task 2	Task 4	Task 1	Task 3
13	Task 3	Task 1	Task 2	Task 4
14	Task 3	Task 2	Task 1	Task 4
15	Task 3	Task 4	Task 1	Task 2
16	Task 3	Task 1	Task 4	Task 2
17	Task 3	Task 2	Task 4	Task 1
18	Task 3	Task 4	Task 2	Task 1
19	Task 4	Task 1	Task 2	Task 3
20	Task 4	Task 2	Task 1	Task 3
21	Task 4	Task 3	Task 1	Task 2
22	Task 4	Task 1	Task 3	Task 2

Table A.9: Randomised order of tasks for each participant in Experiment 1

A.5.2 Randomised order of tasks in Experiment 2

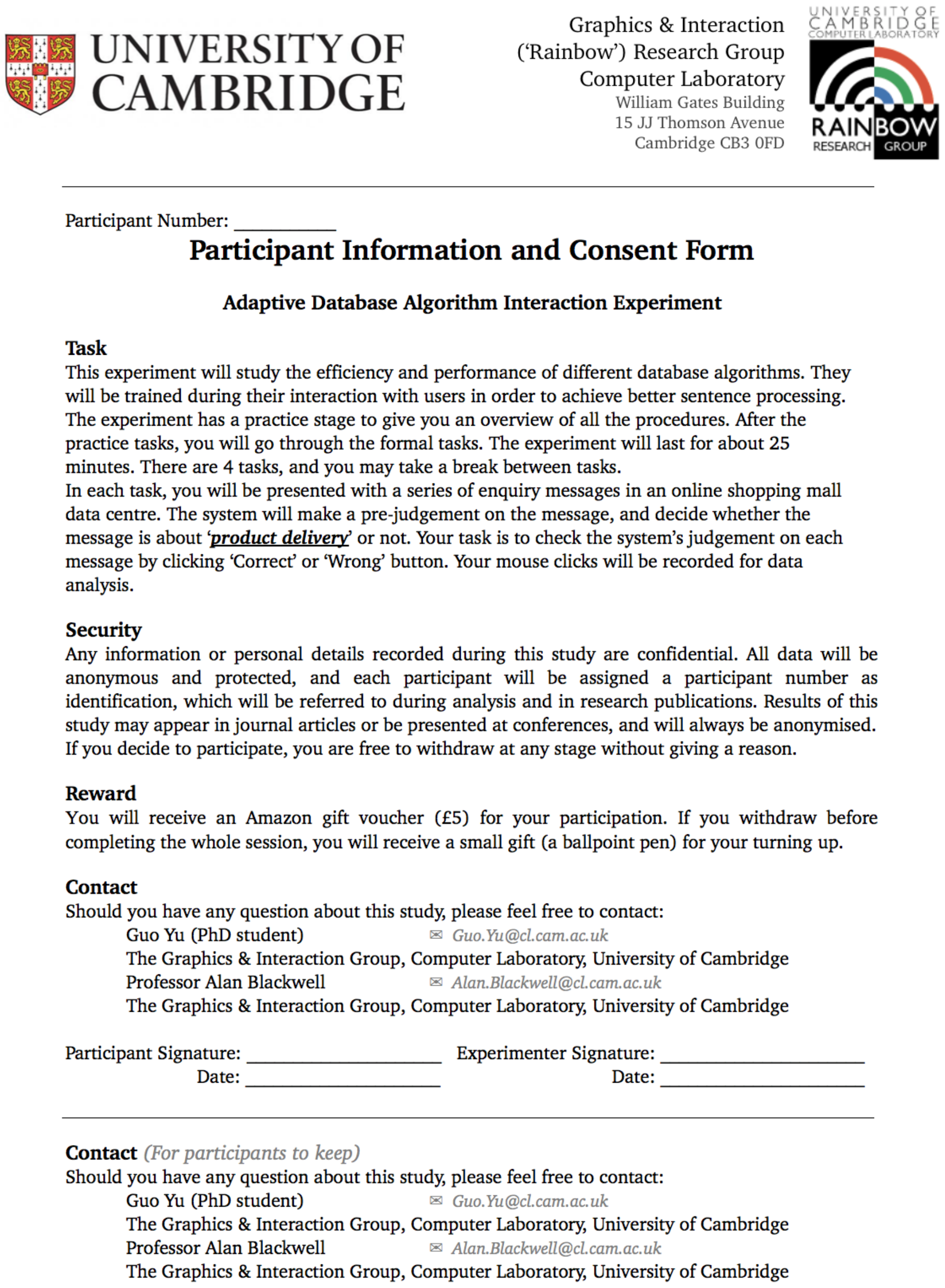
Participant				
1	Task 3	Task 1	Task 2	Task 4
2	Task 3	Task 2	Task 1	Task 4
3	Task 3	Task 4	Task 1	Task 2
4	Task 4	Task 1	Task 3	Task 2
5	Task 4	Task 2	Task 3	Task 1
6	Task 4	Task 3	Task 2	Task 1
7	Task 4	Task 1	Task 2	Task 3
8	Task 4	Task 2	Task 1	Task 3
9	Task 4	Task 3	Task 1	Task 2
10	Task 3	Task 1	Task 4	Task 2
11	Task 3	Task 2	Task 4	Task 1
12	Task 3	Task 4	Task 2	Task 1
13	Task 1	Task 2	Task 3	Task 4
14	Task 1	Task 3	Task 2	Task 4
15	Task 1	Task 4	Task 2	Task 3
16	Task 2	Task 1	Task 4	Task 3
17	Task 2	Task 3	Task 1	Task 4
18	Task 2	Task 4	Task 1	Task 3
19	Task 2	Task 1	Task 3	Task 4
20	Task 2	Task 3	Task 4	Task 1
21	Task 2	Task 4	Task 3	Task 1
22	Task 1	Task 2	Task 4	Task 3

Table A.10: Randomised order of tasks for each participant in Experiment 2

APPENDIX B

EXPERIMENT MATERIALS FOR EXPERIMENT 3 (CHAPTER 6)

B.1 Consent form for Experiment 3



UNIVERSITY OF CAMBRIDGE

Graphics & Interaction
('Rainbow') Research Group
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD

UNIVERSITY OF CAMBRIDGE COMPUTER LABORATORY
RAINBOW RESEARCH GROUP

Participant Number: __________

Participant Information and Consent Form

Adaptive Database Algorithm Interaction Experiment

Task
This experiment will study the efficiency and performance of different database algorithms. They will be trained during their interaction with users in order to achieve better sentence processing. The experiment has a practice stage to give you an overview of all the procedures. After the practice tasks, you will go through the formal tasks. The experiment will last for about 25 minutes. There are 4 tasks, and you may take a break between tasks.
In each task, you will be presented with a series of enquiry messages in an online shopping mall data centre. The system will make a pre-judgement on the message, and decide whether the message is about '***product delivery***' or not. Your task is to check the system's judgement on each message by clicking 'Correct' or 'Wrong' button. Your mouse clicks will be recorded for data analysis.

Security
Any information or personal details recorded during this study are confidential. All data will be anonymous and protected, and each participant will be assigned a participant number as identification, which will be referred to during analysis and in research publications. Results of this study may appear in journal articles or be presented at conferences, and will always be anonymised. If you decide to participate, you are free to withdraw at any stage without giving a reason.

Reward
You will receive an Amazon gift voucher (£5) for your participation. If you withdraw before completing the whole session, you will receive a small gift (a ballpoint pen) for your turning up.

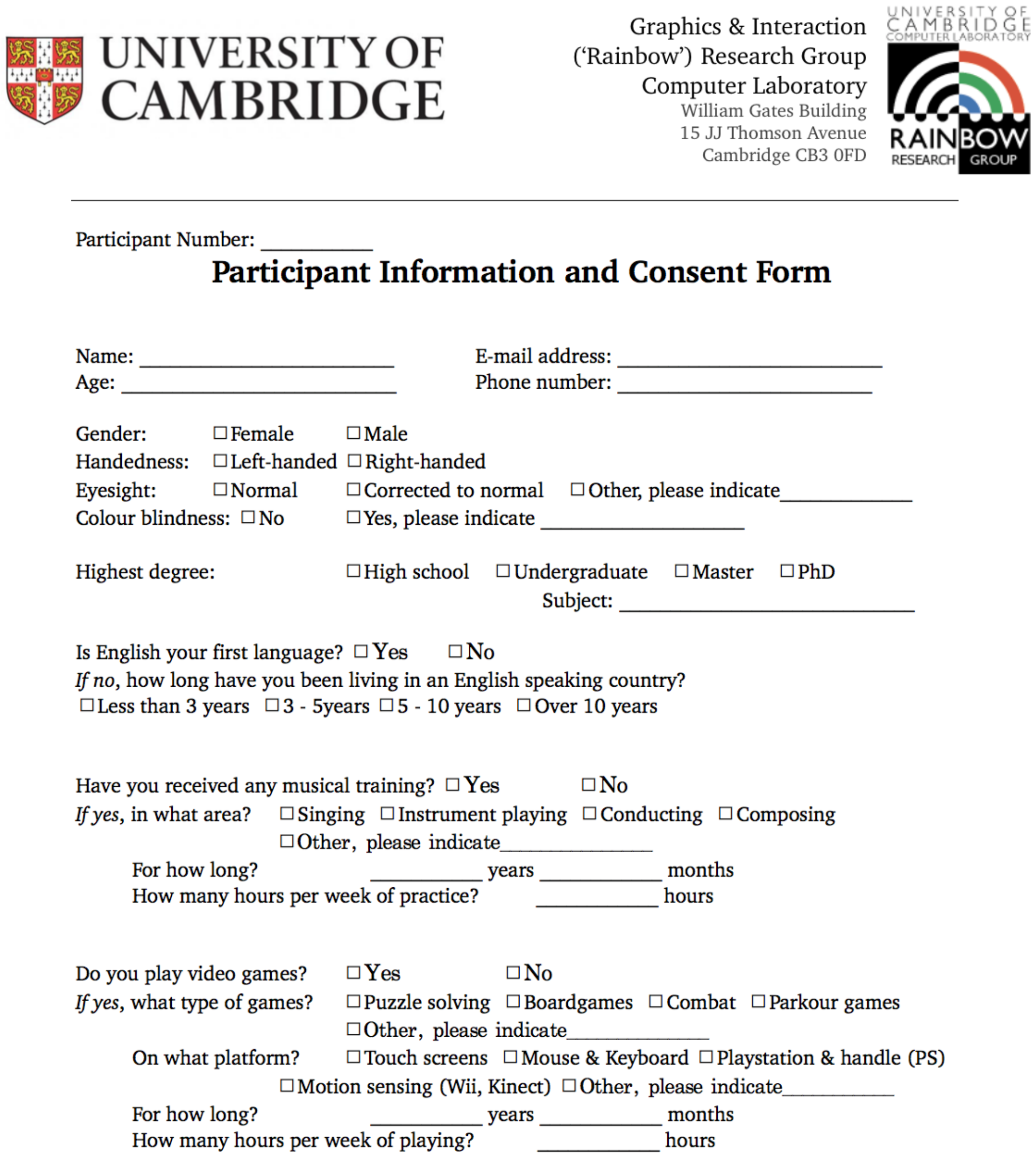
Contact
Should you have any question about this study, please feel free to contact:
Guo Yu (PhD student) ✉ *Guo.Yu@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge
Professor Alan Blackwell ✉ *Alan.Blackwell@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge

Participant Signature: ____________________ Experimenter Signature: ____________________
Date: ____________________ Date: ____________________

Contact *(For participants to keep)*
Should you have any question about this study, please feel free to contact:
Guo Yu (PhD student) ✉ *Guo.Yu@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge
Professor Alan Blackwell ✉ *Alan.Blackwell@cl.cam.ac.uk*
The Graphics & Interaction Group, Computer Laboratory, University of Cambridge

Figure B.1

B.2 Participants' information questionnaire for Experiment 3



UNIVERSITY OF CAMBRIDGE

Graphics & Interaction
('Rainbow') Research Group
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD

UNIVERSITY OF CAMBRIDGE COMPUTER LABORATORY
RAINBOW RESEARCH GROUP

Participant Number: __________

Participant Information and Consent Form

Name: ______________________ E-mail address: ______________________
Age: ________________________ Phone number: ______________________

Gender: □Female □Male
Handedness: □Left-handed □Right-handed
Eyesight: □Normal □Corrected to normal □Other, please indicate___________
Colour blindness: □No □Yes, please indicate _________________

Highest degree: □High school □Undergraduate □Master □PhD
Subject: _________________________

Is English your first language? □Yes □No
If no, how long have you been living in an English speaking country?
□Less than 3 years □3 - 5years □5 - 10 years □Over 10 years

Have you received any musical training? □Yes □No
If yes, in what area? □Singing □Instrument playing □Conducting □Composing
□Other, please indicate_____________
For how long? __________ years ___________ months
How many hours per week of practice? ___________ hours

Do you play video games? □Yes □No
If yes, what type of games? □Puzzle solving □Boardgames □Combat □Parkour games
□Other, please indicate_____________
On what platform? □Touch screens □Mouse & Keyboard □Playstation & handle (PS)
□Motion sensing (Wii, Kinect) □Other, please indicate__________
For how long? __________ years ___________ months
How many hours per week of playing? ___________ hours

Figure B.2

B.3 Experiment task briefing scripts for Experiment 3

General Introduction:

Thank you very much for taking the time to participate in this study!

The experiment will study the efficiency and performance of different database algorithms. They will be "trained" during their interaction with users in order to achieve better language processing.

Background:

An online shopping mall has a data centre. Recently they developed a few machine learning algorithms, which can process customers' enquiry messages, and automatically label messages into several categories, such as "delivery", "exchange and return", aamembership", and so on.

However the performance of those algorithms are quite poor at the moment, and the system often makes wrong judgements. Therefore they are now recruiting people to manually "train" the algorithms, to make them better.

Your tasks:

As one of the first steps, the data centre wants to let the algorithms judge whether an enquiry message is about "product delivery" or not.

So your task is to check the system's judgement:

- If that message is about "product delivery" and the system says so too, then you click the "Correct" button, in this way you can reinforce the correct formula of the system.
- If that message is about "product delivery" but the system says it's not, then you click the "Wrong" button, in this way you can rectify the wrong formula of the system.

Similarly,

- If that message is NOT about “product delivery” and the system says it isn’t as well, then you click the “Correct” button, in this way you can reinforce the correct formula.
- If that message is NOT about “product delivery” but the system says it is, then you click the “Wrong” button, in this way you can rectify the wrong formula.

Over time, the algorithms will be trained by you and its performance will improve.

This experiment has a practice stage to give you an overview of all the procedures. After the practice tasks, you will go through the formal tasks. The experiment will last for about 20 minutes. There are 4 tasks, and you may take a break between tasks.

There are 2 kinds of task. You will be reminded of each one before it starts, so you don’t need to remember complicated instructions, but I will just go over now the task so that you know what kind of action will be involved.

- In one kind of task, you will need to click the “Start Task” button first. Then the system will start presenting a message. The system will make a pre-judgement on the message, and decide whether that enquiry message is about “product delivery” or not. Your task is to check the system?s judgement on each message by clicking the “Correct” or “Wrong” button. The system will keep presenting more messages one by one, and they may pile up as new messages arrive.
- In the other kind of task, you will also need to click the “Start Task” button first. Then the system will present an enquiry message and whether it’s about “product delivery” to you. Again you will need to determine whether the system’s judgement is “Correct” or “Wrong” by clicking the buttons. The difference here is that you will need to click the “Show Next” button when you want to move on to the next messages.

Defining “Product Delivery”:

One more thing before the experiment: the definition of “product delivery”.

Any messages regarding how, when, to where the order is shipped to the customer is considered as “product delivery” category. (keywords: deliver, parcel, post, receive, shipping, address, courier, etc.)

Complaints and enquiry about membership, product information, return and exchange, promotion, customisation, and other issues are not in “product delivery” category.

Sometimes the messages can be quite vague. If you find it hard to tell whether a message is about “product delivery” or not, don’t stress, just make your best guess, then move on to the next message.

Don’t worry about remembering all of these - you will receive instructions before each task.

Do you have any questions?

B.4 Messages and labels used in Experiment 3

	Task 1 messages	Initial labels	Designed truth
1	Hi can I exchange this superman cape with a batman mask?	0	0
2	Should I use my home or work address for anytime delivery?	1	1
3	How often do you have extra discount for students?	1	0
4	The latest order on my account was not placed by me, why?	0	0
5	How big and how heavy is this portable bicycle pump?	1	0
6	What's the difference between edition 6 and 7 of this book?	0	0
7	Could you shorten the sleeves of the new blazer I got here?	1	0
8	Hey can I ask when I can expect my parcel to be delivered?	0	1
9	Any chance you could help me chase the courier up?	0	1
10	What's the fabric of this coat? Can it be machine washed?	0	0
11	What should I do if I lost both my login name and password?	0	0
12	Could you please help me check if this bag is in stock?	0	0
13	I'd like to cancel my last order, how should I proceed?	0	0
14	If I need to return an item, will you offer a prepaid label?	1	0
15	Could you repair the music player I bought from you?	1	0
16	Would it be cheaper if I combine two orders in one delivery?	1	1
17	I'm not in later today, could you deliver it tomorrow?	1	1
18	Could you deliver it to the nearest convenience store?	1	1
19	How would you charge for delivering fragile objects?	0	1
20	Could you please confirm that my order is with courier now?	0	1
21	Can I ask for a discount after I get the membership?	1	0
22	I wonder if I get two pairs of customised headphones?	0	0
23	Is this leather belt reversible? And how do I reverse it?	1	0
24	Why does the water still taste bad after we used the filter?	1	0
25	Hi there, can I expect to get my order refunded by tomorrow?	1	0
26	How much does it cost to replace the ink carriage?	0	0
27	I can't find my parcel, could you check where you put it?	0	1
28	Hello, can I know how much it cost to post with 2nd class?	1	1
29	Will I enjoy faster service I sign up for your newsletter?	1	0
30	hey could you check if this shirt is available in store?	0	0

Table B.1: Short messages used in Experiment 3, their random initial labels and designed truth (e.g. Task 1, Participant 1). The value "1" stands for "product delivery", "0" for "NOT product delivery". Participants were presented with initial labels, and their expected labels should be consistent with the designed truth.

	Task 2 messages	Initial labels	Designed truth
31	The desk lamp is not working, can I return or exchange it?	0	0
32	What's the material of this bracelet, sterling silver?	1	0
33	Why does this perfume bottle looks different from your ads?	0	0
34	How long will this floral scent candle (in large jar) last?	0	0
35	My daughter said the dress didn?t fit, can I get a size 4?	1	0
36	What the length and width of your giftwrapping paper?	1	0
37	Does this double air mattress come with an electric pump?	1	0
38	I need the boots tonight, why haven't them been delivered?	1	1
39	Can I upgrade my delivery option and how much should I pay?	0	1
40	Could you sent my three orders altogether on the same day?	1	1
41	Hey, will this CD be on promotion next month? Should I wait?	1	0
42	Can you engrave customised word on the back of this watch?	1	0
43	Do I have to use branded heads for an electronic toothbrush?	0	0
44	Hello, could you please post my order in five days?	0	1
45	I was thinking if you could leave the parcel in the garden?	1	1
46	How many bars of turkish delights are there in one box?	0	0
47	How will the logistics affect when I could have my order?	1	1
48	hey, just to check if you provide a UK plug for this razor?	1	0
49	Will you help me to install the air conditioner for free?	1	0
50	The photoframe was bent when it arrived, can I change one?	0	0
51	Can I get twenty feet of garden hose when it's in stock?	0	0
52	How will you let me know when the order is delivered?	1	1
53	Is this reindeer phone case available for iPhone 6s?	0	0
54	Is it okay if I want to change my delivery address?	0	1
55	Hello, do you still have the old castle themed Playmobil?	1	0
56	Where can I check my order status on my account page?	1	0
57	Can I track the delivery online? What's the webpage link?	0	1
58	Is this chest of drawers safe under EU regulations?	0	0
59	I want to choose premium delivery, will it be safer?	0	1
60	What is the maximum volume of this food processor?	0	0

Table B.2: Short messages used in Experiment 3, their random initial labels and designed truth (e.g. Task 2, Participant 1). The value "1" stands for "about product delivery", "0" for "NOT about product delivery". During labelling, participants were presented with initial labels, and their expected labels should be consistent with the designed truth.

	Task 3 messages	Initial labels	Designed truth
61	What type of capsules does this espresso machine need?	1	0
62	Could you please check if there's a delay in delivery?	0	1
63	The sunglasses are hurting my nose, could you repair it?	0	0
64	Can I expect to receive it before I leave for holiday?	1	1
65	I returned this blazer ages ago, when can I have my refund?	0	0
66	Why isn't my product review displayed on the webpage?	1	0
67	Just wondering if you've left my parcel at the reception?	1	1
68	My order is meant to be a gift, will it be with me tomorrow?	1	1
69	This cast iron tray is too heavy, do you have a light one?	0	0
70	Could you wrap this Teddy Bear with pink paper please?	0	0
71	Do you have a cowboy hat in the same style of Westworld?	1	0
72	Hello is it normal that the cleansing wipes smell a bit?	1	0
73	How long do I need to wait if I'm ordering from abroad?	1	1
74	The string of the sushi mat is too loose, is that normal?	0	0
75	Where can I get a coupon for your next sales season?	1	0
76	Can I choose the flavour of chocolate on that cake?	0	0
77	How will you protect my china dishes when you deliver them?	0	1
78	Will it be cheaper I opt for normal packaging and courier?	1	1
79	Any chance you could help me to add a student discount?	0	0
80	The handle of the mug isn't comfortable, can I return it?	1	0
81	Hello, could you confirm that you'll post with 1st-class?	0	1
82	Do I need to sign for my friend's order when it arrives?	0	1
83	Hi any chance you could send my bill to another address?	1	0
84	Hi any chance you could delivery my post to another address?	0	1
85	hey my shirt was coloured, do you know how to bleach it?	0	0
86	Your can opener cut my fingers, how do you compensate me?	1	0
87	Could you add a 'Thank you' tag for this fountain pen?	1	0
88	May I change the colour of the jumper I ordered last week?	1	0
89	How to purchase a gift card for my mom for her birthday?	0	0
90	How long do I have before the warranty of my laptop expires?	0	0

Table B.3: Short messages used in Experiment 3, their random initial labels and designed truth (e.g. Task 3, Participant 1). The value "1" stands for "about product delivery", "0" for "NOT about product delivery". During labelling, participants were presented with initial labels, and their expected labels should be consistent with the designed truth.

	Task 4 messages	Initial labels	Designed truth
91	Are those jars airtight enough to store my homemade jam?	1	0
92	What will happen if I'm not in when the courier buzzes me?	0	1
93	Was wondering if I can exchange this to one size smaller?	1	0
94	Just to check if the backpack I bought is on the way?	0	1
95	Hey will the order arrive by tomorrow afternoon or evening?	0	1
96	This pair of jeans are too long, do you do alternation?	1	0
97	I think I received an extra towel, was I charged for it?	0	0
98	How many days should I wait for a standard class parcel?	1	1
99	What accessories would come along with the new bike?	1	0
100	Where can I get a cocktail shaker with two measuring cups?	0	0
101	Would you help me put up those birdhouses I bought?	0	0
102	How much charcoal (kg) do I need to set up this BBQ?	1	0
103	Could you please leave my parcel to my neighbour if I'm out?	1	1
104	I'd like to get some Ghibli wall stickers, how big are they?	0	0
105	The glass vase is broken when it arrived, I need a refund	1	0
106	How much will three shovels and five pots cost altogether?	1	0
107	The hair dryer I bought is too hot, is that really safe?	0	0
108	The book I ordered looks like a second-hand one, why?	1	0
109	Not sure when I can expect to receive the wine I ordered?	0	1
110	I struggle to fold this umbrella all the time, is it broken?	0	0
111	Does this belt come together with the dress or not?	0	0
112	How much will it cost if I choose premium delivery?	0	1
113	Will this dress still be available during your Xmas sale?	1	0
114	The shoes I ordered smell so bad, are they real leather?	0	0
115	Hey may I ask when you're going to deliver my parcel?	1	1
116	How much water should I put in the pan to make porridge?	1	0
117	How long does it take for this nail polish to dry?	0	0
118	I still can't find my order, where did you put it?	1	1
119	I left the wrong address on the order, has it been sent yet?	1	1
120	I received the wrong item, how to send it back to you?	0	0

Table B.4: Short messages used in Experiment 3, their random initial labels and designed truth (e.g. Task 4, Participant 1). The value "1" stands for "about product delivery", "0" for "NOT about product delivery". During labelling, participants were presented with initial labels, and their expected labels should be consistent with the designed truth.

B.5 Sample intervals used in Experiment 3

	Sys-ii(Task 1)		**Sys-pr**(Task 2)		**Usr-Sys**(Task 3)		**Usr-r**(Task 4)	
	t_{Sys-ii,R_k}	t_{Sys-ii,S_k}	t_{Sys-pr,R_k}	t_{Sys-pr,S_k}	$t_{Usr-Sys,R_k}$	$t_{Usr-Sys,S_k}$	t_{Usr-r,R_k}	t_{Usr-r,S_k}
3	11.726	3.031	5.454	4.401	2.594	4.280	4.701	4.240
4	8.994	4.363	2.852	4.408	5.633	4.577	1.854	5.495
5	5.774	5.045	6.737	4.404	4.466	3.096	4.118	2.272
6	4.425	3.421	3.816	4.409	1.674	3.095	1.947	4.627
7	5.201	5.199	4.955	4.405	5.761	6.138	1.628	2.391
8	4.317	2.924	3.764	4.407	5.084	2.189	2.276	2.088
9	2.553	4.405	6.304	4.405	4.959	2.189	2.381	2.703
10	6.899	6.438	3.098	4.414	4.337	2.189	2.716	2.753
11	5.651	3.704	6.913	4.401	8.506	6.264	4.341	3.344
12	2.487	4.780	3.730	4.407	4.859	4.842	2.830	4.871
13	4.265	3.190	6.247	4.403	4.523	4.845	2.526	3.232
14	2.749	5.516	6.771	4.412	4.362	9.017	13.550	3.064
15	4.068	3.329	4.753	4.402	9.646	5.029	5.429	13.969
16	6.226	4.434	2.027	4.407	6.692	4.865	4.453	5.912
17	3.847	5.284	1.829	4.406	3.385	4.860	7.413	4.952
18	2.646	6.129	1.423	4.405	8.490	10.151	1.638	7.864
19	1.978	4.604	3.217	4.407	6.406	3.891	2.468	2.049
20	2.486	2.812	5.115	4.406	3.721	3.893	1.805	2.895
21	10.278	3.344	3.379	4.408	2.625	3.888	1.644	2.121
22	6.579	5.907	2.998	4.405	2.361	4.233	2.430	1.991
23	4.504	3.331	2.118	4.408	11.812	3.125	16.053	2.777
24	7.207	5.793	2.712	4.406	10.815	2.861	3.885	16.464
25	5.316	3.283	2.586	4.406	10.761	2.861	4.214	4.503
26	2.205	6.295	6.342	4.405	8.997	2.860	5.316	4.594
27	3.791	3.750	3.763	4.410	7.712	2.862	18.877	5.863
28	2.303	3.081	6.834	4.402	6.453	2.858	3.014	19.320
29	7.200	4.919	6.368	4.410	8.912	11.326	2.262	3.328
30	5.066	4.286	5.740	4.405	4.167	6.953	4.006	2.568

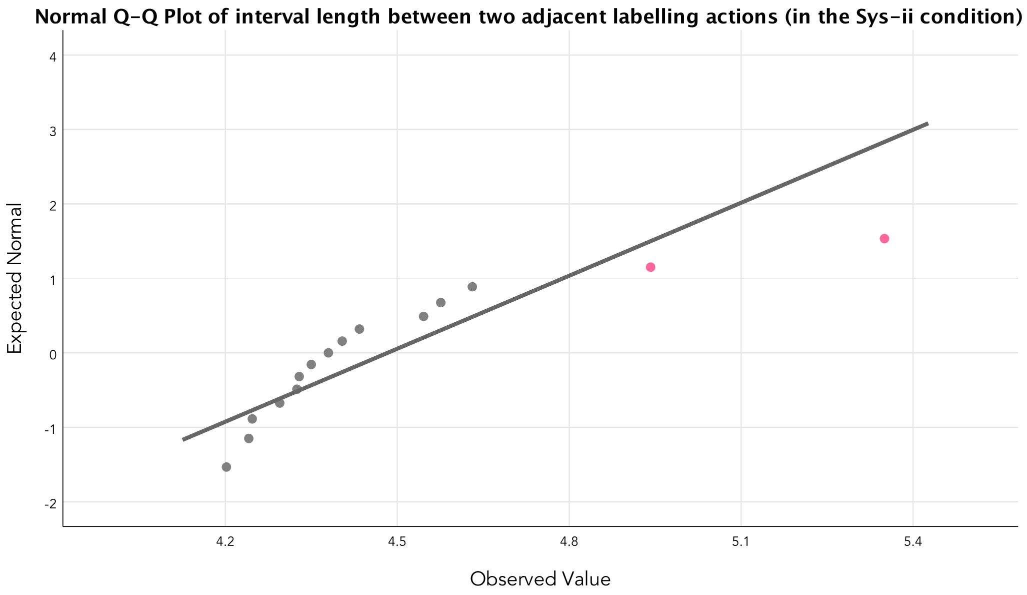
Table B.5: Inter-onset intervals (sec) calculated in Experiment 3 (Participant 1, all tasks). They were used to calculate the windowed cross-correlation coefficient between message updating intervals and the participant's response intervals.

B.6 Randomised order of tasks in Experiment 3

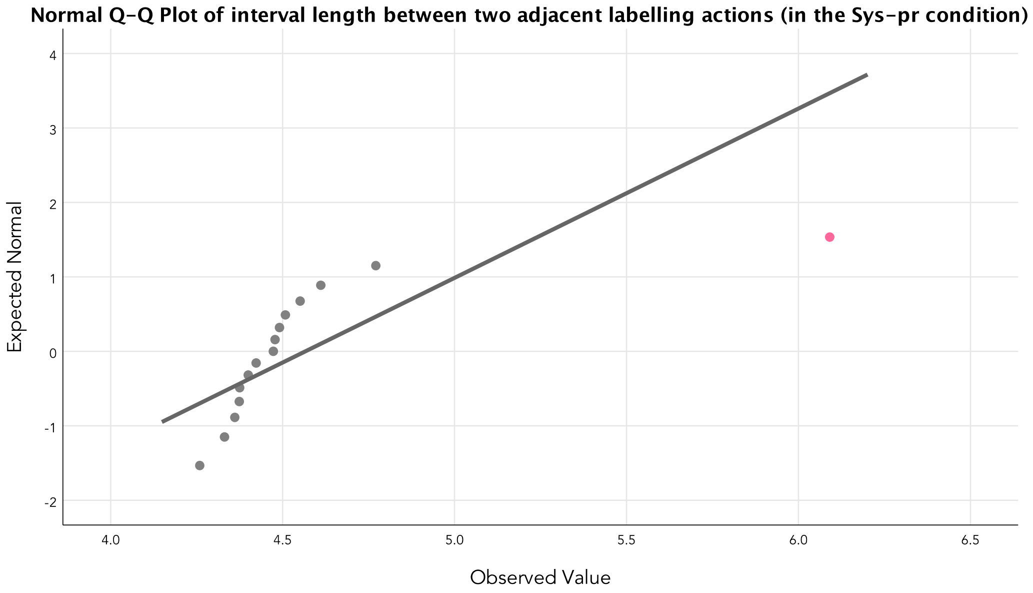
Participant				
1	Task 1	Task 2	Task 3	Task 4
2	Task 1	Task 2	Task 4	Task 3
3	Task 1	Task 3	Task 2	Task 4
4	Task 1	Task 4	Task 3	Task 2
5	Task 1	Task 4	Task 2	Task 3
6	Task 2	Task 1	Task 3	Task 4
7	Task 2	Task 3	Task 4	Task 1
8	Task 2	Task 4	Task 1	Task 3
9	Task 2	Task 1	Task 4	Task 3
10	Task 2	Task 3	Task 1	Task 4
11	Task 3	Task 1	Task 2	Task 4
12	Task 3	Task 2	Task 1	Task 4
13	Task 3	Task 4	Task 2	Task 1
14	Task 3	Task 1	Task 4	Task 2
15	Task 4	Task 1	Task 2	Task 3

Table B.6: Randomised order of tasks for each participant in Experiment 3

B.7 Q-Q plots

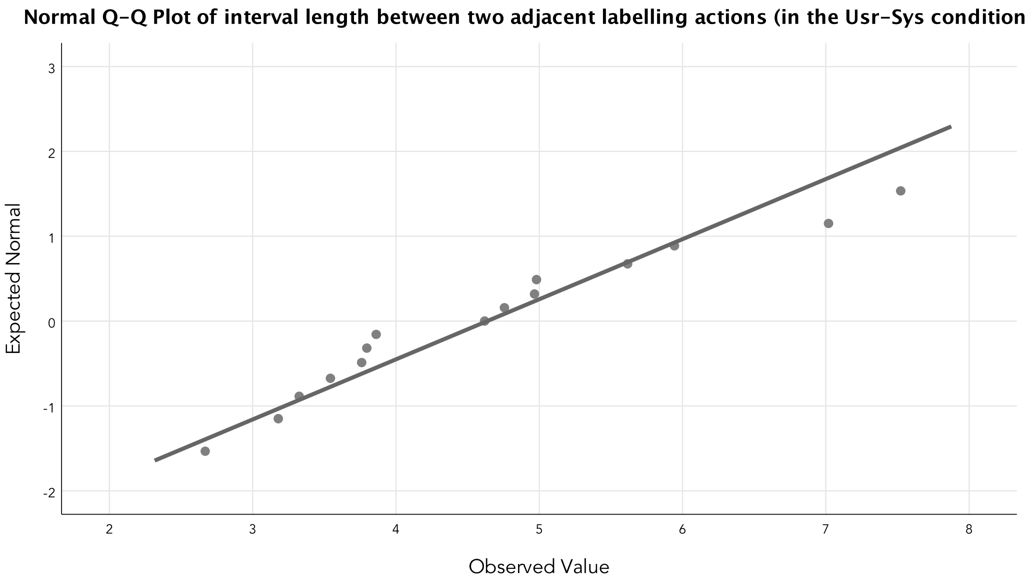


(a)

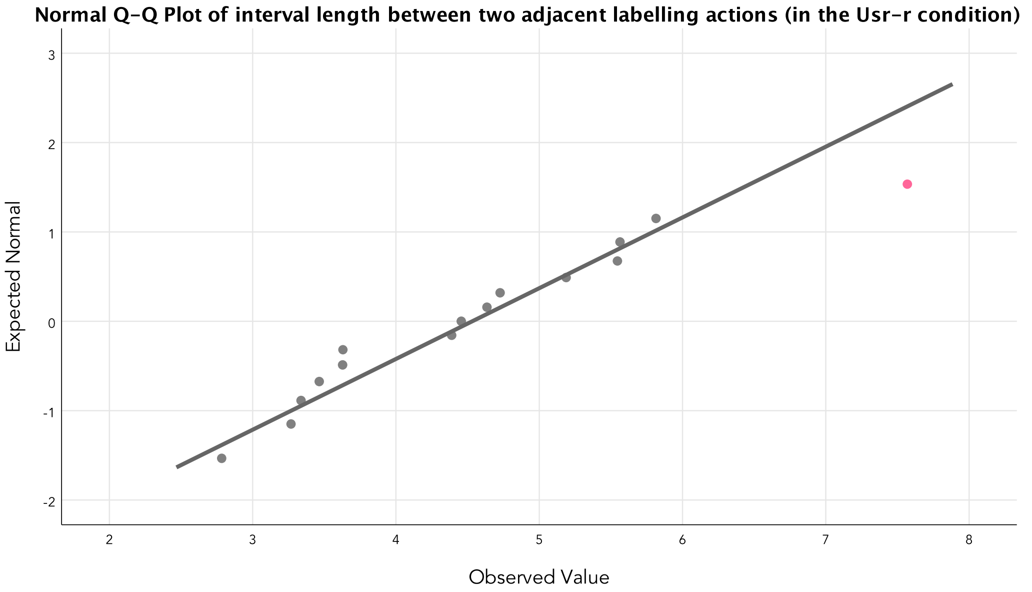


(b)

Figure B.3: Q-Q plots for the interval length between two adjacent labelling actions in the (a) **Sys-ii** condition and (b) **Sys-pr** condition in Experiment 3

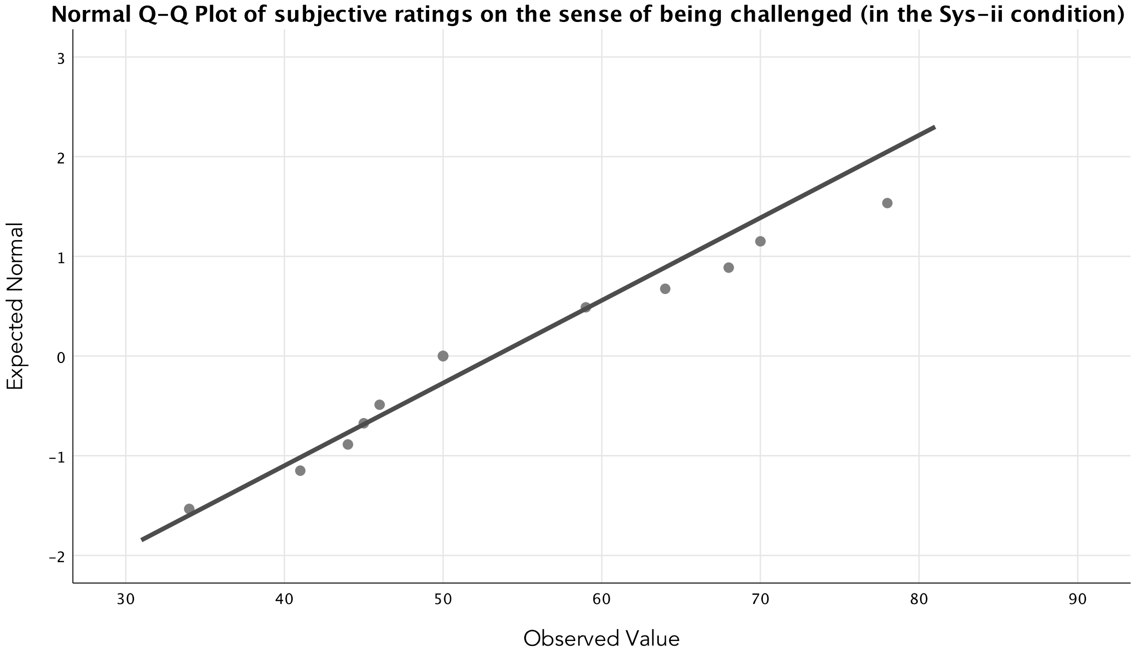


(a)

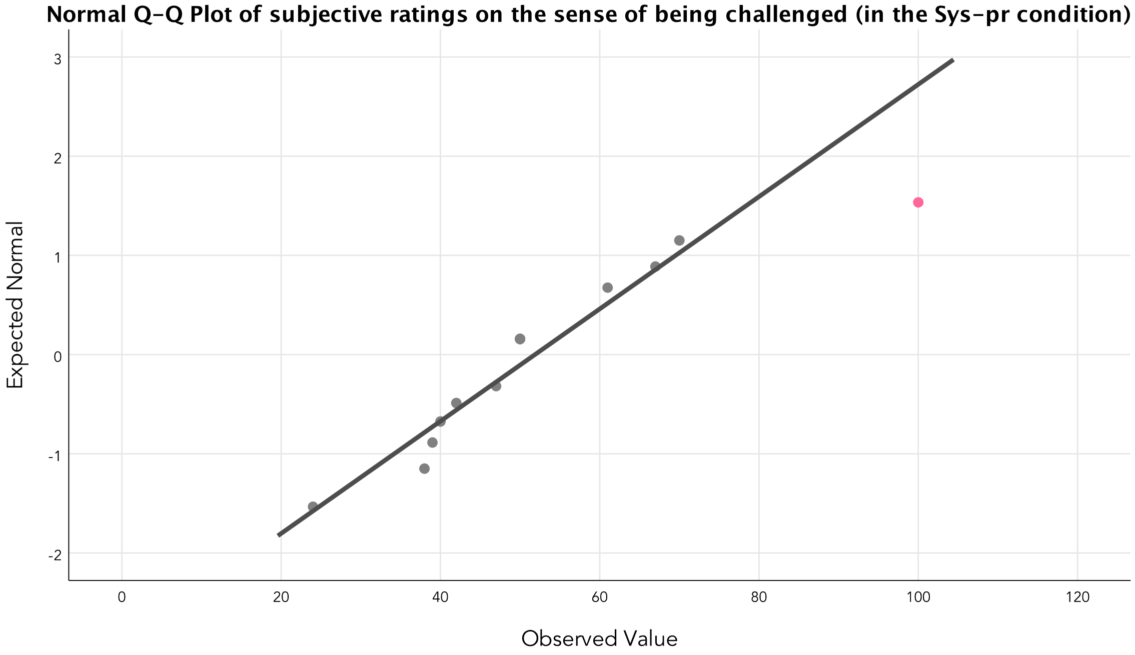


(b)

Figure B.4: Q-Q plots for the interval length between two adjacent labelling actions in the (a) **Usr-Sys** condition and (b) **Usr-r** condition in Experiment 3

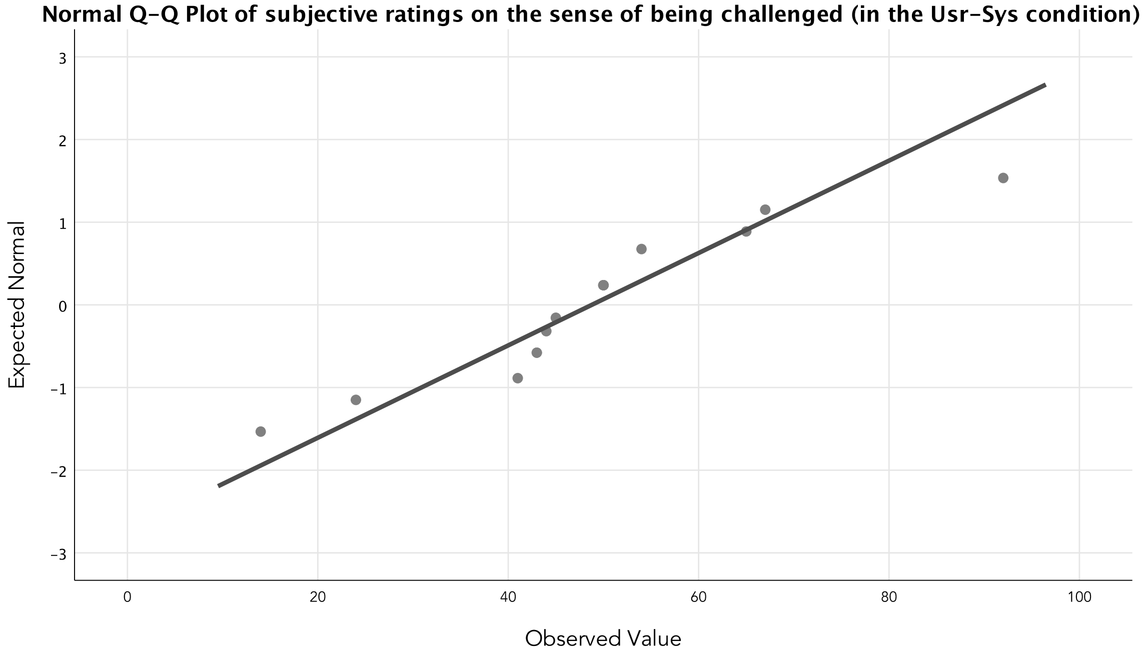


(a)

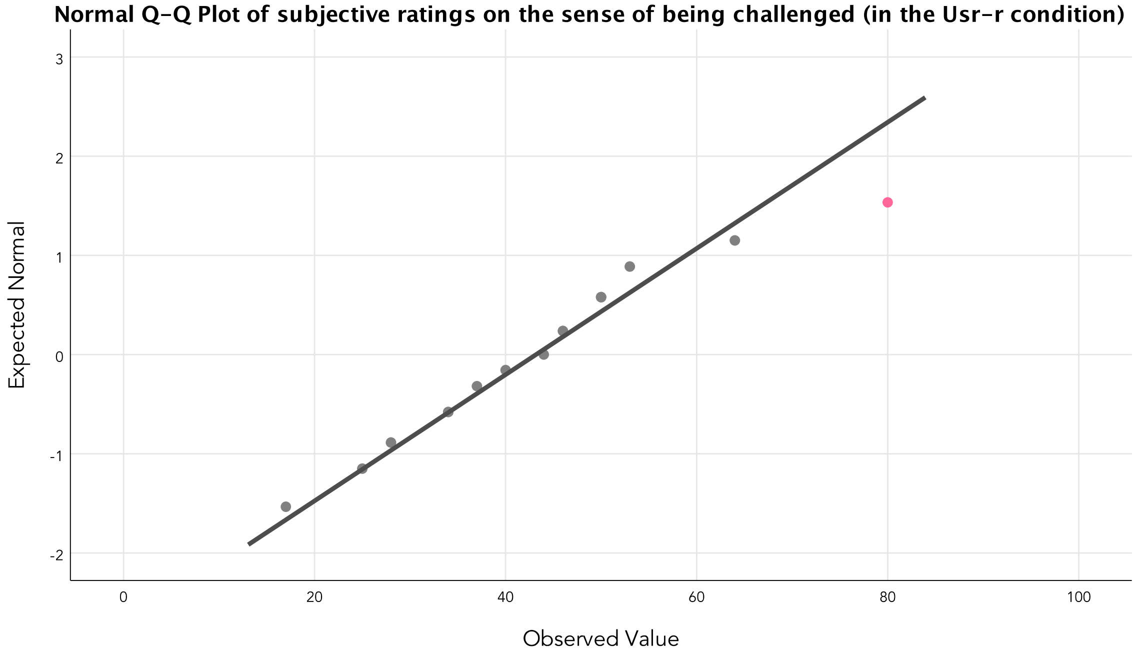


(b)

Figure B.5: Q-Q plots for the subjective ratings on the sense of being challenged in the (a) **Sys-ii** condition and (b) **Sys-pr** condition in Experiment 3



(a)



(b)

Figure B.6: Q-Q plots for the subjective ratings on the sense of being challenged in the (a) **Usr-Sys** condition and (b) **Usr-r** condition in Experiment 3