Understanding Strategic Adaptation in Dual-Task Situations as Cognitively Bounded Rational Behavior

Christian Pieter Janssen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
UCL

April 2012

I, Christian Pieter Janssen confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In this thesis I explored when people interleave attention in dual-task settings. The hypothesis is that people try to perform in a cognitively bounded rational way. Performance is limited by constraints that come from the task environment and cognition. If, given these constraints, multiple strategies for interleaving tasks are available, then people will interleave tasks in a way that aligns with their local priority objective (Chapter 3), or which maximizes the value of an objective payoff function that evaluates performance (Chapter 4). This hypothesis was tested using a combination of experimental studies and computational cognitive models. Across a series of studies, the interplay between different constraints was investigated. In Chapters 5 and 6, I developed mathematical models to study what task combinations in general allowed for "ideal payoff manipulations" to study task interleaving. The work contributed to the existing literature in four ways: (1) it provided an overarching theory of skilled human dual-task performance and tested this in relatively applied settings, (2) the theory was formalized in computational cognitive models that can predict performance of unobserved strategies and that can bracket the (optimal) performance space, (3) linear and logarithmic tasks were identified as an ideal combination for achieving ideal payoff manipulations, and (4) results demonstrated that in multitasking situations attention is not necessarily interleaved solely at chunk boundaries and other "natural breakpoints", but that this depends on a person's priorities. The work has implications for driver distraction research, in that it helps in systematically understanding the performance trade-offs that people face when multitasking. Moreover, the modeling framework could be used for model-based evaluation of new mobile interfaces. Finally, the demonstration that priorities can strongly influence multitasking performance highlights the importance of public safety campaigns that emphasize awareness of driver safety. Limitations and further implications are discussed.

Index of Chapters

Detailed Index

List of Tables

List of Figures

Acknowledgements

I want to acknowledge the EPSRC, who funded the research reported in this thesis, through grant EP/ G043507/1, awarded to Duncan Brumby, PI.

But… of course I also want to thank many people! It is frequently stated that PhD research is, at times, a solitary experience. For several years you research a tiny problem about which you care a lot, while others do not care (yet!). May I therefore start by thanking you, reader, for taking the time to read my thesis! At the same time, I disagree that research is a solitary experience. Research always flourishes through collaboration and interaction with others. I have been fortunate to be surrounded by a great set of people, who I want to thank here.

First, and foremost I want to thank my daily supervisor, Duncan Brumby, for being a great mentor. Although I already tend to set high standards and expectations to my work, you managed to encourage me to set them even higher, but also to explore new avenues, communities, and opportunities.

I also want to thank my second supervisors, John Dowell and Nick Chater. Your different backgrounds has provided me with great insights, and has helped me to communicate my research to multiple communities.

Thank you also Andrew Howes. Through the years, you became more and more involved with my research. Your eye for detail and passion for research are an inspiration.

A big thanks also goes to Richard Young. You always had critical questions and insights. When during meetings some of us struggled to formulate a question or comment, you were able to rephrase it, and explain why it was one of the most relevant questions of the day.

Also a big thanks to all other members at UCLIC. Thank you Anna, for taking part in my intermediate exams. Thank you Ann and Yvonne, for leading UCLIC. Thank you Louise and Romy, for making sure that everything always runs smoothly. Thank you Justin, for being part of the team researching 'interactions on the move'. And thanks to all others who joined me at 'formal' meetings such as our reading group (with cookies instead of biscuits), or at informal meetings such as nights at the Birkbeck bar, karaoke, or musicals (Priscilla is fabulous!). Thank you Abdi, Aisling, Andrea, Aneesha, Annina, Charlene, Chris, Dominic, Eddie, Eduardo, Enzian, Harry, Jenn, Jo, Jon, Jonathan, Maartje, Paul, Sandy, Sarah, Shakil, Simon, Stephann, and Stephen.

I also had the pleasure to supervise bright BSc and MSc students. In particular, work that was conducted together with Rae Garnett is part of Chapter 3 of this thesis. George Farmer's MSc work built on the work I report in Chapter 4, and has shaped my thinking about these problems, datasets, and models. Also thanks to all other students: Nina, Samantha, Elaine, and Shuo. I hope you learned as much from me as I learned from you.

Thanks to the organizers and participants of the PALS PPG group for doing a great job. In particular, thank you, Rose, Zed, Gerrie, and Chris. I could always count on you to also show up at social events.

Thanks also go to those participants in my studies who showed up on time and listened to instructions.

There are also many people outside of UCL who I want to thank. Thanks to all that provided me with their thoughts about my work at conferences, workshops, and other meetings. Thank you to the organizers and participants of the doctoral consortia at ICCM 2010 and CHI 2011.

Thank you, Dario Salvucci. Not only did you give me useful feedback at multiple occasions, you also provided Duncan and me with the initial code for the driving simulator that is used in Chapter 3. Having this available at the start of my PhD made running the first series of studies a lot easier.

Thank you Wayne Gray, for the great continued collaboration since I was part of your lab in 2008. A word of thanks also goes to the Cognitive Modeling group and the Artificial Intelligence department of the University of Groningen. Being part of such a great and dynamic research group as BSc and MSc student has provided essential fundaments for my current research. In particular, thank you: Hedderik, Niels, Fokie, Leendert, and Jelmer.

During my PhD I was also fortunate to be an intern at PARC and at Microsoft Research. I want to thank my mentors, Peter Pirolli (PARC), and Shamsi Iqbal (Microsoft). It was great to work with you, to gain new knowledge and skills, and to see how the knowledge and skills that I gained during my PhD research could be applied to other settings. Also a big thanks to the other people that I've met during these internships. Thank you Kathy for making my life so much easier by being a wonderful host and friend.

Yes, I'm a workaholic, and many people have helped me to maintain a balance between work and non-work. Thank you to all my

friends who have supported me. Thank you for putting up with me not being with you as much as I'd like to. In particular, a big thanks to those of you who took the effort to visit me in London: Merel, Nanda, Jolie, Myrthe, Katrin, Martin, and Jelmer.

Thank you also to my family. Thank you mom and dad for your continuous love and support, even when times were difficult for you. Mom, thank you for being a fantastic example of a hard worker. Dad, thank you for being a great example of good entrepreneurship. Thank you also Ingrid and Wolfgang, for welcoming me wholeheartedly into your family. Last, thank you Tobi for so many things. For encouraging me to follow my dreams, for balancing me out when I'm either too excited or feeling down, for giving me joy every day, and for putting up with me living too far away from you, with a budget airline in between us. Let's hope that will change soon.

List of Publications

My research has resulted in the following publications (sorted by type and year of publication). Research that is described in this thesis is marked with an asterisk.

Journal Papers

* Janssen, C. P., Brumby, D.P., & Garnett, R. (2012). Natural break points: The influence of priorities, and cognitive and motor cues on dual-task interleaving. *Journal of Cognitive Engineering and Decision Making, 6*, 5-29. (reported in Chapter 3)

Janssen, C. P., & Gray, W.D. (2012). When, what, and how much to reward in reinforcement learning based models of cognition. *Cognitive Science, 36*, 333-358.[1]

* Janssen, C. P., Brumby, D. P., Dowell, J., Chater, N., & Howes, A. (2011). Identifying optimum performance trade-offs using a cognitively bounded rational analysis model of discretionary task interleaving. *Topics in Cognitive Science, 3*, 123-139. (reported in Chapter 4)

* Janssen, C. P., & Brumby, D. P. (2010). Strategic adaptation to performance objectives in a dual-task setting. *Cognitive Science, 34*, 1548-1560. (reported in Chapter 3)

Conference papers (peer-reviewed)

Brumby, D. P., Davies, S., Janssen, C. P., & Grace, J. J. (2011). Fast or safe? How performance objectives determine modality output choices while

[1] This research was initiated while I was a MSc student at the University of Groningen and a visiting scholar at Rensselaer Polytechnic Institute in 2008. It was continued while I was a PhD student at UCL.

interacting on the move In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole & W. Kellogg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 473-482). New York, NY: ACM Press.

Farmer, G. D., Janssen, C. P., & Brumby, D. P. (2011). How long have I got? Making optimal visit durations in a dual-task setting. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 862-867). Austin, TX: Cognitive Science Society.

Janssen, C. P., Brumby, D. P., Dowell, J., & Chater, N. (2010). A cognitively bounded rational analysis model of dual-task performance trade-offs*.* In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 103-108). Philadelphia, PA: Drexel University.

(winner Allen Newell Award for best student paper)

Janssen, C. P., Brumby, D. P., & Garnett, R. (2010). Natural break points: Utilizing motor cues when multitasking *Proceedings of the 54th annual meeting of the Human Factors and Ergonomics Society* (pp. 482-486). San Francisco, CA: Human Factors and Ergonomics Society.

Brumby, D. P., del Rosario, N., & Janssen, C. P. (2010). When to switch? Understanding how performance tradeoffs shape dual-task strategy*.* In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the International Conference on Cognitive Modeling* (pp. 19-24). Philadelphia, PA: Drexel University.

Selected Position Papers (peer-reviewed)

* Janssen, C. P., Howes, A., & Brumby, D. P. (2012). Towards optimal payoff manipulations. In *Proceedings of the 11th International Conference on Cognitive Modeling*. Berlin: Universitaetsverlag der TU Berlin. (reported in Chapter 5)

Janssen, C. P., Brumby, D. P., & Howes, A. (2012). Towards a better understanding of adaptive multitasking by individuals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts*. New York, NY: ACM Press.

Janssen, C. P., & Brumby, D. P. (2011). Design and make aware: Virtues and limitations of designing for natural breakpoints in multitasking settings. In *Proceedings of the Automotive User Interfaces 2011: Workshop on Cognitive Load and In-Vehicle Human-Machine Interaction*.

Janssen, C. P. (2011). Understanding multitasking as an adaptive strategy selection process. In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole & W. Kellogg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1041-1044). New York, NY: ACM Press.(CHI Doctoral Consortium)

Book reviews

Janssen, C.P. (2012) The quest for a unified theory of multitasking. The multitasking mind, Salvucci, Taatgen. Oxford university press (2011). *Cognitive Systems Research*, *13*, 95-97.

List of Contributions to the Literature

This thesis aims to make the following contributions to the literature. Each contribution is discussed in detail in the referenced Chapters.

<u>General contributions, described in Chapter 7:</u>

7.1 A theory of skilled dual-task behavior, tested in two relatively applied settings

7.2 Refinement of a methodology to predict human multitasking performance and application to two dual-task scenarios

7.3 Identification of linear and logarithmic tasks as an ideal task combination for further study of ideal payoff manipulations

7.4 An investigation of interleaving at natural breakpoints

<u>Contributions made in Chapter 3</u>:

3.1 A series of critical tests to investigate whether people interleave if and only if they reach a chunk boundary

3.2 Interleaving at natural breakpoints offers valuable speed-accuracy trade-offs

3.3 Motor cues can form natural breakpoints

3.4 Critical reflection on "rationality" in distracted driving

Contributions made in Chapter 4:

4.1 Introducing a formal method to identify 'optimum' strategies in dynamic concurrent dual-task scenarios

4.2 A formal analysis of how task constraints, individual differences, and objective influence dual-task interleaving

4.3 A re-appreciation of the flexibility of human performance

Contributions made in Chapter 5:

5.1 A mathematical analysis of how constraints and payoff function systematically influence performance and payoff curves

Contributions made in Chapter 6:

6.1 An exploration of the usefulness of different task types for studying interleaving with ideal payoff manipulations

Chapter 1. Introduction

1.1. The prevalence of multitasking

Multitasking – performing two or more tasks at roughly the same time or in succession - is a prevalent aspect of our daily lives. At home, teenagers chat with friends while they write e-mails and watch TV (Foehr, 2006). In the office, workers switch between tasks about every two to three minutes (González & Mark, 2004). In the hospital, doctors are getting distracted by their mobile devices (Richtel, 2011). And on the road, drivers have been observed to perform all kinds of tasks while driving, from making phone calls to updating their social network status (e.g., Crowd Science, 2009; Diels, Reed, & Weaver, 2009). In fact, multitasking is so ubiquitous that some have argued that it is an integral part of human nature (Salvucci & Taatgen, 2011).

The ability to multitask has the intuitive appeal that "slack time" can be used to work on other tasks (see also Rattenbury, Nafus, & Anderson, 2008). For example, while waiting for a friend to respond to a chat message, an e-mail can be written. This can be faster than a situation in which you wait for the chat response, doing nothing, before writing the e-mail.

However, multitasking can also lead to dangerous situations. For example, the dangers of performing side tasks while driving have been noted for over two decades (e.g., Alm & Nilsson, 1995; Brookhuis, De Vries, & De Waard, 1991; Green, 2003; Horrey & Wickens, 2006; Lee & Strayer, 2004; McKnight & McKnight, 1993; Reed & Green, 1999; Strayer & Johnston, 2001). Unfortunately, despite this knowledge, people still engage in distracting secondary tasks while driving (e.g., Crowd Science, 2009; Diels et al., 2009; The Economist, 2011).

Might there be other reasons then for why people engage in such dangerous tasks? And how can people be encouraged to move their attention away from distracting side-tasks? Understanding these and other issues can inform the public debate on the benefits and dangers of multitasking. Ideally, it can also facilitate the design, prototyping, and evaluation of mobile technologies that are used in multitasking contexts. In this thesis I will make a contribution towards this large objective, by investigating a relevant sub question. I will investigate people's flexibility in multitasking performance, and how four general factors systematically influence when people interleave their attention between two tasks. These factors are task characteristics (Chapter 3-6), cognitive characteristics (Chapter 3-5), individual differences in skill (Chapter 4), and priorities or objectives (Chapter 3-6), see Chapter 2 for more details.

1.2. Research question: When do people interleave tasks?

One characteristic aspect of multitasking is that people can almost never pay complete attention to two tasks at the same time, due to amongst others physiological limitations (e.g., the eyes can only look in one direction at a time) (cf. Salvucci & Taatgen, 2008, 2011; Wickens, 2002, 2008). Therefore, multitasking requires people to interleave their attention between tasks. But what influences when attention is interleaved? Are people flexible in how they interleave attention, or is this pattern mainly dominated by the structure of the task?

In this thesis I will investigate this research question. Specifically, I will study four general factors that influence when people interleave tasks, and that thereby influence dual-task performance: task characteristics (Chapter 3-6), cognitive characteristics (Chapter 3-6), individual differences in skill (Chapter 4), and priorities or objectives (Chapter 3-6). Following work by Howes, Lewis, and Vera (2009) on skilled performance, the hypothesis is put forward that these four factors put bounds (limitations) on performance, and that given those bounds people will try to perform optimally, or rationally. As such, I try to understand human multitasking as Cognitively Bounded Rational Behavior.

In a series of studies, I will observe *when* people interleave their attention between two tasks. I will then develop computational cognitive models (or simply: cognitive models) to explore how performance changes when different patterns of interleaving are applied. By exploring this wider set of "strategies" for interleaving, I will try to explain *why* people interleaved in a certain way. I hypothesize that people perform boundedly rational. That is, the hypothesis is that people perform the best they can, given the bounds put on performance by the local context.

The above hypothesis on "optimality" should not be read as a claim that people can perform any combination of tasks without deterioration of performance. As said before, in every multitasking context there are systematic bounds on performance. I will investigate whether people do the best they can, *given these bounds*.

Performing boundedly optimal, unfortunately, also does not necessarily align with performing "safe". As will be discussed in detail in Chapter 3, sometimes people prioritize good performance on a distracting secondary task over good performance on a critical primary task. For example, they might prioritize dialing a number quickly over driving safely. As will be shown in Chapter 3, such performance can be considered "optimal" if performance on the driving task is the best possible given the fixed criterion for fast completion of the dialing task (cf., Navon & Gopher, 1979; Norman & Bobrow, 1975). However, this does *not* mean that this level of driving performance is the overall safest possible. The implications of this conclusion are discussed in Chapter 3.

1.3. Limitations

There are many settings in which multitasking has been studied, and inherently my thesis can only cover a small subset. In Chapter 2 I will introduce a framework by Salvucci and Taatgen (2011) to classify my work in the larger set of multitasking studies. Roughly, it is constrained in the following ways. First, I investigate *dual-task situations with concurrent multitasking:* where one task is interleaved for performance on a second task within seconds to minutes. Interleaving is mostly done in a discrete way, meaning that participants can only perform one task at a time.

The experiments capture aspects of driver distraction at *different levels of application*. In the studies reported in Chapter 3, participants drive a simulated car while manually dialing a phone number. In the studies reported in Chapter 4, participants track a cursor while typing digits on a keyboard. Both tasks contain a monitoring task (steering or tracking) and a manual typing task (on a phone or a keyboard). As the tasks are only *simulations* of real-world driver distraction, care should be taken when translating findings from my studies to the real world. I will discuss these limitations in the relevant Chapters.

The theories are specified at the cognitive band (Newell, 1990). This means that theory is described using cognitive processes that take between several milliseconds and several seconds. The description is *functional*, meaning that I describe the functions that are achieved by a person (e.g., achieving a goal, pressing a button), without making strong claims about how this process is biologically realized in the brain.

Within the broader Human-Computer Interaction (HCI) community the work should be seen within the tradition set by Card, Moran, and Newell (1983) of formally modeling human behavior. The emphasis is on understanding the cognitive science aspects of multitasking. Other branches of HCI, such as design and ethnography, will not be explored. Perhaps evident, the research is not a "psychology of driving". It is meant to address general aspects of multitasking that can apply to a wide range of domains, of which driver distraction is a useful illustrative example.

1.4. Structure of this thesis

The remainder of this thesis is structured as follows. The literature review in Chapter 2 will describe the domain, research question, and methodology in more detail. In Chapters 3, 4, 5, and 6, I report results from experimental and modeling studies on multitasking. In Chapter 3, I investigate how task, cognition, and objective influence dual-task interleaving and performance in a dialing-while-driving setting. Specifically, I will investigate under what conditions people interleave at “natural breakpoints” in the task structure. In Chapter 4, I will introduce a payoff function as a formal way of capturing a user’s objectives. Using a tracking-while-typing scenario, I will investigate how a payoff function and individual differences in skill constrain performance. In Chapters 5 and 6, I will discuss mathematical models to reflect on the general limitations of using payoff functions to study dual-task interleaving. Specifically, I will investigate whether manipulations of the payoff function can, in principle, make any arbitrary strategy optimal.

As may be evident, the task environments become more controlled and less applied throughout the thesis. This is because findings that were made in each Chapter sparked new research questions that were more appropriately addressed using more controlled experimental or modeling settings.

Each Chapter will discuss the contributions, implications, and limitations of the studies in that Chapter. In Chapter 7, I summarize my findings and discuss the broader contributions, implications, and limitations of this research. A list of the contributions that my research makes to the literature is also provided at the beginning of this thesis, with references to the Chapters that cover that contribution.

Chapter 2. Literature Review

Abstract

This literature review discusses the domain, research question, and methodology of the research in this thesis. The domain is concurrent dual-tasking of an analog monitoring task and a discrete typing task. Within this domain, I investigate when people interleave their attention; that is I investigate what strategy people choose for interleaving two tasks. Following the work by Howes, Lewis, and Vera (2009), I hypothesize that this choice can be understood as being cognitively bounded rational. That is, the strategy choice and resulting performance is subject to constraints that come from the task environment and cognition. I hypothesize that within those constraints people try to apply the strategy that maximizes utility, as evaluated using an explicit priority or an explicit payoff function. As a methodology I combine experimental studies with cognitive models.

2.1. The domain: Concurrent dual-tasking of a continuous monitoring task and a discrete typing task

Multitasking has been studied in a variety of domains and settings. This makes it hard to pin down an exact definition of the word "multitasking". In all cases, it involves performing more than one task, on which a user works either concurrently or in succession. To help in the classification, Salvucci and Taatgen (2011) introduced three orthogonal continua to distinguish different multitasking situations. In

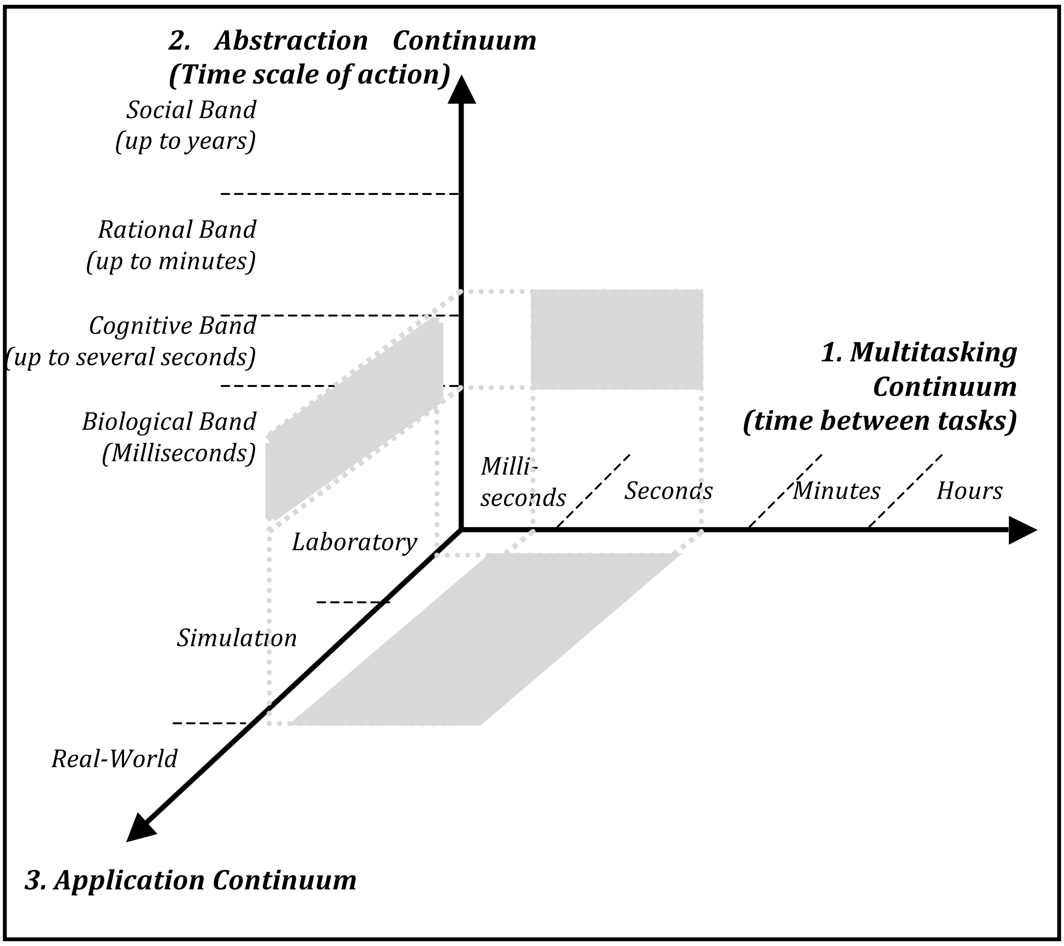

Figure 2.1: The three multitasking continua, as identified by Salvucci & Taatgen (2011).
The grey areas project the area that is occupied by the studies in this thesis onto two-dimensional space. Note that the boundaries between different sections on the continua and between my research area and other areas are not strict boundaries but approximations.

this section I describe these continua and position my work within them. Figure 2.1 provides a graphical representation of the three continua, with the area that is covered by my research projected onto two-dimensional space. The division into sections is meant as a rough indication of areas, not as an absolute boundary.

The first continuum is the multitasking continuum (Salvucci & Taatgen, 2011; Salvucci, Taatgen, & Borst, 2009). It classifies multitasking situations based on the amount of time that is spent on one task before moving attention to the next task. I investigate concurrent task interleaving: situations where attention is moved between tasks within a couple hundred of milliseconds up till a couple of seconds.

The second continuum is the abstraction continuum, which identifies the level of abstraction at which the theory is posited. Salvucci and Taatgen (2011) distinguish regions using Newell's four bands of cognition: the biological, cognitive, rational, and social band (Newell, 1990). Each band differs in the timescales over which the associated processes take place, from neurological processes (biological band) up to completion of an entire task (rational band) and beyond. The chosen level influences both the level of detail of the theory and the type of data that is needed to validate it. For example whether neurological data (biological band), or behavioral data (cognitive band) is needed. My theory is posited at the intersection of the cognitive and the rational band. The cognitive band describes processes that take place over hundreds of milliseconds up to a couple of seconds, for example eye-movements and key presses. The rational band is used to explain behavior based on the goals (or objectives) that are pursued by a person, and the rewards (or payoff) that can be obtained this way.[2]

---

[2] Marr's three levels classification is also a commonly used framework to specify level of abstraction (Marr, 1982). The mapping to this framework is less straightforward. It is somewhere around the computational level (as it explores the rationale for applying particular strategies) and the representation/algorithmic level (as it specifies the steps that the model goes through when executing a strategy).

The third continuum is the application continuum, which describes how applied the setting is in which the theory is tested. All studies in this thesis simulate aspects of driver distraction. Specifically, all involve a continuous monitoring and control task (e.g., keeping the car in the middle of the lane) and performing a discrete typing task (e.g., manually dialing a phone number).

The studies are placed at different levels of application, which I will now describe. In Chapter 3, I report how people interleave their attention between manually dialing a phone number and steering a simulated vehicle in a low fidelity desktop-based driving simulator. This setting is less applied than driving studies in a high-fidelity driving simulator (e.g., as in Iqbal, Ju, & Horvitz, 2010; Strayer, Drews, & Johnston, 2003), studies on a test-track (e.g., Horrey & Lesch, 2009), and studies on the road (e.g., Brookhuis et al., 1991).

The task environment in Chapter 4 is on the lower end of the application continuum. In this task environment participants have to track a cursor using a joystick and type in digits using a keyboard. Tracking tasks have been used frequently to study multitasking performance (e.g., Ballas, Heitmeyer, & Pérez-Quiñones, 1992; Chong, 1998; Chong & Laird, 1997; Gopher, 1993; Hornof, Zhang, & Halverson, 2010; Kieras, Meyer, Ballas, & Lauber, 2000; Lallement & John, 1998; Martin-Emerson & Wickens, 1992; Salvucci & Taatgen, 2008; Strayer & Johnston, 2001).

A benefit of my particular task environment was that the interleaving of attention could be measured directly. In addition, because of its abstraction, participants did not bring in natural ways of interacting with the task (in contrast to driving, where people have "natural ways" of performing the task, such as staying in the lane boundaries). This allowed exploration of a broader spectrum of performance.

A more abstract description of tasks will be used in Chapters 5 and 6. In Chapter 5, I report a mathematical model of a tracking-while-typing task, in which movement of the cursor is predicted using Pascal's triangle. In Chapter 6, I use an even more abstract setting, in which tasks are described using mathematical equations of the underlying performance functions (e.g., linear, logarithmic, exponential), without relating the models to a concrete task environment.

2.2. The research question: When do people interleave attention between two tasks?

2.2.1. How are strategies selected?

Theories of strategy selection and their commonalities

The core question of this thesis is to better understand when people interleave their attention between two tasks and to offer an explanation for this behavior. To investigate this, I will use cognitive models (see section 2.3) to explore performance of a variety of different "strategies" for interleaving. This analysis will be used to explain why participants applied particular strategies.

With a strategy I do not refer to a deliberate reasoning process, as was done in more traditional problem solving studies. Rather, I consider it to be a sequence of mental and physical actions that leads up to a specific behavior, which by an outside observer can be labeled as a coherent strategy (Young, 1978). This is a similar approach as in for example Gray, Sims, Fu, and Schoelles (2006), Howes et al. (2009), and

Walsh and Anderson (2009). In my domain in particular, a strategy is characterized by the amount of time spent on one task before moving attention to the next task.

Often, there are many strategies available for performing a task. In a multitasking context these typically range from strategies that interleave frequently between tasks, to strategies that do not interleave at all. Given that there are multiple strategies, how do people decide which strategy to apply? Many theories of strategy selection exist.

In previous work I have classified the commonalities in the literature on strategy selection as follows (Janssen, 2008; see also Janssen & Gray, 2012). Strategy selection always involves a *choice* among alternatives. Note that, given my definition of a "strategy", such a choice does not necessarily involve a deliberate (symbolic) reasoning process (Newell & Simon, 1972). Rather, it can also be achieved through experiencing and learning the (subsymbolic) value of different actions, in achieving goals (Sutton & Barto, 1998).

What the alternatives (or individual strategies) are differs between settings, based on the goal that needs to be achieved. For example, a goal can be to make a photocopy of a page, to solve a problem such as the Tower of Hanoi, or to dial a phone number while steering a car. Despite the differences in what goals are achieved, all theories on strategy selection tend to agree that competing strategies always work towards completion of goals (e.g., Erev & Barron, 2005; Erev & Gopher, 1999; Gonzalez, Lerch, & Lebiere, 2003; Gray et al., 2006; Howes et al., 2009; Lovett, 1998; Marewski & Schooler, 2011; O'Hara & Payne, 1999; Rieskamp & Otto, 2006; M. J. Roberts & Newton, 2001; Siegler, 1991; Sperling & Dosher, 1986; Walsh & Anderson, 2009; Young, 1978).

Different theories also tend to have different ways of representing strategies (see Sperling & Dosher, 1986, for a discussion of some alternatives). In some sense, different representations of strategies can be seen as describing a phenomenon at different levels of abstraction (Marr, 1982; Newell, 1990).

What makes people then settle down on specific strategies? One idea is that people are flexible in their decisions (e.g., Erev & Gopher, 1999; Gopher, 1993; Moray, Dessouky, Kijowski, & Adapathya, 1991; Navon & Gopher, 1979; Norman & Bobrow, 1975; Norman & Shallice, 1986). For example, when driving a car and performing a task, people can adjust the speed of the car to allow for more or less time to perform secondary tasks (Cnossen, Meijman, & Rothengatter, 2004; Iqbal et al., 2010).

Cognitively Bounded Rational Analysis: a theoretical framework for exploring strategy selection

Computational theories of strategy selection are starting to appreciate this flexibility more and more (e.g., Gray et al., 2006; Howes et al., 2009; Janssen & Gray, 2012; Kieras, Meyer, Ballas, & Lauber, 2000; Marewski & Schooler, 2011; Meyer & Kieras, 1997a, 1997b, 1999; Smith, Lewis, Howes, Chu, & Green, 2008). The basic idea in all these theories is that the appropriate set of strategies for any specific context is constrained by internal and external forces. In the next sections I will discuss some of these constraints in more detail.

To help in the classification, I will follow the distinction made by Howes, Lewis, and Vera (2009), in their theory of cognitively bounded rational behavior. This theory is chosen because it is most explicit about the role of constraints in performance. The theory states that for skilled human performance, strategy selection can be understood as a utility maximization problem that is constrained by the task environment and cognition. That is, it is hypothesized that, given the constraints on performance, people try to select those strategies that maximize their gains.

The theory has previously been tested in the psychological refractory period (PRP) paradigm (Howes et al., 2009; Howes, Vera, Lewis, & McCurdy, 2004; Lewis, Vera, & Howes, 2004; Vera, Howes, McCurdy, & Lewis, 2004). In this paradigm, participants have to respond to two stimuli in a specific order. The timing of the secondary stimulus is manipulated experimentally using a stimulus onset asynchrony (SOA). The typical result, the PRP effect, is that people's response time to the stimulus of the primary task is unaffected by the SOA, while the response time to the secondary task is dependent on the SOA. At short SOAs the response to the secondary stimulus is delayed. Howes, Lewis, and Vera (2009) demonstrated how the PRP effect can be explained as a participant's adaptation to the constraints that come from the task (e.g., the SOA) and cognition (e.g., the noise in motor movement). Within these constraints, people try to optimize their performance score.

However, in some respects the PRP task is simple: stimuli appear at their own pace and single responses need to be made. Slightly more complex are dynamic discretionary task interleaving scenarios, where participants need to decide themselves when to switch attention from one dynamic task to another. In this thesis I will contribute to the literature by testing the theory of cognitively bounded rational analysis in such settings.

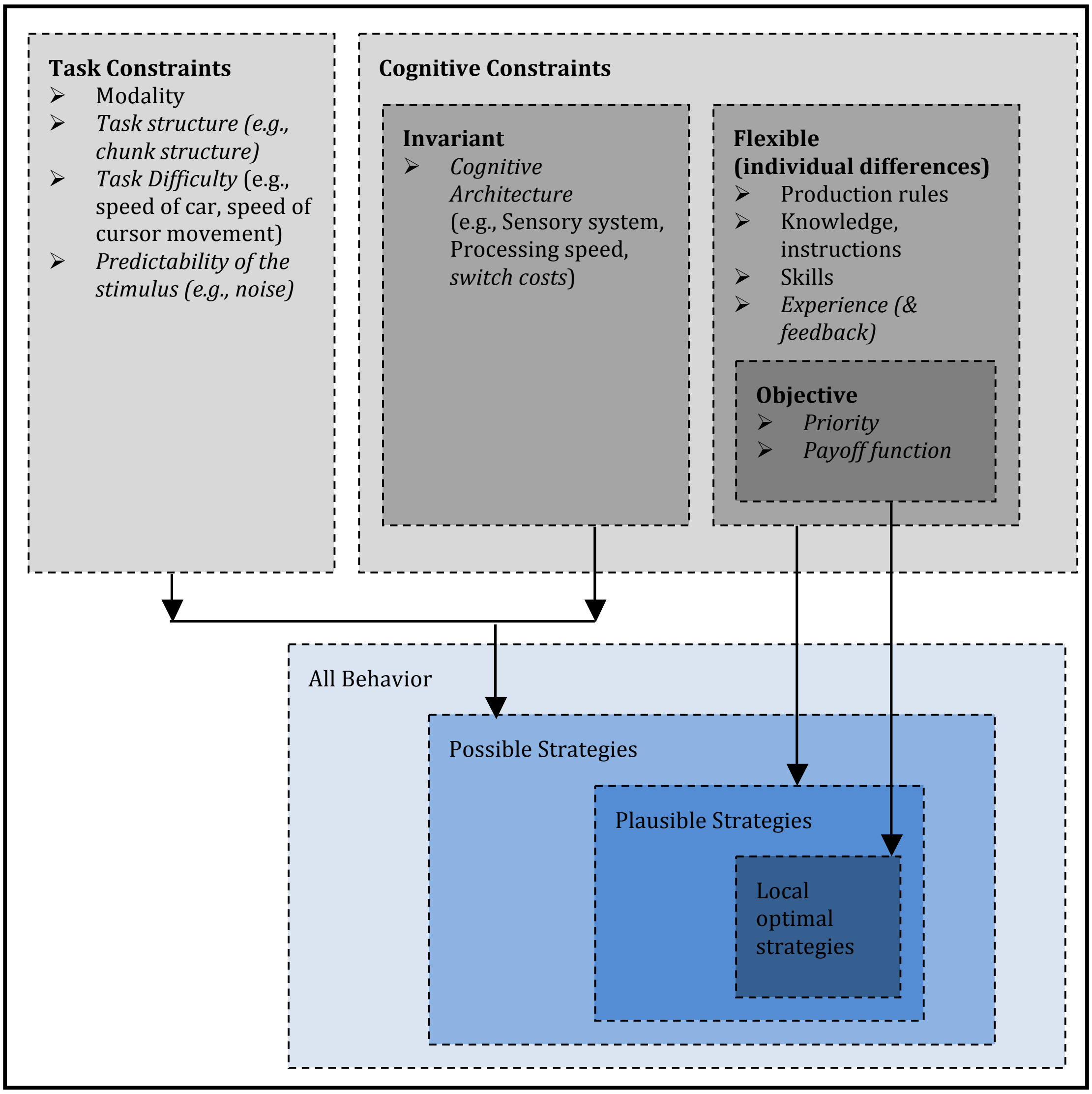

Figure 2.2: The task constraints and cognitive constraints that narrow down the space of all possible behavior to those strategies that are locally optimal. Examples of constraints that are investigated in this thesis are written in italics.

Figure 2.2 (inspired by Figure 4 in Howes et al., 2009) summarizes the general factors (grey boxes) that constrain performance (Venn diagram with blue boxes). These are task constraints and cognitive constraints. Within cognitive constraints I distinguish invariant factors and flexible factors. Within the flexible factors, a special role is reserved for objectives. Below I will go through each of these constraints, explain how they put limits on performance, and provide relevant examples for a multitasking context. This list is not meant to be exhaustive, but to illustrate some of the most common and interesting constraints.

2.2.2. Task constraints (or: data-limitations)

The blue Venn diagram in Figure 2.2 represents the collection of behaviors (labeled "all behavior") that can be achieved across various agents (e.g., animals, humans, artificial agents) in various task environments. Task constraints narrow this space down to a set of "possible strategies", as each task environment, or interface, only allows a specific set of interactions. That is, behavior is data-limited (Norman & Bobrow, 1975). Specific examples of task constraints are mentioned next.

A straightforward example is the modality in which a task is presented and in which it can be interacted with. Audio-vocal tasks require different interactions (i.e., sound, speech) than visual-manual tasks (i.e., glances, button presses) (Wickens, 2002, 2008).

Similarly, the difficulty of a task also restricts performance. In a driving environment for example, the demands posed by the driving environment have been noted to influence performance (e.g., Cnossen et al., 2004; Iqbal et al., 2010). For example, a car that is driving at high speed might be harder to control than a car driving at low speed. Similarly, driving in a busy metropolis such as London might be more challenging than driving on a deserted highway.

Another factor might be the pacing and time pressure of a task. Tasks that have a high pace, or in other forms require a fast response, inherently will need more urgent attention (Moray et al., 1991). Conversely, tasks that are slow paced might be annoying (e.g., waiting for a web browser to load, Nah, 2004), promote less frequent interaction with the interface (e.g., Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Pook, & Rao, 1997; Gray et al., 2006), or even lead to increased distraction (e.g., driving on a deserted road, Atchley & Chan, 2011).

Related to the above factors is the predictability of a stimulus. The more noisy stimuli are, the more difficult it is to predict their effects and to respond appropriately. Studies have shown that increased noise makes it harder to identify optimal strategies, for example in the Psychological Refractory Period paradigm (Kopecky, 2008), in (rapid) motor-decision tasks (e.g., Maloney & Mamassian, 2009; Maloney & Zhang, 2010; Trommershauser, Maloney, & Landy, 2003a, 2008; Zhang, Daw, & Maloney, 2011), and in simple speeded two-choice decision tasks (e.g., Bogacz, 2007; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz, Hu, Holmes, & Cohen, 2010; Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010). See also a discussion in Tseng, Howes, and Lewis (submitted). These findings are in line with predictions from signal detection theory that a stimulus (and the necessary response) is easier to detect if the level of noise surrounding the stimulus is low (Swets, Tanner, & Birdsall, 1961). In Chapter 4, I will directly manipulate the difficulty and predictability of the task, by changing amongst others the level of noise in the stimulus.

Before that, in Chapter 3 I will look into a different type of task constraint: the task structure. As will be discussed in more detail there, it is believed that the task structure can provide “natural” breakpoints for interleaving tasks in a dual-task setting (e.g., Bailey & Iqbal, 2008; Bogunovich & Salvucci, 2010; Iqbal & Bailey, 2007, 2010; Salvucci, 2005). In a driving context in particular, it has been proposed that people tend to interleave solely at these natural breakpoints (Salvucci, 2005). In Chapter 3, I will perform a series of experiments to test the generality of this hypothesis.

Note that the effect of task constraints should always be seen in relation to the characteristics of an individual. Indeed, the formation of (unit) tasks is a man-made construct (Card et al., 1983). Similarly, the perceived difficulty of a task can be different between individuals. Despite this potential ambiguity, I classified the above factors as task constraints, as they are foremost characteristics of the system, which can be changed across systems, but are harder to change within the individual.

Another caution that should be taken is that task constraints are not always fixed in the real world. Indeed, in more naturalistic environments people have been observed to decrease the limitations put on them by the task environment, by making changes to their task environment (e.g., Kirlik, 1998) (called environment enrichment in Pirolli, 2007). For example, when writing a literature review, the most relevant papers might be gathered directly next to one’s computer (see also Pirolli & Card, 1999, for a discussion of studies that noted environment enrichment in office settings). In simulated driving studies, people have been observed to compensate for unsafe multitasking by for example reducing their speed (e.g., Cnossen et al., 2004; Iqbal et al., 2010).

The task environments that I used in the current studies did not allow for such structural modifications of the environment. This choice was made to allow for more control over the experiment. In general, it is expected that many of these potential changes to the environment are made at a distinct point in time (e.g., buying a bigger screen to be better able to read from it, or reducing car speed when entering a village). Given these changes, one can recharacterize the task constraints and their changed effect on the set of available behaviors.

2.2.3. Cognitive constraints: invariant characteristics (or: resource limitations)

Cognitive constraints can further narrow down the strategy space. I will distinguish invariant and flexible components of cognition (similar to e.g., Meyer, Glass, Mueller, Seymour, & Kieras, 2001). Invariant components are those that are rather fixed within an individual, but occasionally differ between individuals. As such, people do not have strategic control over these factors. Flexible components do vary within and between individuals, and some of these can be controlled strategically. I will discuss both types in isolation below.

The invariant components of cognition put limitations on what strategies are possible *for a particular agent* (see Figure 2.2). For example, the characteristics of a human's sensory system limit how much one can sense (e.g., what proportion of the visual field can be seen in detail), and the processing speed (e.g., production rule firing rate) sets a minimum cut on how fast a person can respond to a stimulus. More generally, the cognitive architecture (Newell, 1990) puts resource-limits on performance (Norman & Bobrow, 1975).

A noteworthy aspect of cognitive constraints is again noise. Similar to how noise of the task environment can influence performance (e.g., Kopecky, 2008; Swets et al., 1961), the cognitive system also has noise in the performance of actions. This has been the topic of many recent research efforts in the cognitive science community (e.g., Howes et al., 2009; Trommershauser et al., 2003a; Trommershauser, Maloney, & Landy, 2003b; Trommershauser et al., 2008), see also a discussion in Tseng et al. (submitted).

For example, in the work by Trommershauser, Maloney, and Landy, participants had to touch a target on a touch-screen within a specific time frame. They were rewarded if they hit the target, but received a penalty when they hit a penalty area. The nearness (and overlap) of the reward and penalty areas was varied between settings.

Due to noise in the motor system, participants would not always hit the exact spot that they aimed for. Therefore, if the reward and penalty areas were close together, people had to consider where they aimed and whether this could result in a penalty due to motor noise. Results showed that people took these aspects of noise into account in aiming their movements so as to achieve rewards, while avoiding penalties (Trommershauser et al., 2003a, 2003b, 2008).

2.2.4. Cognitive constraints: flexible aspects (or: individual differences)

Task constraints and invariant cognitive constraints put hard constraints (Gray et al., 2006; Howes & Young, 1997) on performance, because people can often not exert strategic control over these constraints. Flexible aspects of cognition are "soft constraints" (Gray et al., 2006; Howes & Young, 1997), that can change within an individual and over multiple tasks and situations (Meyer et al., 2001). In that way, they narrow down the set of possible strategies to a set of *plausible* strategies (see Figure 2.2).

In the modeling framework of Meyer and colleagues (Meyer et al., 2001) the flexible characteristics (which they call "software") is the set of available production rules (or condition-action pairs, Newell & Simon, 1972). The set of available production rules and the knowledge of the user can differ between settings. Based on the knowledge that a participant has about the task, and the instructions they received, they might only know about a subset of the available strategies. For example, imagine a situation where an interface allows for both manual and verbal interaction, but where a user only knows about the manual interaction. In this case, it is not plausible that the user will use any strategy that involves verbal interaction.

Similar to knowledge, skills might also influence performance. Recent empirical studies have found individual differences in multitasking ability (Ophir, Nass, & Wagner, 2009; Watson & Strayer, 2010), but the sources of these differences are unknown. Some researchers have proposed that there might be a general "multitasking skill" that can be trained (e.g., Gopher, 1993), whereas others have proposed that no such explicit control component is needed to explain performance (e.g., Salvucci & Taatgen, 2011). In Chapter 4 I will investigate how differences in *measurable* skills (e.g., typing speed) change the ways in which people can perform in dual-task scenarios.

Skills and performance can change with experience. This is due to at least two effects: (1) practice with the task and (2) changes of preferences. Practice effects arise with experience, because people become more skilled at the tasks they do. For example, with experience, people typically have to rely less on instructions and memory to guide actions and decisions (e.g., see Taatgen, Huss, Dickison, & Anderson, 2008, for a discussion and computational model of skill learning).

For the development of preferences, a particular useful framework is the theory of reinforcement learning (Sutton & Barto, 1998). This framework has been applied increasingly in cognitive science (Daw & Frank, 2009) and neuroscience research (e.g., Cohen, 2008; Schultz, 2006). Reinforcement learning frameworks provide a computational way of describing how the value of strategies (or actions, or action components) can be associated to experienced feedback and rewards. It is assumed that components of the strategy that preceded a reward contributed to the value of this reward. In this process, the temporal distance between the components and the reward is taken into consideration. Components that were executed closer to the experienced reward get more strongly associated with the experienced reward than more distant components.

More generally, the development of preferences can then be described as follows. If a participant has had many successful encounters with a strategy in the past, these strategies tend to have a high utility value, and tend to be *exploited* in future situations. However, if successes have been varied, or experience with alternative strategies is low, the utility of these strategies will be low, and people tend to *explore* more alternatives (the exploration-exploitation trade-off, see for example Sutton & Barto, 1998). This consideration of the sample of experiences for future decisions is also in line with other theories of decision making, such as decision by sampling theories (e.g., Stewart, Chater, & Brown, 2006).

In this thesis, I will not provide a detailed theory of how participants change their performance over time. This is a separate research topic in itself. However, two critical insights from reinforcement learning theories will be taken into account. The first insight is that people adapt their performance to experienced feedback and rewards. The second insight is that the nature of the reward (e.g., is time or accuracy rewarded) might influence what aspects of a task are optimized (Janssen & Gray, 2012).

2.2.5. Objectives, priorities, and payoff functions

The above constraints on performance can sometimes still allow for a variety of strategies to be applied. In a multitasking situation in particular, each strategy requires trade-offs to be made in terms of how well each task is performed, for example in terms of task time or number of errors made (Navon & Gopher, 1979; Norman & Bobrow, 1975). In order to narrow the space of possible strategies further down, an explicit priority or objective payoff function can be introduced. These help in assessing the relative success of a strategy for attention allocation between tasks. In the Venn diagram in Figure 2.2, objectives narrow the set of plausible strategies down to a set of *local optimal strategies*.

One way of inducing such effects is by stating an explicit priority objective. In dual-tasks scenarios, such a statement can say which of the two tasks needs to receive more attention compared to the other task.

The more specific the performance criterion for one task is (e.g., "spend 75 percent of your time on task A"), the narrower the space of strategies that adhere to these constraints becomes (e.g., Gopher, 1993; Gopher, Brickner, & Navon, 1982). I will use a similar method in Chapter 3. Specifically, it will be shown how stating an explicit priority avoids participants setting their own priorities, and how this can help to assess whether people complied with that priority.

Although priorities are useful, as illustrated in Chapter 3, their use also has limitations. In Chapter 4, I will discuss these limitations and propose the use of quantitative, explicit, objective payoff functions as an alternative method for identifying locally optimal strategies. These payoff functions combine performance on both tasks into a single feedback score in a consistent way. Applying a payoff function makes it easier to identify the optimum strategy: this is the one that achieves on average the highest payoff value. It does not involve subjective comparison of performance on two tasks.

Payoff functions have been used in experimental psychology before (e.g., Neth, Khemlani, Oppermann, & Gray, 2006; Schumacher et al., 1999; Wang, Proctor, & Pick, 2007, 2009) and are being applied more and more in computational frameworks. For example, in studies on the Psychological refractory period (Howes et al., 2009; Howes, Vera, & Lewis, 2007; Howes et al., 2004; Lewis et al., 2004; Vera et al., 2004), dynamic task interleaving (Hornof & Zhang, 2010; Neth, Khemlani, & Gray, 2008), motor movement (Jarvstad, Rushton, Hahn, & Warren, in press; Juni, Gureckis, & Maloney, 2011; Maloney & Mamassian, 2009; Maloney & Zhang, 2010; Trommershauser et al., 2003a, 2003b, 2008; Warren, Graf, Champion, & Maloney, 2012), and vision (Ballard & Sprague, 2007; Reichle & Laurent, 2006; Tatler, Hayhoe, Land, & Ballard, 2011). The work in this thesis builds on this preceding work and extends it by applying payoff functions to a dynamic task, while at the same time using the payoff function to make predictions about locally optimal performance.

2.2.6. Summary of research question

To summarize, my proposal is that multitasking is a form of skilled behavior, and as such is subject to several constraints (Howes et al., 2009). As illustrated in Figure 2.2, the local task context and invariant aspects of cognition constrain what types of strategies (or behaviors) are in principle *possible* for a particular type of agent. Flexible characteristics of cognition (e.g., differences in knowledge and skills) influence what types of interactions are *plausible* for a specific agent. Finally, payoff functions can be used to assess which set of strategies is *optimal* within a local context. These are the strategies that maximize the payoff (or utility), given the constraints. The hypothesis is that people will try to apply these optimal strategies (Howes et al., 2009).[3] Note that, due to the explicit consideration of constraints, the theory explicitly acknowledges that people's performance is not unbounded (Simon, 1956).

To constrain the scope of this thesis, I will not investigate in detail how performance adapts over time (i.e., how people learn to adapt performance). This is a large research question in itself. In Chapter 4 I will reflect on how theories of learning can enhance and complement the understanding of performance that is given there.

[3] A similar notion has also been proposed for artificial agents (Russell & Subramanian, 1995).

2.3. The approach: Cognitive modeling using cognitively bounded rational analysis

2.3.1. The rationale for cognitive models

I formalize my theory of multitasking (i.e., as specified in section 2.2 and Figure 2.2) in cognitive models. Cognitive models specify theories (about behavior in a specific task) in computer simulations. By specifying a theory in a computer simulation the claims get precise (cf. Newell, 1973; Newell, 1990), as they have to be specified as executable code. In addition, models can be adapted to make predictions for novel settings and environments. In effect, this helps to draw implications of the theory beyond the specific task setting at hand (e.g., Anderson, 2007; Gray, 2007; Kieras, in press). For these and other reasons, cognitive models have also been used frequently in the field of human-computer interaction, since the introduction by Card, Moran, and Newell (1983).

Throughout the literature, a variety of modeling frameworks has been applied to multitasking settings. Some of the bigger, process oriented frameworks are ACT-R (Anderson, 2007; Anderson et al., 2004), specifically the threaded cognition approach (Salvucci & Taatgen, 2008, 2011), SOAR (Newell, 1990) as in for example Chong (1998), Chong and Laird (1997), and Lallement and John (1998), and EPIC (Kieras & Meyer, 1997; Meyer & Kieras, 1997a), as in for example Hornof and Zhang (2010), Kieras et al. (2000), Meyer et al. (2001), and Meyer and Kieras (1999). I will use the framework of cognitively bounded rational analysis (Howes et al., 2009), and discuss that in more detail below.

2.3.2. Cognitively bounded rational analysis

Cognitively bounded rational analysis has been applied to multitasking situations, namely to study the Psychological Refractory Period task (Howes et al., 2009; Howes et al., 2004; Lewis et al., 2004; Vera et al., 2004) and to study driver distraction (Brumby et al., 2010; Brumby, Howes, & Salvucci, 2007; Brumby, Salvucci, & Howes, 2007, 2009). The models in this thesis follow the structure of the driver distraction models, which will be described in more detail in Chapter 3.

A feature of the framework is that it is explicitly designed to systematically explore performance of alternative strategies for executing a task and to evaluate which strategy is best in a local context. In order to do this, the framework requires specification of the relevant constraints on performance (see Figure 2.2). In a second step a set of plausible strategies for task performance that adheres to these constraints is derived. For each of these strategies the performance can be predicted. Finally, if an objective function is available then the performance of each strategy can be assessed, and the best performing strategies can be identified. The performance of these strategies can then be compared with human performance to assess whether humans achieved optimal performance as predicted by the model.

2.3.3. Comparison of CBRA with other modeling frameworks

The cognitively bounded rational analysis framework shares characteristics with other modeling frameworks, but also has differences. It shares with all the frameworks mentioned above (i.e., ACT-R, EPIC, SOAR) a methodology of formalizing cognitive behavior in computational models of the underlying processes (at the cognitive level, Newell, 1990).

As in the EPIC framework, the cognitively bounded rational analysis framework explicitly allows exploration of performance for *alternative* ways of performing tasks. Similar to the ACT-R framework (particularly the rational analysis part of it, Anderson, 1990), skilled behavior is thought to be optimal behavior. However, both approaches differ in what is optimized. ACT-R models assume that people adapt their performance to the (statistical) structure of the environment, following its foundations in rational analysis (Anderson, 1990; Oaksford & Chater, 1998)[4]. In contrast, cognitively bounded rational analysis models hypothesize that performance is optimized to payoff or utility, given the constraints put on performance by both the environment and cognition.

An important difference between cognitively bounded rational analysis and EPIC, SOAR, and ACT-R is that the framework does not require a theory to be specified in terms of specific fine-grained procedures (i.e., as a sequence of production rules). Rather, using constraint satisfaction techniques, a set of possible strategies can be defined and performance of each strategy can be tested.

This provides a way of dealing with an important problem in cognitive architecture research: the architecture-strategy credit assignment problem (Howes & Young, 1997), see also Howes et al.

---

[4] For more recent discussions on rational analysis see for example Griffiths, Chater, Kemp, Perfors, and Tenenbaum (2010), and Jones and Love (2011)

(2009). The architecture-strategy credit assignment problem describes the difficulty for most cognitive modeling research to identify the source of a good fit between model and human performance. A good fit could arise either due to an appropriate choice of constraints or due to an appropriate choice for the strategy (i.e., the sequence of production rules), given the constraints. The focus of most cognitive architecture research tends to be on the role of constraints, and not on the role of alternative strategies (although strategies are an important component of EPIC, Meyer & Kieras, 1997a, 1997b). In contrast to other approaches, the cognitively bounded rational analysis approach considers both explicitly.

To summarize, features of the models that I develop are that (1) no detailed assumptions are made that go beyond the level of detail that can be measured in an experiment (e.g., there are no assumptions about production rules), (2) constraints are defined explicitly, and (3) a consideration is given to the performance of alternative strategies. As a result of these features, the models in this thesis can make contributions to the literature, while at the same time staying away from having to make strong claims about highly debated aspects of multitasking performance, such as the role of cognitive control (for relevant discussions see Cooper, 2010; Kieras et al., 2000). This is a feature, as a relatively small set of explicit constraints together can explain relatively complex behavior. This enables testing, including falsification, of the theory and the underlying constraints through a series of critical experiments.

2.4. Conclusion

In this literature review I have set out the domain, research question, and the approach that I will take in the remainder of this thesis. It is hypothesized that multitasking can be understood as cognitively bounded rational behavior. Performance is subject to constraints that come from the task environment, invariant cognitive aspects, flexible cognitive aspects (e.g., skill differences), and an objective. In Chapter 3 I will investigate how task, invariant cognitive aspects, and objectives influence dual-task performance in a dialing-while-steering set-up. In Chapter 4 I will in addition look at the role of flexible cognitive aspects (individual differences in skill) and use a payoff function instead of a priority objective. The context in that Chapter will be a tracking-while-typing task. In Chapter 5 and 6 I will reflect on the suitability of payoff functions for studying task interleaving through mathematical modeling.

Chapter 3. When do People Interleave Tasks in a Dialing-while-Steering Setting? On the Role of Natural Breakpoints, Priorities, and Cognitive and Motor Cues.[5]

Abstract

In this Chapter I report a series of experiments and cognitive models that are used to investigate when people interleave their attention between manually dialing a phone number and steering of a simulated vehicle. Preceding work in a similar setting (Salvucci, 2005) has argued that people tend to interleave at cognitive chunk boundaries of the phone number, as these form "natural breakpoints" in the task. In experiment 3A, I investigate how this interleaving pattern changes with a change in the frequency of chunk boundaries. In experiment 3B, I demonstrate how the frequency of interleaving is dependent on the priority of the user. Finally, in experiment 3C, I demonstrate that motor cues can also act as "natural breakpoints" for interleaving tasks. Taken together, the results support the idea that people can strategically control the allocation of attention in multitask settings to meet specific performance criteria. Their performance is not solely

[5] This Chapter combines work from two published papers with one other experiment (3A). Throughout the Chapter text, analyses, and figures have been taken from these papers:

- Janssen, C. P., & Brumby, D. P. (2010). Strategic adaptation to performance objectives in a dual-task setting. *Cognitive Science, 34*, 1548-1560.
- Janssen, C. P., Brumby, D. P., Garnett, R. (2012). Natural Break Points: The Influence of Priorities, and Cognitive and Motor Cues on Dual-Task Interleaving. *Journal of Cognitive Engineering and Decision Making, 6, 5-29*.

Rae Garnett collected the data of experiment 3C using the experiment developed by Christian Janssen. Christian Janssen conducted the analyses.

guided by the position of the chunk boundaries, or more generally, by the task structure. Four contributions to the literature are made. First, I perform a series of critical tests on the theory that people interleave attention if and only if they reach a chunk boundary, which demonstrates the strategic control that people exert over their actions. Second, I demonstrate an advantage of interleaving at natural breakpoints: this offers an important speed-accuracy trade-off. Third, I demonstrate that motor cues can also act as natural breakpoints. Fourth, the studies provide a critical reflection on "rationality" in distracted driving. Limitations and further implications are discussed.

3.1. Introduction

3.1.1. "Natural breakpoints" in driver distraction settings

When do people interleave two tasks when multitasking? Many preceding studies have shown that interleaving at subtask boundaries is especially useful (e.g., Bailey & Iqbal, 2008; Bogunovich & Salvucci, 2010; Miyata & Norman, 1986; Payne, Duggan, & Neth, 2007; Salvucci & Bogunovich, 2010). Interleaving here, rather than at other positions in the task, offers at least three advantages. First, when completing a subtask there is a reduction in mental workload (Bailey & Iqbal, 2008; see also, Salvucci & Bogunovich, 2010). This is because during the execution of a subtask localized state information is often required (Borst, Taatgen, & Van Rijn, 2010) and suspending the task would require that information to be held in memory (Altmann & Trafton, 2002). On completing a subtask, localized state information is no longer needed; hence there is no need to actively maintain it in memory while the other task is attended to. A second advantage of interleaving at subtask boundaries is that task resumption lags following an interruption are shorter when a task is suspended at the completion of a subtask (Altmann & Trafton, 2002). Finally, as the number of resources that are needed for the primary task are reduced at the subtask boundary, these resources become available for other tasks, as predicted by multiple resource theories of cognition (e.g., Salvucci & Taatgen, 2008, 2011; Wickens, 2002, 2008). As such, subtask boundaries form "natural breakpoints" for interleaving tasks. But do people always interleave attention if and only if they reach such a natural breakpoint? That will be investigated in this Chapter.

The particular context in which this will be studied is driver distraction. The effect of subtask boundaries on interleaving in distracted driving has been investigated in a series of studies (Brumby, Howes et al., 2007 ; Brumby, Salvucci et al., 2007; Brumby et al., 2009; Salvucci, 2005). In these studies, people had to steer a simulated vehicle, while also typing in a previously rehearsed phone number. The phone numbers that were used in these studies all had a typical Northern American structure, with digits clustered in groups of three and four digits, as in for example: 123 – 456 – 7890. The studies found that people typically interleaved dialing for steering in between these groups, or chunks, of digits.

Following more general theories of the hierarchical representation of memory (Anderson, Bothell, Lebiere, & Matessa, 1998), it is assumed that this pattern reflects the hierarchical representation of the phone number in memory (e.g., Salvucci, 2005). One reason for why people might interleave at these points is that after dialing a chunk of digits the next chunk needs to be retrieved from memory. As this retrieval takes time, the hands and eyes are not needed for dialing, and can therefore be used for driving (Salvucci & Taatgen, 2008, 2011; Wickens, 2002, 2008). In that sense, chunk boundaries can be considered a subtask boundary, and a "natural breakpoint" in multitasking scenarios.

A study by Brumby, Salvucci, and Howes (2009) found that whether people interleaved at these natural breakpoints depended on the priority that drivers had. In a simulated driving study, it was found that if participants' priority was to dial the number as fast as possible, they tended not to interleave at the natural breakpoints of the phone number. However, when their priority was to drive safely, participants would interleave at the natural breakpoints.

3.1.2. The objective of this Chapter

The studies in this Chapter are set out to investigate the generality of these findings. A shortcoming of the above studies is that only Northern American phone numbers were used, with regular chunk boundaries. It might be that people interleaved at these points because they offered an appropriate frequency for interleaving – every three digits. Study 3A is therefore set out to investigate how the frequency of chunk boundaries within a phone number influences the interleaving pattern. Two numbers were used. One conformed to a typical French structure, with chunk boundaries occurring roughly every two digits (i.e, xx – x – xx – xx – xx - xx). The other number conformed to a typical British (or UK) structure, with only one chunk boundary (i.e., xxxxx – xxxxxx). Given these two numbers, will people only interleave at the chunk boundaries? If so, then driving performance in the French condition should be safer, while dialing performance should be slower, due to more frequent interleaving that is facilitated by the greater number of natural chunk boundaries in the number.

In study 3B and 3C, further investigation will be made in situations where a UK-style phone number is used. In experiment 3B, I investigate the role that priorities play on interleaving decisions. Brumby and colleagues (Brumby et al., 2009) found that the driver's priority strongly influenced if and when driving was interleaved for dialing a Northern American phone number. Is this also the case when the phone number has few chunk boundaries, such as in the UK style number?

Finally, in study 3C, it is investigated whether motor cues can also act as "natural breakpoints". Motor cues will be introduced by having participants dial phone numbers that contain sets of repeating digits. Do people still interleave only at the chunk boundary, if a series of repeating digits crosses the chunk boundaries? For example, when dialing 07722 - 229944, the continuing series of 2s surrounding the chunk boundary might encourage one to prolong typing while the finger is on top of the 2. Is this the case? And how is this influenced by the priority of the driver?

In all studies roughly the same experimental set-up will be used. Experimental data will be complemented with a modeling analysis to investigate what steering and dialing performance might look like when people applied different strategies for interleaving the two tasks. This is then used to investigate whether participants' strategies incorporated useful speed-accuracy trade-offs.

The remainder of this Chapter is structured as follows. First, I will give a general description of the experimental set-up and the general structure of the cognitive model. This is followed by reports on the three experiments and their modeling counterparts. In the general discussion, the findings will be summarized and the contribution of this Chapter to the literature will be discussed. Implications, limitations, and possible extensions of the work will also be discussed.

3.2. General experiment structure and analysis

All three experiments were conducted with the same driving simulator and mobile phone. The general structure of the experiments is described here.

Participants

All participants were recruited using the UCL subject pool, and participated for monetary compensation. All participants had their driver's license for a minimum of two years. No other restrictions were applied.

Materials

Participants had to perform two tasks: (1) steer a simulated vehicle, and (2) manually type in a previously rehearsed phone number. Figure 3.1 shows a picture of the set-up for the dual-task scenario. The steering task required participants to navigate the center lane of a straight, three-lane highway environment. The simulation environment was displayed on a 30-inch monitor and controlled using a Logitech G25 steering wheel. Participants were required to steer the vehicle to maintain a central lane position. The view in the simulator was from a first-person perspective - as if looking through the windscreen of their

Figure 3.1: The set-up of the driving simulator.

car. The vehicle drove at a constant speed of 88.5 km/h behind a lead vehicle at a fixed distance of 30 meters. No acceleration or braking was required. To encourage safe lane keeping, safety cones were placed at either side of the driver's central lane. Noise was added to the vehicle dynamics, causing it to gradually drift in the lane. This meant that the participant had to actively control the vehicle's lateral position and heading to maintain a central lane position. The simulator logged both the vehicle's lateral deviation from the center of the lane and the angle of the steering wheel at a rate of 200 Hz. The simulator software was developed by Dario Salvucci, and first used in Salvucci and Beltowska (2008). I adapted it for the present purposes.

For the dialing task, participants used a Nokia 6300 mobile phone, which was secured in a hands-free cradle mounted on the desk to the left of the steering wheel (see Figure 3.1). Using the phone, participants had to dial previously rehearsed 11-digit phone numbers (see procedure). To start and finish dialing, participants had to press '#' on the phone. Any errors that were made needed to be corrected. Pressing the '*' button deleted the last digit from the sequence of digits dialed. Therefore, mistakes made earlier required multiple corrections. This time consuming action encouraged accurate performance. The experimental software logged all keypress data from the phone (through Bluetooth communication) and integrated it with data from the driving simulator.

Procedure

Participants always performed four types of tasks: (1) single-task steering, (2) phone number practice, (3) single-task dialing, and (4) dual-task steering + dialing trials. Details of each task are discussed below.

Single-task steering trials In the single-task steering trials (30 seconds each) the objective was to keep the car as close to lane center as possible. Feedback on mean absolute lateral deviation (i.e., drift from lane center) was given after each trial. Every fifth trial, average performance was reported. The single-task steering trials served as a way for the participant to get used to the driving simulator, as well as a way to measure base-line driving performance.

Phone number practice trials Within a single experiment, all participants had to dial the same number(s). Different numbers were used across experiments. To ensure that each number was memorized in accordance with the intended chunk structure, the learning process was controlled. Two slightly different training regimes were used. Both regimes emphasized 'learning by doing': the number was learned by systematically typing it in on the phone.

In the first training regime, participants got the to-be-learned number presented on the simulator screen with all digits covered by an X, except for those from the current chunk (e.g., two examples are xx-x-15-xx-xx-xx and 07722-xxxxxx). Subjects had to dial this chunk on the phone. As soon as a chunk was dialed, it was covered up and the next chunk was revealed and had to be dialed (e.g., xx-x-xx-73-xx-xx or xxxxx-229944). This training was done for 10 trials per number. This regime was used in experiment 3A and 3C.

The second training regime is an extension of the first. Here, participants practiced dialing each chunk of digits separately. For each chunk of digits they received a note that had this chunk written on it. Participants were encouraged to learn this sequence of digits by dialing the digits on the phone for three minutes. Participants were encouraged to use their left hand for this practice, as later in the experiment they would also use their left hand for typing the number. After the first chunk of the number was trained in this way, the note was taken away and a second chunk of digits was trained in a similar way. Participants were then told that the two chunks together formed a phone number, and they could practice the combination for 10 trials using the procedure with covered-up digits as explained for the first training regime above. This regime was used in experiment 3B.

Single-task dialing trials In the single-task dialing trials, participants had to dial the previously rehearsed number from memory without any cues. Each time that a number was typed in correctly, participants received feedback on their dialing time (in seconds). Every fifth trial, feedback on average performance was given.

Dual-task steering + dialing trials In the dual-task trials, participants had to dial the phone number from memory, while also steering the simulated vehicle. A dual-task trial finished once the number was completely keyed in, or after 60 seconds, whichever came first. Per trial, the phone number was dialed once.

In experiments 3B and 3C participants received an explicit priority instruction before each block of trials, to prioritize either safe driving (steering-focus) or rapid completion of the dialing task (dialing-focus). To emphasize these instructions, participants received specific feedback in each priority condition at the end of every trial, and received average performance feedback after every fifth trial. In the steering-focus condition, feedback was given on the vehicle's root mean squared error (RMSE) lateral deviation. In the dialing-focus condition, feedback was given on the time taken to dial a number (the time between the two presses of the '#'). Priority was manipulated as a within-subjects variable. Where needed, the experimenter reminded the subjects about their priority in between trials.

In experiment 3A, participants did not receive an explicit priority instruction. They were instructed to drive as safely as possible while also dialing the phone number. They received feedback on RMSE lateral deviation at the end of every trial, and feedback on average performance after every fifth trial.

In cases where multiple phone numbers were used (experiment 3A and 3C), all trials that used the same phone number were clustered together. All experiments were run as within-subjects experiments. The within-subjects factors were phone number (experiment 3A and 3C), and priority (experiments 3B and 3C).

Analysis

Data was excluded from participants whose dialing time or driving performance was greater than two standard deviations from the overall participant mean. Trials on which dialing errors were made were also excluded. For trials that involved dialing (single- and dual-task), performance was investigated after the first '#' key was pressed up until the last digit of the phone number was keyed in. That is, only performance during dialing was investigated.

Two measures for overall performance were explored: average total dialing time, and mean absolute lateral deviation (i.e., average drift from lane center). The measure of absolute lateral deviation was used in two ways. First, the average lateral deviation while dialing was calculated as a measure of overall performance. Second, the lateral deviation position on each keypress was inspected to allow investigation of how lateral deviation changed as dialing progressed.

Two measures of fine-grained performance were explored: interkeypress interval time, and total number of steer counts between keypresses. Interkeypress interval time was measured as the time between a previous keypress and the current keypress. For example, interkeypress interval for digit 5 was the time interval between the fourth and fifth keypress.

Interkeypress interval was used to explore at which moments in time significant delays in dialing occurred. It was assumed that a relative large interkeypress interval in single-task trials marked a chunk boundary, and that a further increase of the interval in dual-task trials marked interleaving of attention (similar to, Salvucci, 2005).

Total number of steer counts was used to measure whether relatively more corrective steering movements were made at certain interkeypress intervals compared to others. A steering movement was defined whenever there was a change in the angle of the steering wheel after at least three consecutive samples (at a frequency of 50 Hz) at which the steering wheel was held at a constant angle. Varying this threshold will clearly affect the absolute number of steering movements identified in any given sample. For example, a lower threshold will identify small, random, movements of the steering wheel also as steering movements. Therefore, caution should be taken when interpreting these absolute counts. However, what is interesting, is the *relative* difference in these steer counts at different points in the number. In general, I assume that a high steering count at an interkeypress interval (compared to steer counts at other intervals) represents a period in which the participant made many adjustments to the angle of the steering wheel. This is taken as a good indicator that the participant was actively attending to the driving task.

For most analyses, Analyses of Variance (ANOVAs) were applied. For all analyses a significance level of .05 was used.

3.3. General description of the model

For each experiment, human performance was compared to predicted performance of a cognitive model. The modeling approach is an extension of the modeling work by Brumby, Salvucci, and Howes (Brumby, Howes et al., 2007; Brumby, Salvucci et al., 2007; Brumby et al., 2009). Their framework is based on the cognitively bounded rational analysis framework (Howes et al., 2009). The model is used to derive performance predictions for a variety of ways in which the dual-task can be completed. The model represents basic task operators (e.g., dialing a single digit, or performing a steering control update) as discrete processing units that are limited by a serial bottleneck. This means that the dialing task interferes with steering control processes.

The model framework contained a steering control model and a model of dialing. A central processing bottleneck was assumed which meant that only a single task could be attended to at a time. Given this constraint on performance, the modeling analysis focused on systematically exploring the space of possible dual-task interleaving strategies. Between the models only one free parameter was varied: the time needed to type in a digit. These times were fit to measured single-task data for each experiment.

Steering model

As in the experiment, the steering model simulated a car moving at a constant speed down a straight highway. Whenever the model was not actively steering (e.g., dialing the phone number or switching between tasks, see "Dialing model"), the current lateral velocity of the vehicle was altered every 50 msec by adding a new value sampled from a Gaussian distribution with a mean of 0.00 m/s and a standard deviation of 0.13 m/s. This represented the way a car might drift about in the lane when a driver has taken their eyes off the road.

Active steering was modeled using a simple control function that changed the heading (or lateral velocity in m/s) of the vehicle based on its position in the lane (LD, measured as absolute deviation from lane center, expressed in meters), shown in Equation 1:

Lateral Velocity = 0.2617 x LD^2 + 0.0233 x LD – 0.022

(Equation 3.1)

The equation was developed by Brumby, Salvucci, and Howes (2007) based on an analysis of previously collected lateral deviation data (Salvucci, 2001; Salvucci & Macuga, 2002). It captured the basic idea that a driver will make sharper corrective steering movements as the car gets farther from the center of the lane. Whereas, when the car is very close to the center of the lane only very minor corrections to heading are required.

In normal (uninterrupted) driving conditions, the model adjusted the heading, or lateral velocity, of the vehicle once every 250 msec (based on Equation 3.1). This timing estimate approximates the time needed for updating a lateral position, and is similar to timing predictions of other discrete driving models (e.g., Salvucci, 2001; Salvucci & Macuga, 2002). As the steering wheel in the simulator had a maximum angle, lateral velocity was limited to a maximum value of 1.7 m/s. To reflect variability in human performance (as observed in the analysis by Brumby, Salvucci et al., 2007), a random value was added to the lateral velocity on every 50 msec update. The value came from a Gaussian distribution with a mean of 0.00 m/s and a standard deviation of 0.10 m/s.

Dialing model

The dialing task was modeled at the level of how long it took to press a single key. This time was approximated using measured data from single-task dialing trials. This free parameter varied between models, as the phone number and its characteristics varied.

It was assumed that at the chunk boundaries the next set of digits needed to be retrieved, which increased dialing time with 100 msec for the first digit of that chunk (Brumby et al., 2009). In a UK-style number (experiments 3A-C) there are chunk boundaries at positions 1 and 6 (e.g., xxxxx - xxxxxx). In a French-style number (experiment 3B), there are chunk boundaries at positions 1, 3, 4, 6, 8, and 10 (e.g., xx - x - xx - xx - xx - xx).

Dual-task model

To model dual-task behavior it was assumed that at any moment in time the model was either dialing, steering, or switching attention between the two tasks. When the model switched between tasks there was a time cost of 200 msec. This cost captured the time needed to switch attention between the phone and the road. If the model interleaved at other positions than the chunk boundary, it was assumed that the position within the chunk of digits at which to resume needed to be retrieved. This additional switch cost was set to 100 msec (Brumby et al., 2009). Figure 3.2 gives a schematic representation of the states that the model cycles through: dialing, switching attention, driving, switching attention, experiencing optional costs, and dialing again.

For each model run, the model started at 0.45 meter away from lane center.[6] This value was fitted to the human data. The measured

---

[6] For model 3C this differs from the value used in Janssen, Brumby, and Garnett (2012). There, the value was set similar to values in Brumby et al. (2009) for consistency. Here, the value was revised to reduce the number of free parameters between the models in this thesis, and to have a better correspondence with human

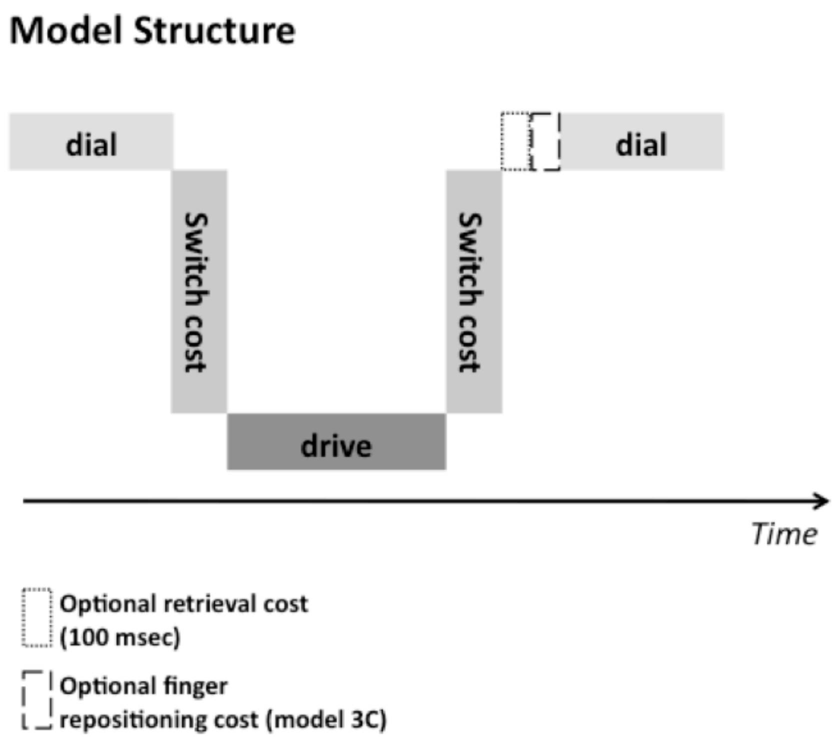

Figure 3.2: Steps that the model iterates through: dialing, switching attention, driving, switching, and then a series of optional costs (see text), before continuing to dial.

values in experiments 3A, 3B, and 3C were: 0.450, 0.447, and 0.431 meter. This measure was the mean value at which the car was positioned when participants pressed the first '#' key.

Exploring the space of dual-task interleaving strategies

With the model, an exploration was made of what performance could look like given the frequency and duration of task interleaving. This was done by systematically varying between model runs the number of digits dialed in one sequence (a strategy), and the time spent on driving

---

data. Although the exact performance of strategy alternatives is different as a result, the main take-away points in that paper are in line with the findings reported here.

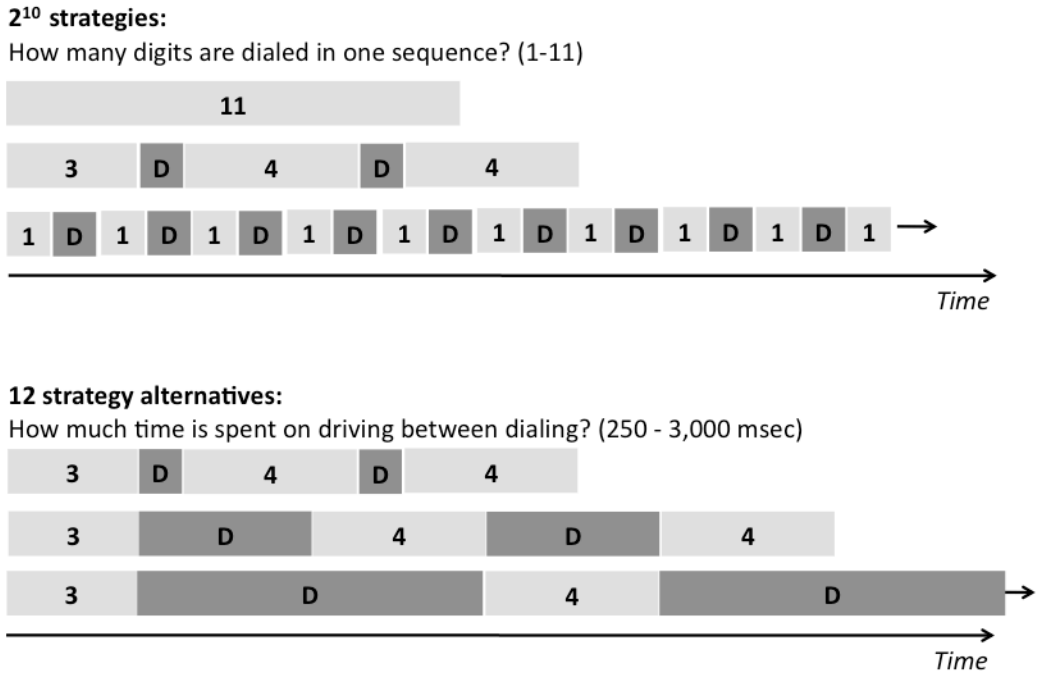

Figure 3.3: Example of the manipulation of strategy
(top: how many digits are dialed before driving?) and strategy alternatives (bottom: how much time is spent on driving?).

in between dialing (a strategy alternative). Six examples are schematically drawn in Figure 3.3. Each phone number had 11 digits. It was assumed that after each digit that was dialed by the model, the model could either dial another digit or attend to the road. This gave a total of 1,024 (2^{10}) distinct dual-task interleaving strategies.

A further element of strategic variability was in terms of how much time the model gave up to steering control before returning to dialing. This time was systematically varied with steps of 250 milliseconds, between 250 and 3,000 milliseconds (12 alternatives) for model 3A and 3C, and between 250 and 5,000 milliseconds (20 alternatives) for model 3B. These steering control alternatives were varied between simulations, but kept constant within a single run to make the space of strategies tractable. In model 3A and 3C this resulted in a total of 12,277 *strategy alternatives*, out of a full set of 13^{10} strategy alternatives for each number to be evaluated. In model 3B this gave 20,461 *strategy alternatives* out of a full set of 21^{10} strategy alternatives.

To get a relatively stable estimate of mean performance, 50 (model 3A and 3C) to 200 (model 3B) simulations were run for each strategy alternative. Mean performance of each strategy alternative is reported for each phone number.

General model result analysis

The predicted performance by the model is expressed in terms of the total time needed for dialing and the average absolute lateral deviation of the simulated car (i.e., how far the car deviated from lane center). The aim of this analysis is to get a sense of the tradeoffs that are involved in the various strategies that participants could have adopted. This can then help to better understand why participants might have applied specific strategies for interleaving.

For all models, plots such as the one in Figure 3.4 are presented. In this "performance space" the average performance of different strategy alternatives (grey dots) is expressed in terms of expected total dialing time (horizontal axis) and mean lateral deviation (vertical axis). In between the two extreme strategies (highlighted with arrows) of no interleaving and max interleaving, there are many strategy alternatives. Of particular interest are the strategy alternatives that are highlighted with the black dots. These are the strategy alternatives that are on the pareto frontier. For these points, the model achieves optimal performance on a secondary task (e.g., dialing), given a specific performance criterion on a primary task (e.g., driving). If horizontal and

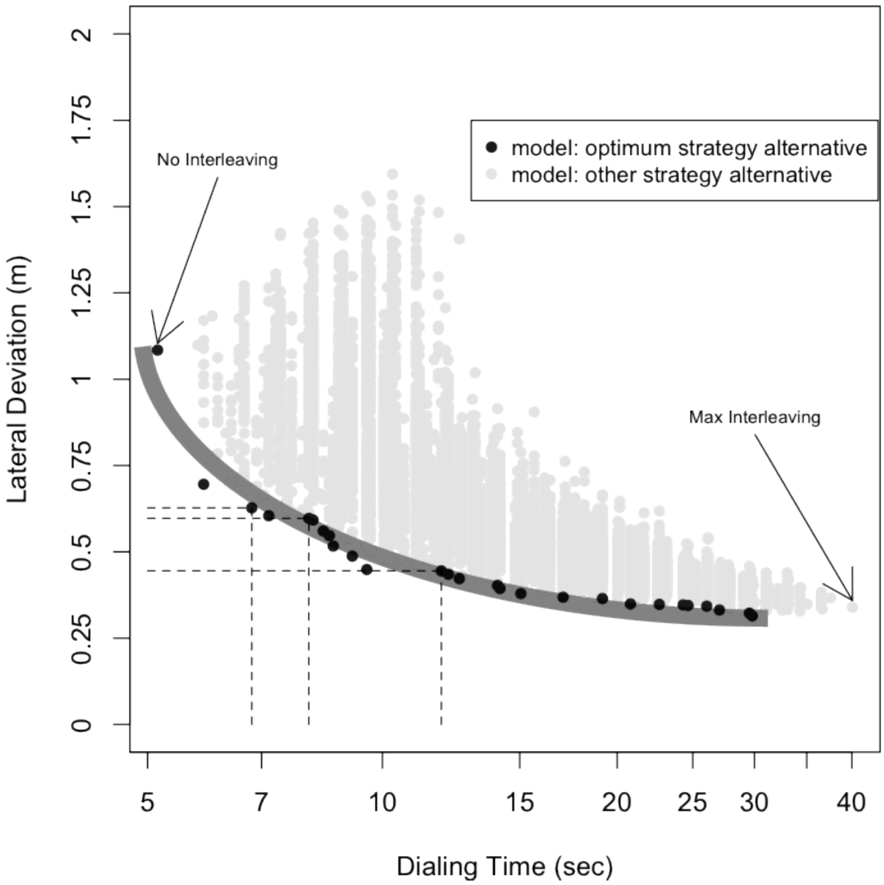

Figure 3.4: An example of a model performance space.
Grey dots are model predictions for strategy alternatives expressed in average lateral deviation (vertical) and total dialing time (horizontal). The black dots are those that are on the pareto frontier. An optimal trade-off curve (black line) can be fitted through these points.

vertical lines are drawn between the points on the pareto frontier and the two axes, as is done with dashed lines for three points in Figure 3.4, then no other points will be encapsulated within the rectangular area defined between these lines for each point on the pareto frontier. For more mathematical details on the pareto frontier, see for example Messac, Ismail-Yahaya, and Mattson (2003).

A curve can be fitted that roughly covers the outside edge of the predicted performance space, and the pareto frontier points with it.

This is illustrated with the dark grey line in Figure 3.4. I will refer to this curve as the *optimal (performance) trade-off curve*, as the points on this curve are on, or close to, the pareto frontier, and therefore make an optimal trade-off between performance on the steering task and performance on the dialing task. This is similar to predictions that experimental studies can give using performance operating characteristics (Navon & Gopher, 1979; Norman & Bobrow, 1975), although here the results are acquired through model predictions rather than human experiments. In a series of experiments, I will test whether human performance is positioned on this curve, and if so, where it is positioned. If performance lies on top of the curve, then participants made an optimal performance trade-off.

3.4. Experiment 3A: Frequency of chunk boundaries

The first experiment was set out to explore what the effect was of the frequency of chunk boundaries in a phone number on dual-task performance and interleaving pattern. If people interleave dialing for steering if and only if they reach a chunk boundary (cf. Salvucci, 2005), then a change in the frequency of chunk boundaries within a number might significantly alter performance. The more frequent the chunk boundaries occur, the better steering performance might be as the steering task is attended to more frequently. This might be at the cost of dialing time. Alternatively, it might be that participants strategically adapt performance to the task at hand, and change the interleaving pattern in both numbers in such a way that performance reaches a specific criterion level for driving performance (e.g., Brumby et al., 2009).

3.4.1. Method

Participants

Thirty participants (eight female) with a mean age of 24.0 years (SD = 4.5 years) participated.

Materials

Two phone numbers were used. One number had one chunk boundary, after the fifth digit (xxxxx-xxxxxx). This will be referred to as the UK number, given the similarities with a typical British telephone number. Another number had five chunk boundaries, roughly after every second digit (xx – x – xx – xx – xx – xx). This is referred to as the French number, given similarities with a typical French telephone number. Each number contained all digits zero to nine, with the constraint that two consecutive digits were not in the same row (i.e., “123”) or column (i.e., “147”) of the phone’s keyboard. This resulted in 01953-842761 (UK number) and 29-0-15-73-48-62 (French number).

Design

A single factor within-subjects design was used, which manipulated the frequency of chunk boundaries in the phone number (French number versus UK number).

Procedure

Participants started with 10 single-task driving trials. There were then two blocks - one for each phone number. Per block, participants performed 10 training trials using training regime 1, followed by 10 single-task dialing trials. This was followed by 20 dual-task trials. Finally, they again performed 5 single-task dialing trials (post-test). The order in which participants experienced the phone numbers was counter-balanced. The total procedure took about 60 minutes.

3.4.2. Results

Whenever single-task data is reported, this is the average data of the pre-test and the post-test combined.

Outliers

From the raw data, data of two participants were excluded. For one participant their overall mean dialing time in dual-task trials (M = 30.03 sec) was more than two standard deviations from the overall mean (M = 13.31 sec, SD = 5.31 sec). For the other participant, their overall driving performance in dual task trials (M = 1.86 m) was more than two standard deviations from the overall mean (M = 0.60 m, SD = 0.28 m). Of the remaining 1,120 dual-task trials, data from 173 trials were removed as participants made a typing error in those trials (mean participant error-rate was 15%, SD = 6%).

Overall performance

If attention was interleaved if and only if the participant reached a chunk boundary in the phone number, then a difference in lateral deviation between the two number conditions would be expected. Lateral deviation in the French number condition should be less than in the UK number condition, as attention would be returned more frequently to steering in the French number condition. At the same time, this might lead to a longer dialing time (depending on the amount of time spent on driving during each steering episode).

The diamonds in Figure 3.7 display mean dialing time and average lateral deviation of the car from lane center for the UK phone number condition (left plot) and for the French phone number condition (right plot). A paired t-test was used for statistical analysis. Surprisingly, there was *no* significant difference in average lateral deviation of the car between the UK number condition (M = 0.54 m, SD = 0.16 m) and the French number condition (M = 0.57 m, SD = 0.18 m), $t(27) = -1.00$, $p = .33$. Similarly, there was no difference in dialing time between the UK number condition (M = 12.70 sec, SD = 4.26 sec) and the French number condition (M = 12.96 sec, SD = 4.71 sec), $t(27) = -0.71$, $p = .48$.

Events at interkeypress intervals

The overall performance data suggests that performance in both number conditions was quite similar. What about the moment-to-moment interleaving patterns? Did participants interleave if and only if they reached a chunk boundary? This was investigated by exploring performance during each interkeypress interval. In the appendix, mean interkeypress interval values for each interval are reported per phone number in single- and dual-task trials (see Figure 3.A.1). For the analysis here, performance was averaged over the type of interval, whether this was a chunk boundary (dark bars in Figure 3.A.1) or not (light bars in Figure 3.A.1). The results are plotted in Figure 3.5, for the UK number (left plot) and the French number (right plot). The Figure suggests that interkeypress intervals were higher at chunk boundaries compared to other positions for both numbers in single- and dual-task trials. Surprisingly, the interval durations increased not only at the chunk boundaries, but also at other positions. This was surprising, as no increase in interkeypress interval was expected if participants interleaved dialing for driving if and only if they reached a chunk

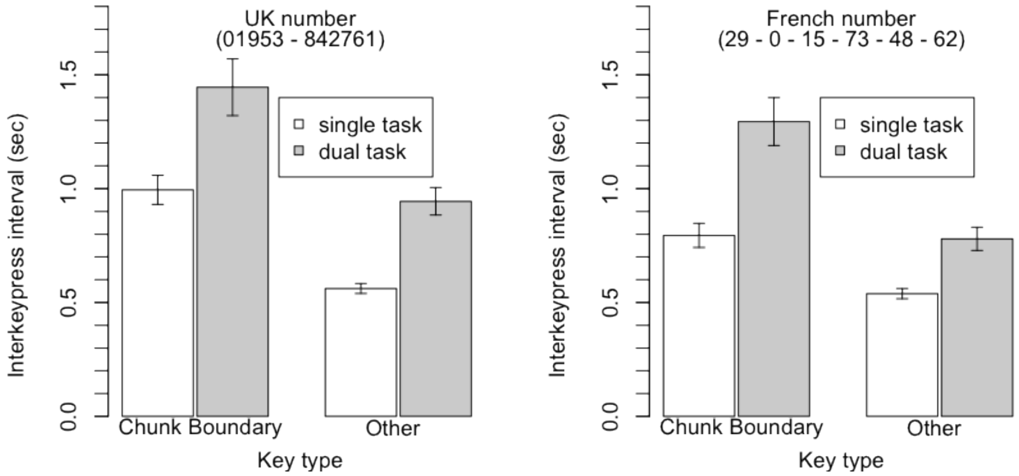

Figure 3.5: Average interkeypress interval time per keytype
in the UK number (left) and French number (right), for both single-task (white) and dual-task trials (grey). Error bars denote standard errors.

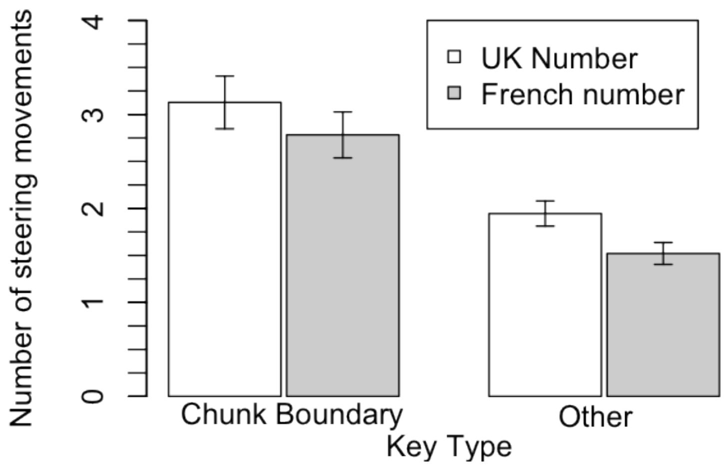

Figure 3.6: Average number of steering movements per key type
in the UK number (left) and French number (right), for both single-task (white) and dual-task trials (grey). Error bars denote standard errors.

boundary (Salvucci, 2005). In general, the interkeypress intervals for the UK number tended to be longer than those for the French number, with the exception of interkeypress intervals at non-chunk boundary positions in single-task. There were also hints of an interaction effect. The interkeypress intervals tended to be highest at the chunk boundary when the number was a UK number, particularly in dual-task trials.

A 2 x 2 x 2 ANOVA (trial type x phone number x digit type) on mean interkeypress interval time confirmed these intuitions. There was a main effect of trial type, $F(1, 27) = 42.35$, $p < .001$, such that interkeypress intervals were higher in dual-task than in single-task trials. There was also a significant main effect of phone number, $F(1, 27) = 23.50$, $p < .001$, such that interkeypress intervals were higher in the UK number compared to the French number. There was also a significant main effect of digit type, $F(1, 27) = 66.87$, $p < .001$, such that interkeypress intervals were higher at the chunk boundary compared to other positions.

There were also significant interactions between trial type and digit type, $F(1, 27) = 9.02$, $p = .006$, and between phone number and digit type, $F(1, 27) = 5.86$, $p = .02$. There was no interaction between trial type and phone number, $F(1, 27) = 1.03$. However, there was a three-way interaction between trial type, phone number and digit type, $F(1, 27) = 8.20$, $p = .008$.

Was this additional time in dual-task spent on steering, as suggested by Salvucci (2005)? For this analysis, the average number of steering movements per key type were investigated. Figure 3.6 shows these values per key type (chunk boundary, or non-chunk) in each phone number (UK and French). The Figure suggests that more steering movements were made at chunk boundaries compared to other points, and that slightly more steering movements were made in the UK number compared to the French number.

A 2 x 2 ANOVA (phone number x digit type) confirmed these results. There was a main effect of digit type, $F(1, 27) = 50.54$, $p < .001$, such that the number of steering movements was higher at the chunk boundary compared to other positions. There was also a main effect of phone number, $F(1, 27) = 19.68$, $p < .001$, such that more steering movements were also made for the UK number compared to the French number. There was no interaction between phone number and digit type, $F(1, 27) < 1$.

3.4.3. Discussion of results

Surprisingly, there were no performance differences between the two number conditions, despite a difference in the frequency of chunk boundaries. An analysis of events at the keystroke level suggested at least two reasons why this might have occurred. First, participants interleaved for a longer period at the chunk boundary in the UK phone number, to make relatively more steering movements, which could improve mean driving performance. Second, in both numbers, participants also seemed to occasionally interleave attention at non-chunk boundaries (see dual-task data in Figure 3.5). Might it be that, rather than being solely guided by chunk structure, participants were guided by speed-accuracy considerations (Brumby et al., 2009)? To investigate this a computational model was developed to explore performance of alternative strategies.

3.5. Model 3A

3.5.1. Model structure and parameters

The computational model followed the general structure outlined before. The only free parameter was the time needed per keypress, which was set at 450 msec. This was close to the first quantile of interkeypress interval times as measured in single-task trials (444 msec). It was assumed that the chunk boundaries were at positions 1 and 6 for the UK number, and at positions 1, 3, 4, 6, 8, and 10 for the French number.

For the steering tasks, the number of steering updates was varied between 250 and 3,000 msec in increments of 250 msec. This (together with the manipulation of strategy) gave a total of 12,277 strategy alternatives. For each phone number I ran 50 simulations per strategy alternative. The performance that is reported below is the average over these 50 simulations.

3.5.2. Model results

Figure 3.7 plots average dialing time (horizontal axis) and mean lateral deviation of the simulated car for a variety of strategies (grey dots) in both the UK number (left plot) and the French number (right plot). For a general description of what these performance spaces depict, see section 3.3, and compare Figure 3.7 with Figure 3.4.

The model suggests that participants did not make an optimal performance trade-off. Their performance is in the middle of the performance cloud, and not on the optimal performance trade-off curve, on the outside of the performance space (compare with Figure

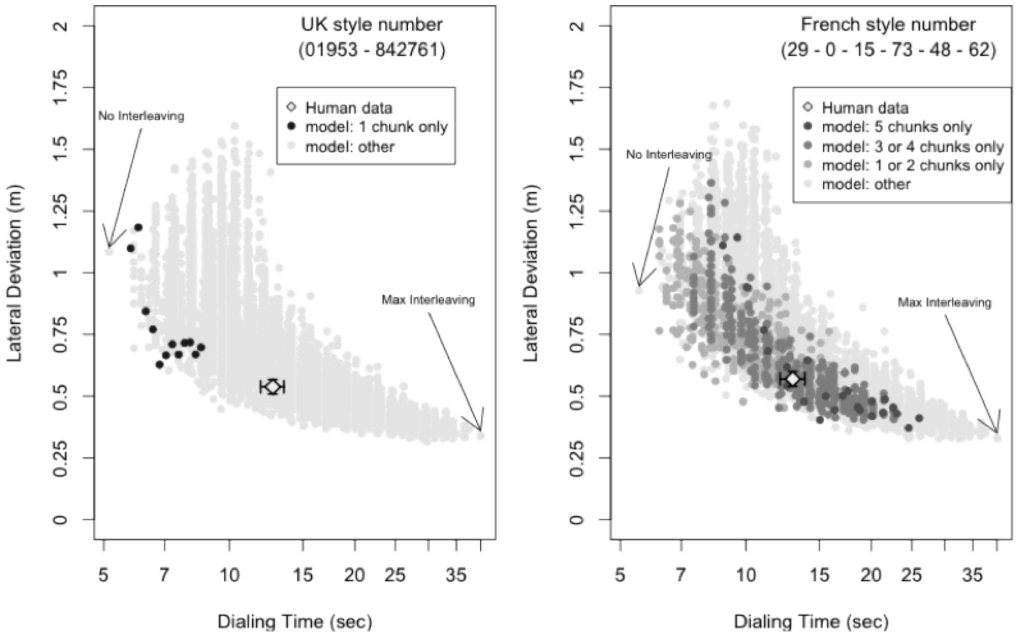

Figure 3.7: Predicted performance space for the UK style number (left) and French style number (right).
Different colored dots highlight strategy alternatives that interleave at a specific number of chunk boundaries (see legend). Diamond shows human mean performance with standard error bars.

3.4). This implies that the lateral deviation score that the participants achieved could also have been achieved using other, faster strategies. Similarly, given the achieved trial time, the participants could have achieved a better lateral deviation score.

Did participants then at least interleave at the chunk boundaries? To investigate this, I highlighted strategies that only interleaved at the chunk boundaries with darker shades of gray. As the French number had multiple chunk boundaries, different shades of gray indicate at how many of the chunk boundaries the model interleaved dialing for steering (see legend). Human performance is highlighted with a diamond.

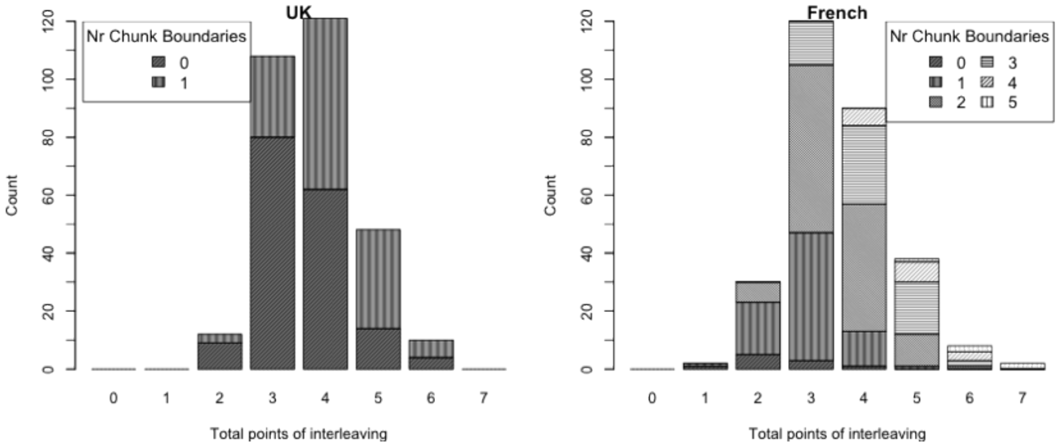

Figure 3.8: Frequency count of the number of points at which the model interleaved (horizontal axis) for strategies that fall inside the standard error bars of the human data
for the UK number (left) and the French number (right). Within each bar, a count is given of how many of the points of interleaving were chunk boundaries.

For the UK number condition, there was only one chunk boundary in the number. As Figure 3.7 shows, the model suggests that human performance was far away from the performance level of these strategy alternatives (black dots), independent of the time spent on driving while interleaving. Performance for these strategies was predicted to be at least twice as fast as the performance that the human participants achieved, and the expected deviation of the car was higher. This suggests that participants strategically decided to interleave more frequently than just at the chunk boundary to improve steering performance.

For the French number condition, it was harder to tease out whether participants interleaved if and only if they reached a chunk boundary based on Figure 3.7. The model suggests that human performance was in a region where multiple strategies could explain performance.

To further this point, I explored the points of interleaving for the model's strategy alternatives that fell inside the standard error bars of human performance. Figure 3.8 shows a histogram of the total number of points of interleaving (horizontal axis) that these strategy alternatives had, for the UK number (left; 299 strategy alternatives) and the French number (right; 290 strategy alternatives). Within each bar, the proportion of strategies that interleaved at different number of chunk boundaries are highlighted (ranging between 0 and 5). For the UK number, all strategies interleaved between two and six times while dialing the number. A little more than half of these strategy alternatives (169 out of 299) used the chunk boundary as a point for interleaving.

For the French number, all strategies interleaved between one and seven times, and in all cases at least half of the points of interleaving were at a chunk boundary. Only a very small fraction of the strategy alternatives (30 out of 290) interleaved at chunk boundaries only, and only one strategy alternative interleaved exactly at all five chunk boundaries. That is, only one strategy alternative interleaved if and only if a chunk boundary was reached, as proposed by Salvucci, (2005).

3.6. Discussion of results

This experiment was set out to investigate whether people interleave dialing for driving if and only if they reach a chunk boundary, as originally proposed by Salvucci (2005). Two phone numbers were used in which chunk boundaries occurred at different frequencies. In the

"French number", chunk boundaries occurred roughly every two digits. In the "UK number", there was only one chunk boundary in the middle of the number. If people interleaved if and only if they reached a chunk boundary in the phone number, then performance between the two conditions should differ. In the French number condition, steering performance should be better at the cost of dialing speed due to more frequent attention to the steering task that the available chunk boundaries allow.

This was not the case. Overall performance was similar in the two conditions, and experimental data suggested that people also performed some steering movements at non-chunk boundaries. A model analysis was then conducted to explore what performance looked like for a variety of strategy alternatives. For the UK phone number, the model suggested that performance could not be achieved by strategies that only interleaved at the chunk boundary – at least one other point of interleaving needed to be added (see Figure 3.8). For the French number, the model suggested that performance could be achieved by a variety of strategies, which either (1) interleaved if and only if a chunk boundary was reached, (2) interleaved at other positions, or (3) interleaved at some chunk boundaries but not all.

Taken together, these results suggest that chunk boundaries only act as *cues* for task interleaving. Participants can strategically decide to follow these cues or not. However, the model also suggests that the competing theory, that people adapt their performance and interleaving pattern to reach a performance criterion (Brumby et al., 2009) was not supported. Human performance was not on top of the predicted optimal performance trade-off curve (compare Figures 3.4

and 3.7). Rather, performance was in the middle of the performance space.

The above comparisons between human and model performance all have the underlying assumption that the cognitive model of dialing-while-steering is (approximately) correct. As the model is built around a series of assumptions, the conclusions might change with a change in assumptions. I return to this in more detail in the general discussion of this Chapter.

If one assumes that the model is correct, then the positioning of human performance in the middle of the model performance space might be explained as follows. Perhaps the priority objective of the participants was not explicit enough, or perhaps different participants had slightly different objectives. It is known that priorities can have a strong influence on dual-task performance (Brumby et al., 2011 ; Brumby et al., 2009; Gopher, 1993; Gopher et al., 1982; Horrey, Wickens, & Consalus, 2006; Iqbal et al., 2010; Levy & Pashler, 2008; Navon & Gopher, 1979; Wang et al., 2007). Specifically, when manually dialing a phone number while driving, the effect of priority interacts with the effect of chunk structure: If the priority is to dial the number as fast as possible, participants might forsake interleaving at the chunk boundary, but when the priority is to drive safely they might use these cues to improve driver safety (Brumby et al., 2009). It might be that participants had no clear priority, or different priorities, in mind when participating in the current experiment. To avoid this, two other experiments were run that had an explicit objective.

3.7. Experiment 3B: Effect of priorities

In the preceding experiment, it was found that people do not interleave dialing for steering if and only if they reach a chunk boundary. Surprisingly, human performance was not on top of a predicted optimal performance trade-off curve. Might this have been because there was no explicit priority stated? This was tested in a second experiment. Participants were asked to dial a UK-style phone number while steering a simulated vehicle. Within participants the priority of the driver was manipulated. In all dual-task trials, participants were required to perform two tasks: dialing and steering. In half of those, they were encouraged to focus mostly on safe driving. In the other half, they were encouraged to dial the number as fast as possible.

Does again performance change as a result of the user's priorities, as was put forward by Brumby and colleagues (Brumby et al., 2009)? And if so, do participants make an optimal trade-off, such that their performance lies on the optimal performance trade-off curve (see Figure 3.4)? A model will be used to investigate this.

3.7.1. Method

Participants

Fourteen participants (four female) took part in the study. They were between 19 and 28 years of age (M = 22.9 years).

Materials

All participants dialed one UK-style phone number (07854-325698).

Design

A single factor within-subjects design was used, in which participants were instructed to either focus on completing the secondary dialing task as quickly as possible (the dialing-focus condition) or to focus on keeping the vehicle as close as possible to lane center (the steering-focus condition).

Procedure

Participants started the experiment by training the phone number using training regime 2. This was followed by 10 single-task driving trials. There were then two experimental blocks. For each block participants completed: (1) five single-task dialing trials, (2) five single-task driving trails, (3) twenty dual-task trials under one of the two priority instructions (i.e., steering-focus or dialing-focus). The second block had the same structure as the first, but with a different priority instruction for dual-task trials. The order of priorities was randomized and counterbalanced across participants. The total procedure took about 60 minutes.

3.7.2. Results

Outliers

No participants were excluded. Of the 650 dual-task trials, data from 85 trials were excluded in which participants made an error on the dialing task (mean participant error-rate was 15%, *SD* = 11%).

Overall performance

Figure 3.9 shows a data plot where the elapsed time of each keypress from the start of dialing (horizontal axis) is plotted against the corresponding absolute lateral distance of the vehicle from lane center (vertical axis). In the first analysis, the focus was on total dialing time,

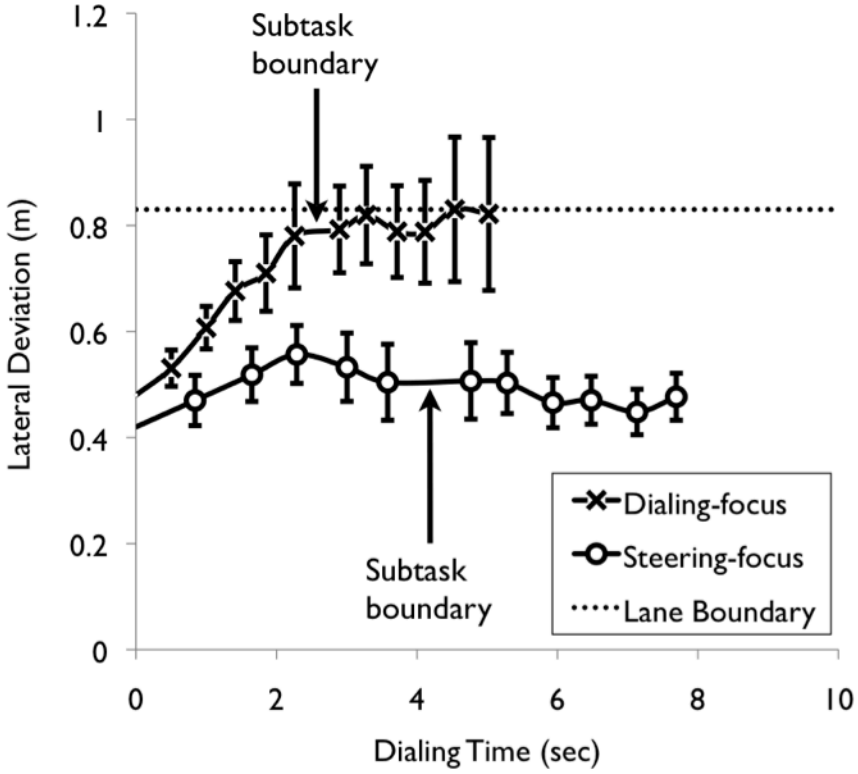

Figure 3.9: Data plot shows changes in vehicle lateral deviation between consecutive keypresses (07854-325698)
under varying dual-task performance objectives. Arrows highlight the subtask boundary for the dialing task. Error bars represent standard error of the mean.

and average lateral deviation during the dialing period. Dialing time was quicker in the dialing-focus condition (M=5.02s, SD=1.39s) than in the steering-focus condition (M=7.69s, SD=1.81s), $t(13) = 7.00$, $p < .001$, resulting in overall worse RMSE lateral deviation in the dialing-focus condition (M=0.74m, SD=0.30m) than in the steering-focus condition (M=0.50m, SD=0.18m), $t(13) = 4.20$, $p < .001$.

Events at interkeypress intervals

More interestingly, the data in Figure 3.9 suggests that when participants prioritized safer driving there was a reduction in lateral deviation after the third digit had been entered. This correction in lateral deviation occurred well before the chunk boundary in the number.

To support the observation that there was a reduction in lateral deviation before the chunk boundary in the steering-focus condition, a 2 x 11 (task priority x digit position) mixed factorial ANOVA was performance on lateral deviation data. As expected, there were significant main effects of task objective, $F(1, 13) = 17.65$, $p < .001$, and digit position, $F(10, 13) = 2.42$, $p < .05$, on lateral deviation. More importantly, there was a significant interaction between task objective and digit position, $F(10, 130) = 4.37$, $p < .001$. Follow-up tests of the simple effect of task priority showed that the divergence in lateral deviation between the two conditions occurred at the third digit position; that is, there was no significant difference in lateral deviation between the two conditions at the first and second keypress ($p = .23$ and .09, for the first and second keypress, respectively), but after the third keypress, the vehicle was significantly farther from the lane center in the dialing-focus condition than the steering-focus condition ($p < .05$, for the third through eleventh keypress).

The data in Figure 3.9 show that when participants were prioritizing safer driving the vehicle reached its maximum lateral distance from the lane center on dialing the third digit. This suggests that participants were attending to the driving task before they reached the chunk boundary. To be more confident that this was not an artifact of considering average data, an investigation was made of the frequency of trials that the maximum lateral deviation was found at a particular digit position. These data are shown in Figure 3.10. It can be seen that for both conditions there were two peaks in the distribution. As might be expected, there was a large set of trials where the vehicle was farthest from the lane center at the completion of dialing (i.e., the right peak in the Figure at digit position 11). Interestingly, there was a second peak for each condition, representing trials in which the vehicle was farthest from the lane boundary partway through the dialing task. The peak of this second distribution varied between conditions: in the dialing-focus condition, the peak was right before the chunk boundary at digit position 5, but for the steering-focus condition, the peak occurred much earlier, at digit position 3.

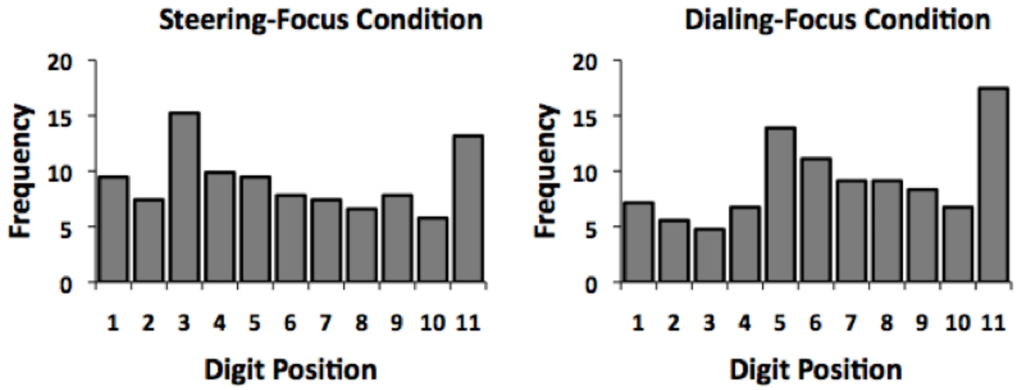

Figure 3.10: Histograms show the frequency of trials that the maximum lateral deviation was found at a particular Digit Position for each focus condition.

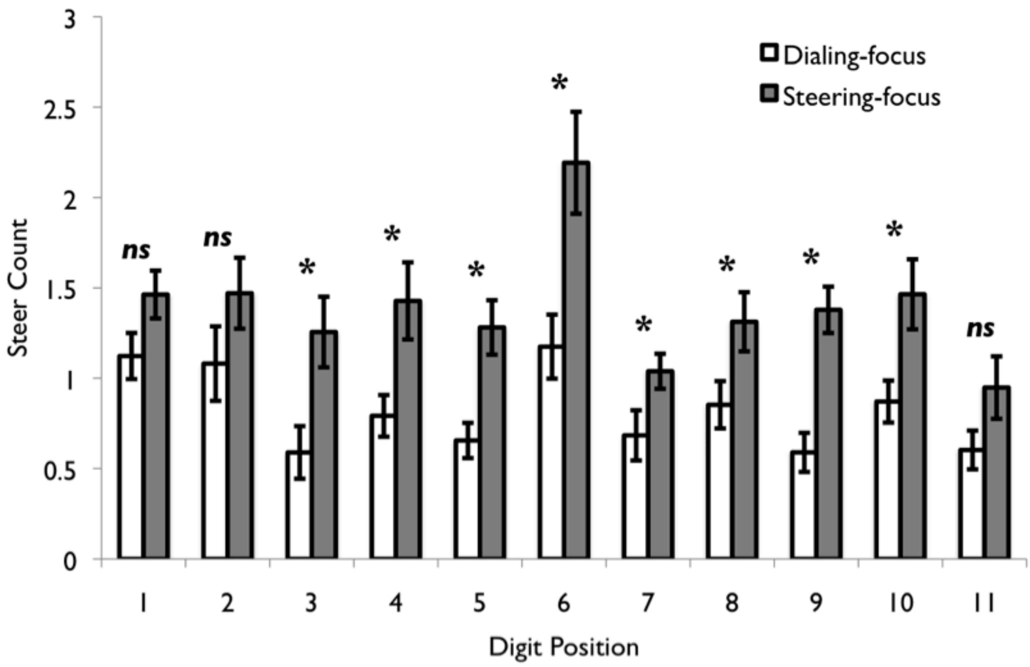

Figure 3.11: Bar chart shows the frequency of the average number of steering movement events for each condition at each digit position.
Error bars represent standard error of the mean. *denotes p < .05; ns denotes a nonsignificant difference between conditions.

To be sure that these improvements in vehicle control were due to active steering on the part of the participant, active steering movements were analyzed. Figure 3.11 shows the average frequency of steering movements for each condition at each digit position. As expected, participants steered more in the steering-focus condition (M = 15.23 events, SD = 4.48 events) than in the dialing-focus condition (M = 9.01 events, SD = 2.86 events), $F(1,13) = 33.47$, $p < .001$. To identify at which specific digit positions participants chose to steer more in the steering-focus condition than in the dialing-focus condition, the simple effect of task priority were studied at each digit position. This analysis showed that there was no significant difference in steer count frequency between the two priority conditions at the first and second digit positions (p = .07 and .15, for the first and second keypress, respectively). But consistent with the lateral deviation data, after the third keypress there were significantly more active steering movements in the dialing-focus condition than the steering-focus condition ($p < .05$ for the third through tenth keypress, with the exception of the final keypress, where there was no effect of task priority, p = .07).

3.7.3. Discussion of Results

The results of this study repeat the result of experiment 3A: participants do not interleave dialing for driving if and only if they reach a chunk boundary. New in this study was that the exact performance depended on the priority of the participant (cf., Brumby et al., 2009). In particular, when participants were instructed to prioritize steering performance, the steering and lateral deviation data showed that participants tended to actively attend to the driving task before the chunk boundary. This increase resulted in better lateral control of the vehicle, but at the cost of dialing task time. In contrast, participants in the dialing-focus conditions appeared to wait until the chunk boundary before attending to the driving task, which allowed the dialing task to be completed more rapidly but at the cost of causing the vehicle to drift farther from the lane center. However, did this adaptive performance result in an optimal trade-off? This was investigated through a modeling analysis.

3.8. Model 3B

3.8.1. Model structure and parameters

The model had the same structure as before. For this task, the time to type in a digit was set at 400 milliseconds. It was assumed that there was one chunk boundary, between digits 5 and 6.

To see if human performance could be reached if people interleaved if and only if they reached a chunk boundary, a wider set of steering update values was explored, between 250 and 5,000 milliseconds, in increments of 250 milliseconds. This resulted in a total of 20,461 strategy alternatives. To get a very reliable measure of performance, the model was run 200 times for each strategy alternative.

3.8.2. Model results

Figure 3.12 shows the predicted performance space. Human data is displayed for the dialing-focus condition (square) and steering-focus condition (diamond). Performance of the participants in both conditions lies on the optimal trade-off curve on the inner side of the

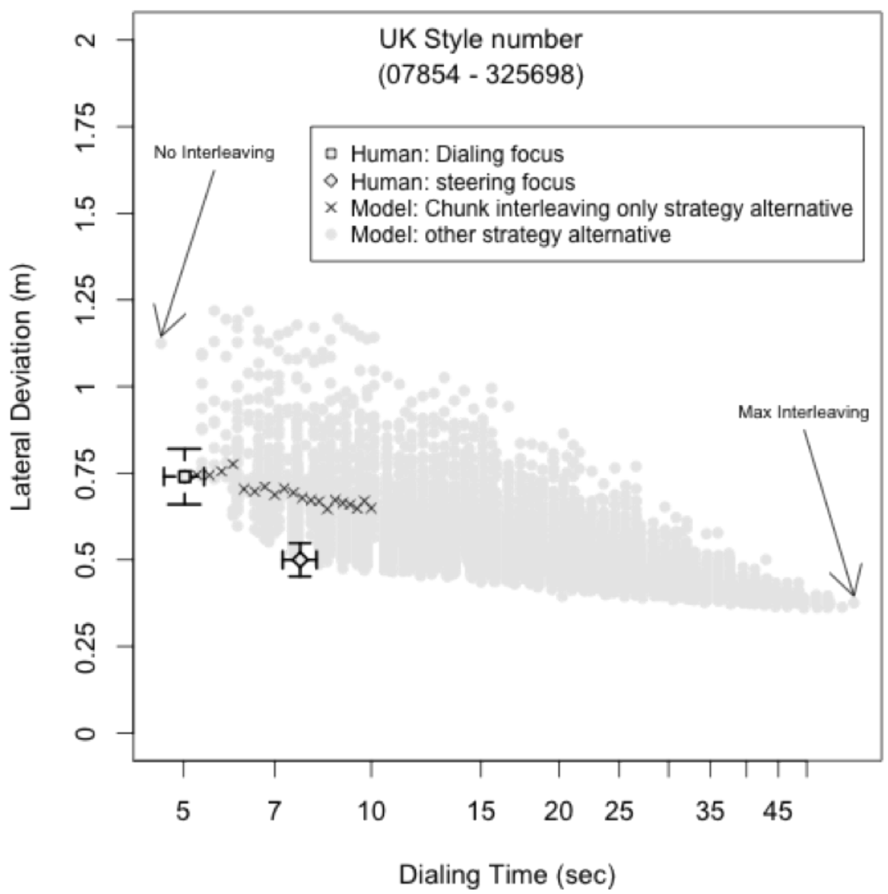

Figure 3.12: Predicted dialing time and lateral deviation for modeled strategies and human data under varying dual-task performance objectives. Error bars represent standard error.

performance space (compare with Figure 3.4). That is, given the objective that was set for one task (e.g., driving), they performed the best they could on the other task (e.g., dialing). In the dialing-focus condition, participants' performance is on the left-most side of the curve, where relatively fast dialing is achieved, at the cost of driving performance. In the steering-focus condition, performance is further down the curve where driving performance is a lot better, at the cost of dialing time. Interestingly, participants are not at the far right of the curve, where driving performance is predicted to be the best possible. Rather, they are in a region where there is (compared to preceding points on the curve) a rapid improvement in driving performance, without a significantly large cost in dialing performance. If participants would apply a strategy that was even further down the trade-off curve (i.e., more towards the point of maximum interleaving), then the relative improvement in driving performance would be small compared to current performance, whereas dialing time would increase with a factor of at least eight (note the logarithmic scale on the horizontal axis).

The model can be used to determine whether interleaving before the chunk boundary was required in order to reach the steering performance level that was observed for human participants in the steering focus condition. Specifically, it might have been possible to achieve the observed lateral deviation by returning attention to steering control only once while dialing, at the chunk boundary, and dedicating a lot of time to steering control at this point. To test this idea, Figure 3.12 highlights the performance of a "chunk-interleaving only" strategy with varying time given up to steering control (black crosses). It can be seen that even when substantial periods of time were given up to steering control (up to 5 seconds), lateral deviation never reached the tipping point in the tradeoff curve, nor, did performance fall within the standard error bars[7] of the human data from the steering-focus condition. Moreover, of the 21 strategy alternatives that fell within the

---

[7] In Janssen and Brumby (2010) we analyzed the strategy alternatives that fell inside the 95% confidence interval. More alternatives fell inside of that area, yet the same conclusion held that the performance in the steering focus condition could not be achieved by interleaving solely at the chunk boundary. Of the 586 strategy alternatives that fell inside the 95% CI of the human data, all but one strategy alternative performed at least one steering update before the chunk boundary.

standard error area of the human data, all interleaved before the chunk boundary. Taken together, this modeling analysis suggests that participants had to interleave tasks more often than that allowed by the chunk boundary in the dialing task to meet the performance objective of minimizing vehicle lateral deviation while dialing.

3.9. Discussion of results

This study had participants dial a UK-style phone number while steering a simulated vehicle. The objective of the participant was systematically influenced through a priority instruction. In one case, participants had to prioritize safe driving over fast dialing, in the other case they had to prioritize fast dialing over safe driving. Participants adapted their performance to comply with the priority instruction. Moreover, the modeling analysis suggested that participants adapted their performance in such a way that they made an optimal performance trade-off. In the case where the objective was to prioritize safe driving, such performance could not be achieved by interleaving solely at the chunk boundary. Instead, participants had to insert additional points of interleaving. This is what they did.

Overall, these results support the hypothesis put forward by Brumby and colleagues (Brumby et al., 2009) that whether people interleave at chunk boundaries or not depends on their priority objective. This is also in line with a wider set of studies that report the idea that people can strategically allocate attention in multitask settings to meet different performance objectives (e.g., Brumby et al., 2009; Gopher, 1993; Gopher et al., 1982; Horrey et al., 2006; Levy & Pashler, 2008; Navon & Gopher, 1979; Wang et al., 2007).

3.10. Experiment 3C[8]: Effect of priorities, and cognitive and motor cues

The preceding studies demonstrated how cognitive cues (chunk boundaries) and priorities influence when people interleave dialing for driving, and how this results in different performance styles. Might other factors also contribute to decisions about when to interleave? For example, might the actions that are required to perform a task influence this? If this is the case, then gathering more knowledge about what sorts of actions encourage or discourage task interleaving can be useful for the design, prototyping, and evaluation of mobile equipment.

To see how actions might influence interleaving, take the following example. Imagine you are driving to work and running late for a meeting. While driving to your destination, you send a message from your smart phone saying 'Running late'. When would you pause while entering this message to check on the road ahead? One strategy might be to interleave attention only after entering a whole word (i.e., only after typing 'Running'). This might seem to be an exceptionally reckless strategy as it takes the eyes off the road for a long period of time. A more conservative strategy might be to interleave after entering individual letters. In this case, the eyes are more frequently returned to the road, which can benefit safety. But what about that middle double N in the word 'running'? Would you check on the road? Or would you quickly key the second "n", as your finger is already on the letter? Perhaps this again depends on how quickly you feel you need to write the message – if you are running ten minutes early of your meeting you might even take the time to stop the car before texting.

[8] This experiment was designed and programmed by Christian Janssen. Rae Garnett collected the data. All analyses in this write-up were conducted by Christian Janssen.

As set out in the example above, a relevant motor cue in a dialing while driving scenario might be whether the finger has to be moved to a new key on the phone or whether a repeated keypress is made. When making a keypress, two phases can be distinguished (Rosenbaum, 1991). First, during the preparation phase, the motor movement is prepared, among other things by acquiring the position of the target that needs to be pressed. Second, during the execution phase, the physical action (i.e., the keypress) is performed. In the case of typing a repeated digit, the preparation phase is relatively short, as the finger is already on top of the to-be-pressed key. This makes it faster to type a repeating digit (e.g., "22") compared to having to relocate the finger to another digit (e.g., "29"), particularly when the digits are relatively far apart on the phone. In this final study I consider whether the act of moving the finger to a different key serves as a cue to interleave attention between tasks.

In addition, an exploration is made of how these basic motor cues interact with the two factors that were studied in the preceding studies: cognitive cues (chunk boundaries) and priorities. As an example, consider dialing a number such as 1233344444. Adhering to the standard North American convention, this number might be represented as 123-334-4444. However, these chunk boundaries conflict with the natural pattern of repeated digits in the number. As a result, it might make sense to represent the number differently in memory by adopting an alternative representational structure such as 1-2-333-44444, which more closely aligns to the natural breakpoints created by repositioning the finger between each series of repeating digits. In this way, basic motor cues have worked to override the default representational structure of the number and as a result one might expect changes in how the number is dialed. In addition, similar to the finding in experiment 3B, this pattern might also depend on the priorities of the user.

In the current study, these three effects (motor cues, cognitive cues, and priorities) were manipulated as follows. As before, participants had to dial previously rehearsed phone numbers, while steering a simulated vehicle, and while having an explicit objective. Through training, participants also learned that the phone numbers had a particular chunk structure. Two phone numbers were used. Each number contained sets of repeating digits, but the numbers differed in the positions at which the repetitions occurred. Critically, in one number the repeating digits crossed the chunk boundary, while in the other number a change in digit corresponded with the chunk boundary. A critical question addressed by this study was: Do people use the chunk boundary to guide task interleaving, even when motor cues favor a delay?

3.10.1. Method

Participants

Twelve participants (eight female) participated. The mean age was 28.8 years (*SD* = 6.0).

Materials

Two 11-digit UK-style phone numbers were used. Both contained sets of repeating digits. In the congruent number, one of the switches between groups of repeating digits was congruent with the position of a switch in chunks (i.e., one group ended after the fifth digit, another started at the sixth digit). In the incongruent number, one of the groups of repeating digits transcended the chunk boundary. The digits of different groups of repeating digits were chosen such that they were spaced far apart on the phone (e.g., “72” and “49” instead of “47” and “78”). This increased the contrast in dialing time between dialing two repeating digits versus dialing two non-repeating digits. The resulting numbers were 07333-888111 (congruent number) and 07722-229944 (incongruent number).

Design

The experiment followed a 2 x 2 (phone number x task priority) within-subjects design. The two phone numbers contrasted in the positions of the repeating digits (see materials). For the manipulation of task priority, participants were instructed to either prioritize safer driving over fast dialing (“steering focus”) or to prioritize fast dialing over safer driving (“dialing focus”).

Procedure

Participants started by performing 10 single-task steering trials. Following this, they performed two blocks of experimental trials. Each block of experimental trials had the same structure: (1) Participants learned a new phone number using training regime 1,[9] (2) they performed 5 single-task dialing trials, (3) they performed 15 dual-task steering + dialing trials with priority condition A (i.e., steering focus or dialing focus), and (4) they performed 15 dual-task steering + dialing trials with priority condition B. The structure in the second block

---

[9] Training regime 1 was used again here as it took relatively less time, which allowed exploration of performance with two numbers and two different priorities. There were also chronological reasons: this experiment was conducted shortly after experiment 3A was completed. If training regime 2 had been used instead, the effect of chunk boundary on dual-task performance that is present in this dataset, would probably have become even more pronounced.

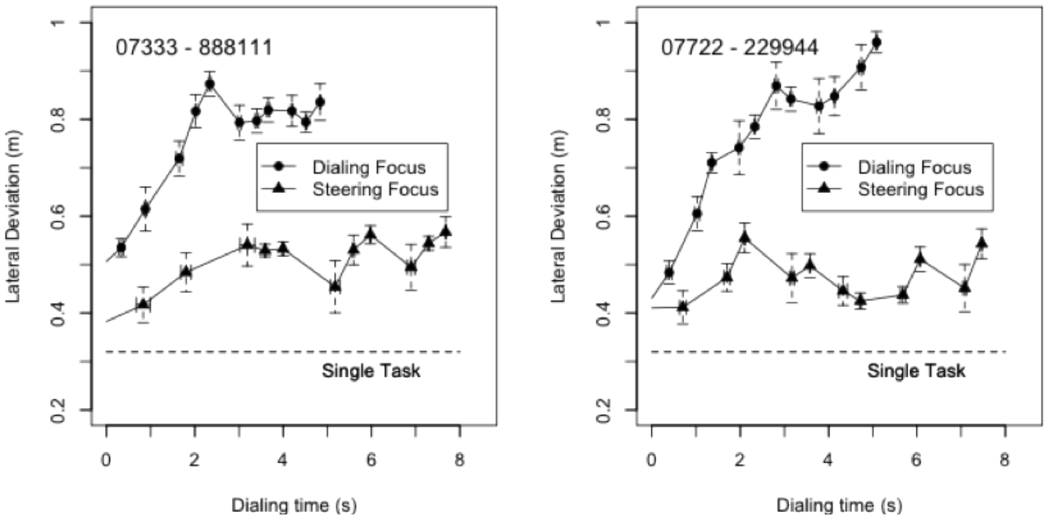

Figure 3.13: Dialing time versus lateral deviation from lane center for the congruent number (07333-888111, left) and the incongruent number (07722-229944, right)
in both the dialing focus (circles) and the steering focus condition (triangles). Error bars represent standardized error. The dashed line shows mean single-task steering performance.

mimicked that of the first block, but for a different phone number. The order of the phone numbers was randomly assigned and counterbalanced across participants. Only one phone number was used per block. The order of the two priority conditions in the first block was also randomly assigned and counterbalanced, and repeated in the second block. The experiment took about 75 minutes.

3.10.2. Results and discussion

Outliers

Data from one participant was excluded from the analysis because their mean dialing time (6.8 sec in pretest; 10.0 sec in dual-task) was greater than two standard deviations from the overall participant mean (pretest M = 4.4 sec, SD = 0.5 sec, and dual-task M = 6.3 sec, SD = 0.8 sec). Of the remaining 660 dual-task trials, data from 112 trials were excluded in which participants made an error on the dialing task (mean participant error-rate was 17%, SD = 3.4%).

Overall performance

Figure 3.13 shows a data plot where the elapsed time of each keypress from the start of dialing (horizontal axis) is plotted against the corresponding absolute lateral distance of the vehicle from lane center (vertical axis). Baseline driving performance in single-task trials is given as a dashed line. The left plot shows performance for the congruent number, and the right plot shows performance for the incongruent number. Similar to the results in the previous study, when participants were instructed to prioritize steering, they were slower at completing the dialing task, taking regular pauses in between keypresses to correct the heading of the vehicle. The net result was that the vehicle's lateral position remained relatively stable while dialing. In contrast, when participants were instructed to prioritize dialing, they were faster at completing the dialing task, taking very few pauses in between keypresses. As a result of taking fewer pauses for steering, the vehicle drifted farther from the lane center.

Table 3.1: ANOVA results for overall task performance measures (dialing time and mean absolute lateral deviation).

Source	Dial time			Average absolute lateral deviation		
	MSE	F	η_p^2	MSE	F	η_p^2
Priority (P)	2.50	30.02**	0.75	0.04	21.24**	0.68
Number (N)	1.01	0.01	0.00	0.01	0.25	0.02
P x N	0.63	0.92	0.08	0.01	1.28	0.11

** $p < .01$

The results from a repeated-measures ANOVA, shown in Table 3.1, lend support to these observations of the effect that task priority had on dual-task performance. There was a main effect of task priority on dialing time, such that dialing times were significantly faster when participants prioritized the dialing task (M = 5.0 sec, SD = 0.8 sec) compared to when they prioritized the steering task (M = 7.6 sec, SD = 1.6 sec). There was also a main effect of task priority on lane keeping performance in the steering task, such that significantly larger mean absolute lateral deviations were observed when participants prioritized the dialing task (M = 0.77 m, SD = 0.23 m) compared to when they prioritized the steering task (M = 0.49 m, SD = 0.12 m). There were no significant main effects of phone number on dialing time or on lane keeping performance, nor were there any significant interactions.

Events at interkeypress intervals

The above analyses suggest that participants were adjusting their strategy dependent on the instructions they were given on how tasks should be prioritized. Surprisingly, there was no overall effect of phone number on global task performance measures. However, what these data do not reveal is the pattern in which the phone number was dialed. Recall that the telephone numbers differed in terms of whether a change in digit was congruent or incongruent with the cognitive *chunk* boundary of the number (i.e., 07333-888111 vs. 07722-229944). For the congruent number there was a change in digit at the chunk boundary, whereas for the other, incongruent number there was no change in digit at the chunk boundary. Did this difference in cognitive and motor cues across the two numbers lead to different patterns of task interleaving?

The data in Figure 3.13 suggests that dialing was interleaved for steering at specific digit types, for example, when there was a change from one series of repeating digits to the next series (e.g., between the eighth and ninth digit in the congruent number, and between the ninth and tenth digit in the incongruent number). To investigate this, interkeypress interval data and steering movement data were analyzed for different digit types. The objective of this analysis was to determine whether there were any reliable differences in task interleaving behavior when participants dialed each of the two phone numbers. One possibility was that participants might have chosen not to utilize the cognitive chunk boundary to switch between tasks when there was no change in digit over that boundary.

Figure 3.14 contains a series of data plots that show interkeypress interval data from each of the dual-task priority conditions along with single-task baseline data for comparison. Mean data is shown for three different types of digits: at the chunk boundary (i.e., the interval between the fifth and the sixth keypress), at the first keypress of a repeated series of digits as long as this was not at a chunk boundary (e.g., “81”), and between a series of repeated keypresses of the same digit (e.g., “88”). Data are shown for both the congruent (white columns) and incongruent number (grey columns). For statistical analysis of these data, digit type (chunk boundary, first of repeating digit, or repeated digit) was incorporated as a new factor into the design, giving a 2 x 2 x 3 (phone number x task priority x digit type) repeated measures ANOVA.

The results of this ANOVA are shown in Table 3.2. It can be seen that there were significant main effects of all three variables (task priority, phone number, and digit type) on the duration of interkeypress intervals. There were also significant interactions between task priority and digit type and between phone number and digit type. Below, these results are discussed in more detail.

In general, interkeypress intervals were significantly elevated when participants prioritized the steering task (M = 820 msec, SD = 150 msec) compared to when they prioritized the dialing task (M = 520 msec, SD = 80 msec). Interkeypress intervals were also generally elevated at the chunk boundary of a number (M = 770 msec, SD = 140 msec) and when a new (non-repeating) digit in the number had to be dialed (M = 860 msec, SD = 160 msec) compared to when a digit was dialed repeatedly (M = 370 msec, SD = 30 msec). This general pattern was also seen in single-task data (see Figure 3.14), though here the pauses were far shorter than those observed in the dual-task conditions.

Table 3.2: ANOVA results for lower-level task performance measures (interkeypress interval and steer counts).

Source	Interkeypress interval			Steering events		
	MSE	F	η_p^2	MSE	F	η_p^2
Priority (P)	0.07	39.48**	0.8	0.51	29.11**	0.74
Number (N)	0.09	6.33*	0.39	0.57	6.35*	0.39
Digit type (D)	0.04	79.85^a**	0.89	0.21	60.36^a**	0.86
P x N	0.04	3.6	0.27	0.19	5.11*	0.34
P x D	0.02	28.94^a**	0.74	0.20	10.65^a**	0.52
N x D	0.04	8.23^a**	0.45	0.20	10.26^a**	0.51
P x N x D	0.02	2.13^a	0.18	0.14	1.39^a	0.12

*df*s are all (1,10), except *a* = *df*(2,20)
* $p < .05$. ** $p < .01$.

These main effects were moderated by a significant interaction between task priority and digit type. This interaction suggested that while participants took longer dialing pauses when they prioritized driving, these pauses were limited to certain points in the number, namely, the natural breakpoints afforded by the chunk boundary and by the change in digit. At chunk boundaries, interkeypress intervals were significantly elevated when participants prioritized driving (M = 970 msec, SD = 230 msec) compared to when they prioritized dialing (M = 580 msec, SD = 110 msec). When there was a change in digit, interkeypress intervals were again significantly elevated when participants prioritized driving (M = 1,080 msec, SD = 250 msec) compared to when they prioritized dialing (M = 640 msec, SD = 120 msec). However, when a digit was repeated, there was little difference between the two priority conditions (M = 400 msec, SD = 40 msec vs. M = 340 msec, SD = 40 msec). Taken together, the results suggest that participants took longer pauses in dialing when they prioritized driving, but only at the natural breakpoints afforded by the chunk boundary and by a change in digit. Relatively little (if any) time was invested in steering when the participant was in the middle of a repeating series of digits – here interkeypress interval time was about equal to the time observed in single-task.

The two telephone numbers that participants dialed differed in terms of whether or not there was a repeated digit running across the chunk boundary. For the congruent number, the chunk boundary corresponded with a change in digit. In contrast, for the incongruent number, there was no change in digit at the chunk boundary. This difference (i.e., whether or not a repeated digit ran across the chunk boundary) had an effect on whether participants chose to pause at the chunk boundary when dialing in dual-task conditions. It can be seen in Figure 3.14 that interkeypress intervals were significantly longer at the chunk boundary when participants dialed the congruent number (M = 930 msec, SD = 230 msec) compared to when they dialed the incongruent number (M = 620 msec, SD = 210 msec), $F(1, 10) = 9.39$, $p < .05$. Critically, it can be seen in the Figure that this difference in how the two numbers were dialed was not present in single-task baseline conditions; that is, the difference in interkeypress interval at the chunk boundary only emerged in dual-task conditions. For all other digit types there were no reliable differences between the congruent and incongruent number. For instance, when there was a change in digit (M = 900 msec, SD = 250 msec vs. M = 820 msec, SD = 90 msec), $F(1,10) = 1.80$, ns, or when a digit was repeated (M = 360 msec, SD = 30 msec vs. M = 370 msec, SD = 20 msec), $F(1,10) = 0.50$, ns. In other words, the only difference in how the numbers were dialed at particular digit *types* occurred at the chunk boundary in dual-task conditions: Participants paused for an extended period of time at the chunk boundary when there was a corresponding change in digit, but did not pause when there was no change in digit at the chunk boundary.

The above analysis indicates at which points in the dialing task participants were choosing to pause. To be confident that participants were using these periods to attend to steering, steering data were analyzed. It was expected that there were more steering movements at points where there were prolonged pauses in dialing (Salvucci, 2005).

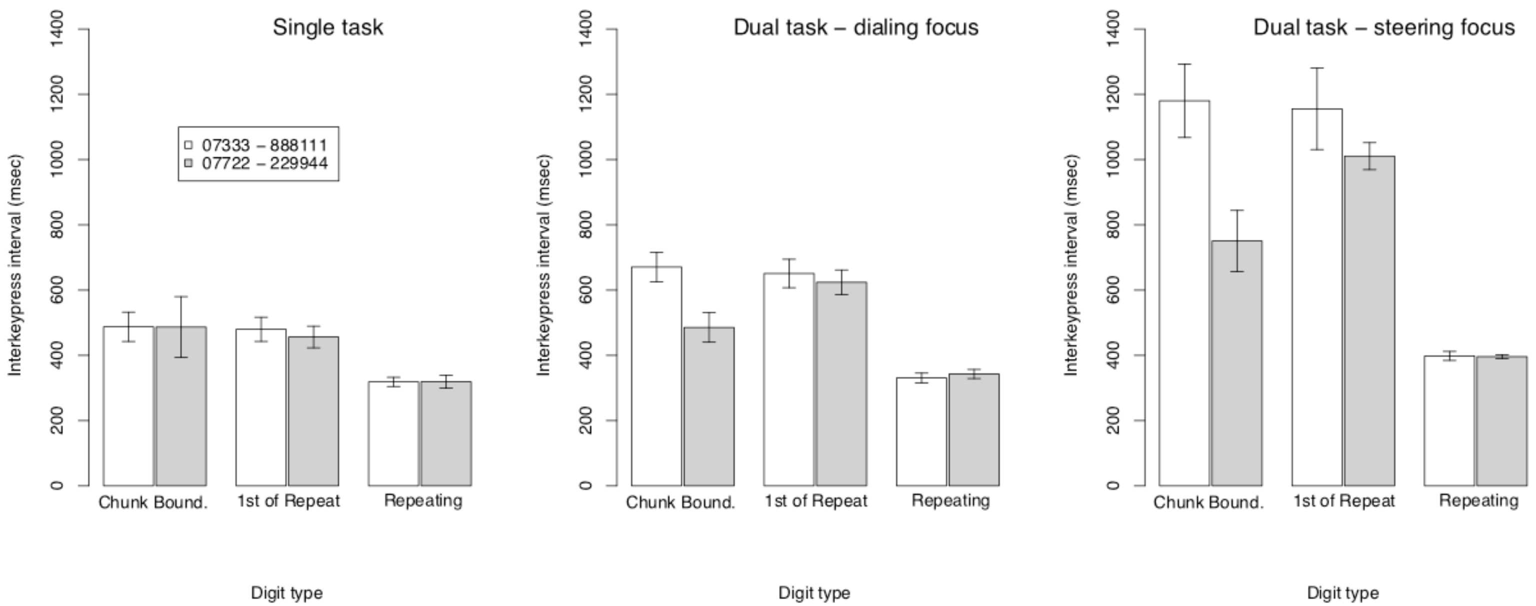

Figure 3.14: Bar plot of mean interkeypress interval in single-task (left), dual-task dialing focus (middle) and dual-task steering focus trials (right).
Bars show mean performance for the congruent (white) and incongruent number (gray), for the digits at chunk boundaries, at the start of a repeating series, and in the middle of a repeated series. Error bars show standardized errors.

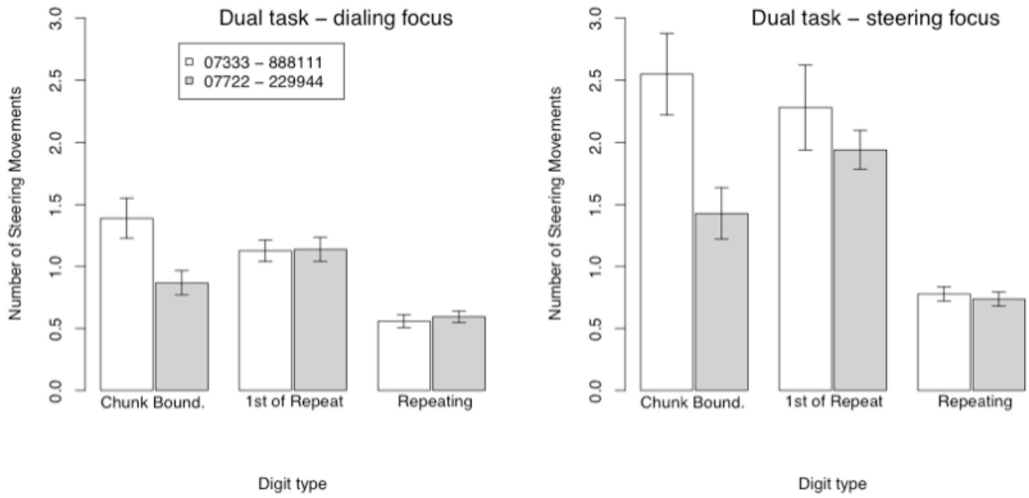

Figure 3.15: Bar plot of mean number of steering movements in dual-task dialing focus (left) and dual-task steering focus trials (right).
Bars show mean performance for the congruent (white) & incongruent number (gray), for the digits at chunk boundaries, at the start of a repeating series, and in the middle of a repeated series. Error bars show standardized error.

Figure 3.15 shows the mean number of steering movements for each of the task priority conditions for each of the three different types of digits in the dialing task: at the chunk boundary, at the first digit of a repeated series of keypresses (if this was not a chunk boundary), and within a series of repeated digits. Data are shown for both the congruent (white columns) and incongruent (grey columns) numbers. The results of a 2 x 2 x 3 (phone number x task priority x digit type) repeated measures ANOVA are shown in Table 3.2. As can be seen in Table 3.2 and Figure 3.15, there were similar effects for the steering data as found in the interkeypress interval data: participants made more active steering movements at positions where the previous analysis found an increase in the duration of the interkeypress interval.

Specifically, participants made more active steering movements when they prioritized driving, and these steering movements occurred mainly at the natural breakpoints in the number (i.e., at the chunk boundary or when there was a change in digit). In contrast, when a digit was repeatedly dialed, there was little difference in observed steering counts between the two priority conditions.

With regards to differences in steering behavior between the two numbers, the main interest was in the interaction between phone number and digit type, because the interkeypress interval data suggested that the only difference in how the two phone numbers were dialed occurred at the chunk boundary. As expected, there was a significant interaction between phone number and digit type, and follow-up tests showed that there were significantly more steering events at the chunk boundary when participants dialed the congruent number (M = 1.97, SD = 0.66) than when they dialed the incongruent number (M = 1.15, SD = 0.44), $F(1,10)$ = 10.58, $p < .01$. This result corroborated the keypress data, as it suggested that participants were choosing to attend to the road only when there was a change in digit at the chunk boundary of the number. There were no reliable differences between the congruent and incongruent numbers when there was a change in digit in the number alone (M = 1.70, SD = 0.68 vs. M = 1.54, SD = 0.35), $F(1,10)$ = 1.15, *ns*, or when a digit was repeated in the number (M = 0.67, SD = 0.17 vs. M = 0.67, SD = 0.15), $F(1,10)$ = 0.01, *ns*.

3.11. Model 3C

3.11.1. Model structure and parameters

The model in general had the same structure as before. However, whereas preceding models had fixed values for the interkeypress times at every keypress, here the model has different values depending on the *type* of keypress. To adapt the model, single-task data was used to measure the average interkeypress times for different digit types. The times used in the model are reported in Table 3.3.

Table 3.3: Model values for the interkeypress times (in msec) per digit type (rows) and phone number (columns), as measured in single-task.

Digit type	Phone number	
	07333 - 888111 (Congruent)	07722 - 229944 (Incongruent)
Chunk boundary	487	487
First of repeating series	480	456
Repeating digit	319	319
Very first digit	501	529
Second digit	501	NA (is a first of repeat)

New to this model compared to predecessors was also the incorporation of a finger-repositioning cost. This cost reflected that if a participant did not continue typing a series of repeating digits in one go, the model had to reorient the fingers on the keyboard, and in that way took a bit longer to dial the digit compared to single-task observations. The experimental set-up did not allow an explicit way of measuring this cost, so it was assumed that this took 170 msec. This is a little higher than the difference between the single task interkeypress interval at the start of a repeating series and within a repeating series. This measured time difference could also be considered a finger repositioning cost. A

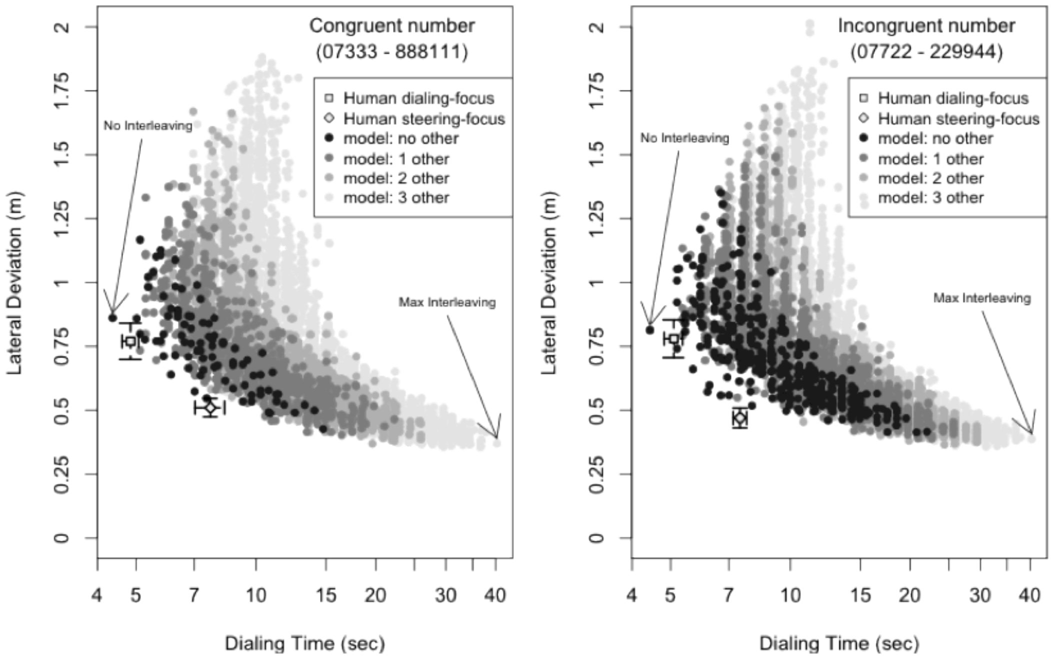

Figure 3.16: Model predictions for total dialing time and mean lateral deviation of the vehicle for strategy alternatives.
Color of the dot indicates how often the strategy interleaved at positions other than the chunk boundary and the first digit of a repeating series (see legend). Human mean performance with standardized error bars is given for the dialing focus condition (square) and steering focus condition (diamond).

choice was made to set this value for the switch cost in the model slightly higher than the measured difference in single-task, as in the dual-task situation the movement might be less well rehearsed, and therefore slower.

3.11.2. Model results

Figure 3.16 shows the performance space for strategy alternatives (grey dots) in the congruent (left) and incongruent phone number (right). The grey scale value of each dot represents a specific strategy type, which will be discussed in more detail below. On top of the model data, human performance is plotted in the steering focus (diamond) and dialing focus (square) conditions.

As in the previous experiment, again performance is near the predicted optimal trade-off curve. In fact, for this model, the human data lies slightly outside of the model space, but very near the optimal performance trade-off curve.[10] Similar to before, performance in the dialing focus condition was on the left side of the trade-off curve, where attention was interleaved infrequently – at most one time. Performance in the steering focus condition was further to the bottom-right on the trade-off curve, but not at the position where driving performance was best (i.e., it was not at the far right). Again, this suggests that participants made a trade-off that achieved reasonable driving performance, while not requiring a very long dialing time (note that the horizontal scale is logarithmic).

Can this trade-off curve also explain why participants tended not to interleave in the middle of a series of repeating digits? To answer this, Figure 3.16 highlights different types of strategies using different shades of grey. The shade was based on the number of times that the strategy interleaved at a point other than either the chunk boundary or the first digit of a series of repeating digits (i.e., at a point other than a natural breakpoint): never (black points), once (dark grey points), twice (lighter grey points), or three or more times (lightest shade of grey). In a sense, the lighter the color, the more 'extra' breakpoints (besides the 'natural' breakpoints) are being inserted.

---

[10] This might be because the value used as a starting point for the drift of the car (0.450 meter) was a little higher than the observed value in human data for this experiment (0.431 meter). If this value was lower, the mean drift of the car could be slightly reduced.

This categorization did not cluster the performance space into completely isolated regions. The slight overlap in performance regions can be explained as each category contained some strategies that differed in the positions of interleaving (e.g., at a repeating digit or not), while having a similar total number of positions at which attention was interleaved (e.g., three times). This made the trial times for these strategies, independent of "grey scale", about equal. Note also that the incongruent number had more black dots compared to the congruent number. This was because the incongruent number had more positions that were 'natural' breakpoints (i.e., either a chunk boundary or the beginning of a series of repeating digits), and therefore the number had more strategy alternatives that interleaved solely at such positions.

The critical observation to take from Figure 3.16 is that solely interleaving tasks at the natural breakpoints in the dialing task (i.e., the black data points), allowed performance to be close to the optimal trade-off curve (i.e., the bottom left arch of the cloud). In other words, the modeling analysis shows that in general it was inefficient to interleave tasks in between dialing a pair of repeating digits. Participants might therefore have avoided those strategy types, as those strategies did not make a valuable speed-accuracy trade-off between dialing speed and driving accuracy.

3.12. Discussions of results

The results of this study and model replicated the findings in experiment 3B: participants' performance changed as a function of their priorities. In addition, performance was adapted in such a way that performance was near the predicted optimal performance trade-off curve. A novel finding of this study was that motor cues can also act as natural breakpoints. Figure 3.16 illustrated why interleaving at these points, compared to others, was beneficial: it brought performance closer to the optimal performance trade-off curve.

3.13. General discussion

3.13.1. Summary of this Chapter

The experiments in this Chapter form a series of critical tests to see whether people interleave tasks if and only if they reach a subtask boundary. Preceding work demonstrated that interleaving at subtask boundaries, or "natural breakpoints", is beneficial in many ways (e.g., Bailey & Iqbal, 2008; Bogunovich & Salvucci, 2010; Miyata & Norman, 1986; Payne et al., 2007; Salvucci & Bogunovich, 2010). I tested some of the boundary conditions of this general claim, using a dialing-while-steering setting in which chunk boundaries form the subtask boundaries (Salvucci, 2005).

Results show that whether people interleave at chunk boundaries depends on (1) their priorities, (2) the number of chunk boundaries, and (3) the availability of other types of "natural breakpoints" (in experiment 3C: motor cues). Participants who set fast dialing as their priority objective interleaved dialing for steering less frequently than participants who set safe driving as their priority objective. To achieve these performance objectives, participants sometimes omitted interleaving at some chunk boundaries (when dialing speed was the main focus), or inserted additional points of interleaving (as in the UK number in experiment 3A and 3B, in particular when safe driving was the priority).

The modeling analysis highlighted an additional advantage of interleaving at natural breakpoints: this gives beneficial speed-accuracy trade-offs. Strategies that interleave at natural breakpoints typically lie closer to the optimum performance trade-off curve than other strategies (see Figure 3.16).

3.13.2. Contributions to the literature

Contribution 3.1: A series of critical tests to investigate whether people interleave if and only if they reach a chunk boundary

The first contribution of this Chapter to the literature is the series of critical tests that I conducted to investigate the hypothesis (Salvucci, 2005) that people interleave dialing for steering if and only if they reach a chunk boundary in the phone number. As summarized above, this is not what people do per se. Rather, the chunk boundary is taken as a *cue* to interleave tasks. Whether people follow up on this cue depends on three factors: the priority of the user, the number of available chunk boundaries, and the availability of other natural breakpoints such as motor cues.

Contribution 3.2: Interleaving at natural breakpoints offers valuable speed-accuracy trade-offs

The results of experiment 3C and the associated model identified another reason why interleaving at "natural breakpoints" typically is beneficial. Interleaving here, rather than at other points, offers valuable speed-accuracy trade-offs (see Figure 3.16).

These trade-offs can be explained as follows. In the single-task dialing task, participants incurred a retrieval cost (experiments and models 3A-3C) or a finger repositioning cost (experiment and model 3C) at every natural breakpoint. In the single-task setting, these costs were not incurred at other points. In a dual-task situation, these costs were still incurred at the natural breakpoints, independent of interleaving pattern. However, they were now also incurred at other points, but *only when a participant interleaved there*. That is, at such "non natural breakpoints", additional costs would be incurred when interleaving. People could avoid these additional costs by interleaving at the natural breakpoints, where the costs would be incurred anyway. Experimental and modeling results suggested that this is mostly what people did when a sufficient number of breakpoints was available to them, and when interleaving was beneficial to their objective.

Contribution 3.3: Motor cues can form natural breakpoints

Preceding studies on the role of natural breakpoints in the driving domain have solely focused on chunk boundaries (Brumby, Howes et al., 2007; Brumby, Salvucci et al., 2007; Brumby et al., 2009; Salvucci, 2005). In experiment 3C I demonstrated that other factors, such as motor cues, can also trigger task interleaving. Moreover, these can offer similar speed-accuracy trade-off advantages.

Contribution 3.4: Critical reflection on “rationality” in distracted driving

Another contribution of this Chapter is a reflection on what constitutes “rational” behavior in a distracted driving task. The modeling analysis suggested that any type of strategy that lies on the optimal performance trade-off curve (see Figure 3.4) is in some way rational. This is because for these strategies performance on a secondary task (e.g., dialing) is the best possible, given a performance criterion for the primary task (e.g., steering). For strategies that do not lie on this curve, there is always an alternative that would achieve a similar score on one measure (e.g., dialing), while achieving better performance on another measure (e.g., steering).

Note that not all strategies that lie on the optimum performance trade-off curve might seem “rational” to an outsider. For example, if a driver sets fast dialing as their priority, the resulting driving performance might not seem rational from the perspective of driver safety. That is, “rational” behavior is not equivalent to *desirable* behavior on the road.

However, the notion of rational, but not desirable, behavior is in line with observations in the real world where people have been noted to continue to engage with various secondary tasks while driving (e.g., Crowd Science, 2009; Diels et al., 2009), and with test-track studies that demonstrate drivers’ unsafe behavior (Horrey & Lesch, 2009; Horrey, Lesch, & Garabet, 2009). This has implications for driver safety, which are discussed below.

Interestingly, in all three studies, drivers that set safe driving performance as their main objective (while also performing a secondary task) did not achieve the safest possible driving performance as predicted by the model. Clearly, performance on both tasks was taken into account, as safer driving performance could only be achieved at the cost of completing the dialing task more slowly.

3.13.3. Implications for driver safety

The study and model results also have implications for driver safety research and for efforts for reducing driver distraction. The finding that people tend to interleave tasks at natural breakpoints has a positive practical implication: efforts on reducing the distracting effect of secondary task devices (e.g., cell phones, PDAs, navigation devices) by incorporating natural breakpoints in the task structure are worthwhile. For example, interfaces could be designed to require repeating motor actions at points where it is desirable to have continued engagement with the interface. At other points these action sets could be made short to encourage task interleaving.

Some researchers have started to work on such systems. For example, in one system, natural breakpoints are considered in deciding when to provide users with system-driven interruptions (e.g., when to provide an e-mail alert, Iqbal & Bailey, 2010). This requires one to be able to identify the natural breakpoints within a task. In practice, this might not always be feasible, as the task might not have a clear (hierarchical) structure, and devices might be appropriated in unanticipated ways.

Unfortunately, even in cases where natural breakpoints can be identified easily (such as in the set-up used in this Chapter), there is another danger lurking: the intended safety benefits might not always arise, due to the driver's objective. The results of the studies clearly showed that whether people interleave at natural breakpoints depends on their priorities. If safe driving is not a priority, participants might completely ignore the cues from natural breakpoints. It therefore seems worthwhile to keep on promoting driver safety - if drivers set safety as their first priority they will be safer compared to situations where they are not, even when they are dual-tasking.

The modeling analysis helped in quantifying how well participants adhered to the priority objective. The model highlighted that even when people were prioritizing driving, they still did not act in the safest possible way. This seems in line with observations in the real world where people continue to engage with various secondary tasks while driving (e.g., Crowd Science, 2009; Diels et al., 2009), and with test-track studies that demonstrate drivers' unsafe behavior (e.g., Horrey & Lesch, 2009; Horrey et al., 2009). Given this demonstrated unsafe behavior, it seems even more worthwhile that appropriate cognitive engineering solutions are sought to minimize distraction and encourage appropriate task interleaving. As was demonstrated here, computational cognitive modeling can help in this effort (cf. Kieras, in press).

3.13.4. Limitations

General limitations

The modeling analysis proved valuable in making judgments about the relative success of different types of strategies (i.e., interleaving only at the chunk boundary, or also interleaving at other positions). A limitation of the studies and models is that, for each priority objective, I was not able to identify which strategy alternative(s) were the overall best. In order to do this, a stricter criterion for both the participants and the model is needed about how to trade-off one task for the other. This can be done in two ways. One option is to set an explicit criterion for one of the tasks (e.g., a specific driving performance level), as was done in for example in Gopher (1993) and Gopher et al. (1982). Alternatively, performance on both tasks could be combined into a single objective score that reflects the desired trade-off, and the model can be used to determine which strategies achieve the highest score. Using this information, optimum performance can be defined and compared to human performance data (cf., Howes et al., 2009). This approach will be taken in subsequent Chapters.

A limitation of the results in experiment 3C specifically is that although it was argued that the findings reflect effects of memory and motor cues on performance, it is unclear whether these effects are directly caused by aspects of motor cues, or by aspects of memory cues, as it is impossible to directly study the mental representation of the number in this task. For example, the breakpoints that occur between groups of digits might also be explained because repeating digits are memorized as one chunk (e.g., "three nines"). Future work should make a more thorough distinction between effects of memory representation and effects of motor cues.

In line with the above view, the model does not make a strict distinction between motor cues and memory cues. Both were parameterized and expressed in the same unit: time. No further claims about the underlying cognitive processes were made, as such an analysis would go beyond the level of detail that the experimental measurements provided (Howes & Young, 1997). Fortunately, even when the exact distinction between what was caused by motor cues and what was caused by memory cues is loosened, an important high-level conclusion still remains: people tend to interleave at natural breakpoints when this complies with their priority objective.

It is to be expected that performance will also change when the driving simulator is made more realistic. A more complex simulator can provide a more challenging environment to the participant, for example as a result of other, unpredictable traffic. However, as the user gains more control over the car, this also gives way for further strategic adaptation to the challenges at hand. For example, by reducing the speed of the car (e.g., Cnossen et al., 2004; Iqbal et al., 2010).

Another limitation of the set-up is that, because the task environment maps to a real-world setting, participants might come into the lab with biases about how the task should be completed. For example, participants might always want to avoid crossing the lane boundary, as this is dangerous to do in a real car. In effect, such biases might override instructions on how to perform the task (e.g., how to prioritize), and limit the possibilities to test people's ability to adapt performance to specific priorities and instructions. In the next Chapter, I will therefore use an alternative task environment that captures the core aspects of driver distraction, but that does not have the potential natural biases associated with the driving simulator set-up.

One final limitation of the task set-up is that it only allowed me to infer the strategy that people applied for task interleaving indirectly, by considering delays in interkeypress interval and increases in steering movements. As a result, no firm statements were made about the exact patterns in which people interleaved; only general patterns were discussed (i.e., why interleaving at natural breakpoints is beneficial). More confidence in the strategies that people applied can be achieved in two ways. First, more fine-grained measurements of performance can be collected, such as eye tracking data. Second, more control can be exerted over the task environment over which tasks are visible to the participant at each moment in time. This latter approach will be taken in Chapter 4.

Model limitations

The conclusions that were drawn in this Chapter about the (sub-) optimality of attention interleaving all presuppose that the cognitive model of dialing-while-steering is correct. However, the model incorporates assumptions that might turn out to be false. Any changes that are made to these assumptions can have implications for the conclusions that are drawn about fine-grained performance, such as conclusions about which specific strategies overlap in performance with human performance. At the same time, independent of the model validation, the experimental results remain intact. For example, the fact that people adapted performance to their priorities, and that the interkeypress interval was longer at some points in the number compared to other points was validated outside of the model.

There are three general assumptions that can be revisited in further modeling efforts. These assumptions are related to: the steering model, moment-to-moment processes, and the strategy space. First, I have assumed that the driving model that was developed by Brumby, Salvucci, and Howes (Brumby, Howes et al., 2007; Brumby, Salvucci et al., 2007; Brumby et al., 2009) based on experimental datasets (Salvucci, 2001; Salvucci & Macuga, 2002) was correct. The model was not independently validated for the driving simulator that was used in the studies reported in this Chapter. Moreover, alternative models for steering control have been proposed in the literature (e.g., Hildreth, Beusmans, Boer, & Royden, 2000; Land & Lee, 1994; Salvucci & Gray, 2004). One research effort could be to explore how predictions of performance differ between these different models.

A second assumption in the model was about moment-to-moment processes. For reasons of simplicity, most processes were assumed to have fixed time costs. For example, a change of the steering wheel angle was made every 250 milliseconds, and every digit that was typed was modeled using a fixed time cost. These mean values are only reasonable approximations of the more complex processes that underlie performance. Each of these processes can be investigated in more detail.

A third assumption in the model is about the strategy space. In the model, strategy alternatives were explored based on two parameters: the number of digits typed within one round, and the time spent on steering when making steering corrections. This division in strategies assumes a serial bottleneck in performance, where only one task at a time can be attended to. In practice though, people might be able to dedicate some of their resources (e.g., eyes) to one task, while dedicating other resources (e.g., hands) to another task (Salvucci &

Taatgen, 2008, 2011; Wickens, 2002, 2008). For example, people might quickly glance at the road while typing in a digit on the phone.

To offer an account for these fine-grained behaviors, the parameters of the strategy space need to be revised. In addition, to validate such strategies, more fine-grained data needs to be collected about performance. For example, eye-movements could be useful for these purposes (e.g., Borst & Taatgen, 2007; Hornof & Zhang, 2010; Horrey et al., 2006; Ralph, Gray, & Schoelles, submitted). Given that the model was used to explore performance of extreme strategies (i.e., no interleaving, and interleaving dialing for driving after every digit) as well as many strategies in between these extremes, it is expected that performance of more refined strategies falls inside the brackets (Card et al., 1983; Kieras & Meyer, 2000) of the explored space.

3.13.5. Extensions

Apart from the limitations mentioned above, there are other extensions possible for the model. The modeling analysis draws attention to the fact that participants can adopt markedly different strategies in a given dual-task setting, and the choice of strategy can have a large impact on performance. It might be valuable to consider how this level of strategic variability is influenced by individual differences in ability and performance. Recent research for instance suggests that there can be large individual differences in dual-task performance, which can sometimes be anticipated based on single-task performance (Watson & Strayer, 2010). There might be scope to consider how differences between individuals, such as the fact that some people dial faster than others, might have lead to differences in dual-task interleaving strategy.

For instance, it might be the case that slow dialers might need to interleave tasks more frequently to reach the same level of performance as fast dialers. One possible way to assess this idea is to adjust parameters in the model to fit individual differences in performance (cf. Howes et al., 2009) and see whether this leads to a marked change in the location of specific strategies within the predicted tradeoff curve. If so, this might provide a possible explanation for why people adopt different dual-task strategies based on differences in single-task performance (Watson & Strayer, 2010). This will be explored in Chapter 4.

The dialing task was relatively simple. However, the core characteristics of this memory based manual task are found in the real world. For example, people have been noted to update their social network status while driving (Crowd Science, 2009). Moreover, by analyzing results with a model, possibilities are created for testing the effects of other types of interfaces and tasks on dual-task performance - model based evaluation (Kieras, in press). This can be achieved by modifying parameter values. Some examples are mentioned below.

First, parameters related to the task constraints can be changed. For example, the ease of access to the interface and its effect on dual-task performance could be manipulated by changing the "switch cost" parameter – the further the phone is away from the participant, the higher the cost. Preceding studies have shown that such information access costs can significantly change interaction with interfaces (e.g., Gray et al., 2006).

Parameters related to cognitive constraints can also be changed. For example, the complexity of the information stored in memory can be manipulated, using the associated parameter for retrieval cost.

Again, adding complexity to the information is likely to influence performance (e.g., Borst et al., 2010), in that people might want to avoid the associated high resumption costs and memory load (Altmann & Trafton, 2002). A model could be used to highlight what the trade-offs would be between continuing a task versus interleaving.

Finally, parameters about motor actions can be changed. In the case of the dialing task, the parameters for interkeypress intervals and those for repositioning the fingers could be changed. The settings for these parameters were crucial for the current findings. As Fitts' law (Fitts, 1954) would predict, the size of buttons on an interface, and their relative distance, can influence the speed with which one can type. The model could be used to test what impact such interface changes would subsequently have on dual-task performance. This can help engineers in making informed decisions about interface design for dual-task scenarios.

3.14. Conclusion

In conclusion, the work in this Chapter helped to better understand when and why people interleave at natural breakpoints in a dual-task set-up. This depends on the frequency and type of breakpoint that is offered, but mostly on the priority objective. Given a clear objective, people will try to interleave in such a way that performance lies on the optimum performance trade-of curve (see Figure 3.4), in a region where the desired criterion for each task is achieved. To achieve this, depending on the priority, people might insert additional points of interleaving or omit interleaving at some natural breakpoints. In general though, interleaving at natural breakpoints compared to other points is beneficial, as this offers useful speed-accuracy trade-offs. Although performance tended to be on an optimal trade-off curve in the studies reported here, this does not mean that performance was always "safe". Safe performance is only achieved when safety is the driver's priority. It is therefore important to continue to promote driver safety.

3.15. Appendix to Chapter 3

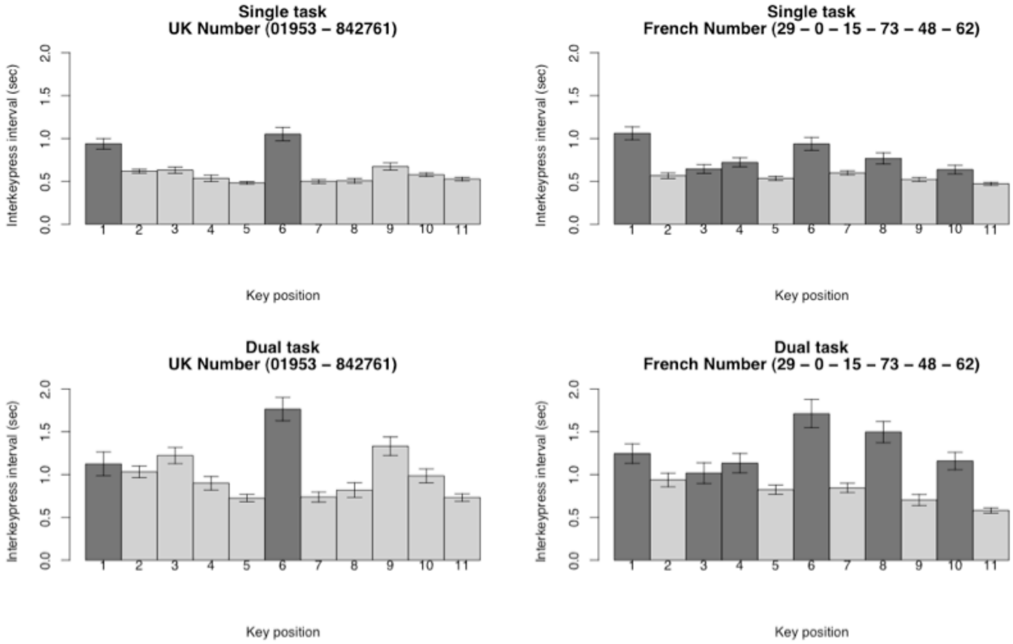

Figure 3.A.1: Interkeypress intervals per digit
in single-task (top) and dual-task trials (bottom) for the UK (left) and French number (right). Color indicates whether the interval was at a chunk boundary (dark grey) or not (light grey). Error bars show standard error.

Chapter 4. Strategic Adaptation to Task, Individual Differences, and Objective Payoff Functions in a Tracking-while-Typing Setting[11]

Abstract

In this Chapter I report the results of two dual-task studies in which participants performed a tracking and a typing task under various experimental conditions. Objective payoff functions were used to provide explicit feedback on how participants should trade-off performance between the tasks, and to make the definition of "optimum" performance less ambiguous. The first study tested whether participants adapted performance to various manipulations of task characteristics, given a fixed payoff function. A simple model of aggregate performance was used to explore performance for a variety of strategies for performing the task, and to see if people achieved the maximum score. In a second experiment, the nature of the payoff function was manipulated. Individual differences in performance arose in this experiment, leading to refinement of the model to capture those performance differences. Overall, the results showed that people adapted their performance to three factors: (1) task (difficulty) characteristics, (2) individual differences in skills, and (3) payoff function. The modeling analysis showed that people adapt their performance in such a way as to try and maximize the payoff value. This is not to say that performance was optimal on every trial. Limitations are discussed. The work makes three contributions to the literature:

---

[11] Parts of this Chapter (esp. experiment 4.1 and model 4.1) have been published as: Janssen, C. P., Brumby, D. P., Dowell, J., Chater, N., & Howes, A. (2011). Identifying optimum performance trade-offs using a cognitively bounded rational analysis model of discretionary task interleaving. *Topics in Cognitive Science, 3*, 123-139.

(1) it introduces a method to identify optimum performance in a dynamic dual-task setting, (2) it demonstrates how dual-task performance adapts to the task, individual differences in skill, and payoff, and (3) it re-appreciates the flexibility of human performance, both when considered within and between individuals.

4.1. Introduction

In the previous Chapter I described a series of studies that investigated when and how people interleave their attention in a dialing-while-steering setting. The results of these studies showed how people adapt their multitasking interleaving strategy to task characteristics (i.e., number of "natural breakpoints" in a phone number) and their objective (i.e., which of the two tasks they gave greater priority to). A model was used to demonstrate how participants made "optimal" performance trade-offs when a clear objective was present. However, as was discussed in the general discussion of that Chapter and in detail below, in some sense this prediction of optimality was under constrained.

Here these weaknesses are being redressed through the introduction of a formal, objective, explicit payoff function that captures performance on both tasks in a single score. This makes it easier for both the participant and the modeler to identify the optimum strategies: These are the ones that on average achieve the highest possible payoff score. A new task paradigm involving a tracking and typing task will be introduced to have more experimental control. In addition, the paradigm will be used to demonstrate how three factors determine the strategies that people use to interleave tasks: task constraints, a payoff function (formalized objective), and individual differences in skill.

4.1.1. Why payoff functions can be useful

Optimal trade-off curves capture how performance on separate tasks varies together systematically (Navon & Gopher, 1979; Norman & Bobrow, 1975). For all points on this curve performance on a secondary task (e.g., dialing time) is the best possible, given a performance criterion on a primary task (e.g., lateral deviation of a car), and vice versa. In Chapter 3 it was observed that when participants changed their priorities, their performance moved along the POC curve accordingly. That is, when safe driving was their priority, a better criterion for driving was set, and given that criterion a fast dialing time was achieved.

However, in some sense this prediction of "optimality" is under constrained. First, the curve is long, and all points on this trade-off curve are optimal for some combination of performance criteria on the tasks. Any claim that people are optimal is therefore relatively weak - if performance moved down a little on the curve, people would still be optimal. Ideally, a prediction of optimality should be given about a specific region on the curve.

A second problem arises from the fact that, typically, performance on both tasks is expressed in different units. For example, in the driving task performance was expressed in meters (the accuracy at keeping the car near lane center), whereas for the dialing task it was expressed in seconds (the speed with which the phone number was dialed). Participants demonstrated to be good at trading-off performance on both tasks, by performing in a way that was predicted to lie on the trade-off curve. Nonetheless, it might have been difficult for them to decide how to trade-off the two units against one another. That is, it might have been difficult to decide how many meters improvement on the driving task was worth an improvement of one second on the dialing task. This is especially hard, given recent empirical evidence that people have difficulty in maintaining internal scales of values, and in making exact judgments about values (e.g., Chater & Vlaev, 2011; Kurniawan et al., 2010; Vlaev, Chater, Stewart, & Brown, 2011).

Related to the above points, in the studies from the previous Chapter and others (e.g., Brumby et al., 2009; Gopher, 1993; Horrey et al., 2006; Levy & Pashler, 2008), the verbal instructions given to participants to prioritize one task over another might have been open to differences in subjective interpretation. This again makes it hard to identify whether people complied with an instructed priority or not.

One solution to this problem of defining an optimal strategy is to use quantitative, objective, explicit payoff functions, which have been used in experimental psychology studies before. They can be used to provide explicit instructions to participants on how the required tasks should be performed. For example, a payoff function might be used to inform participants how they should trade responding quickly to the appearance of stimuli against the risk of making a response error (e.g., Schumacher et al., 1999). Howes, Lewis, Vera and colleagues (Howes et al., 2009; Howes et al., 2004; Lewis et al., 2004; Vera et al., 2004) have taken the use of payoff functions one step further by putting forward the hypothesis that skilled human performance can be understood as a utility maximization problem that is constrained by the task, cognitive architecture, knowledge, and experience. In other words, the idea is to assume that people are boundedly optimal (see also discussion in Chapter 2 of this thesis). A payoff function can be used to identify this optimum solution, as was done successfully in the Psychological Refractory Period (PRP) paradigm (see, Howes et al., 2009, for details).

However, in some respects the PRP task is simple: stimuli appear at their own pace and single responses need to be made. Slightly more complex are the task environments that are studied in this thesis: dual-task dynamic discretionary task interleaving scenarios, where participants need to decide themselves when to switch attention from one dynamic task to another. In these scenarios the use of payoff functions has been limited. For example, payoff functions have been used to motivate participants to perform to a certain criterion (e.g., Hornof, Zhang, & Halverson, 2010), or to demonstrate that participants use payoff as an incentive to spend more time on one task over another (e.g., Wang et al., 2007).

In this Chapter, I will also use a payoff function in a dynamic discretionary task interleaving paradigm. However, I will follow the methodology of Howes and colleagues (2009) and use a payoff function to investigate whether participants adopt the *optimum* strategy for maximizing payoff. This is in line with the original intention of work that inspired research on Performance Operating Characteristics, Signal Detection Theory and Receiver Operating Characteristics, to identify strategies that maximize utility (Swets et al., 1961). It is also in line with recommendations in the human factors literature on task scheduling (see e.g., Moray et al., 1991).

Furthermore, this methodology of using explicit feedback to influence priorities has independently emerged as a technique in various other domains of psychology, most specifically studies of dynamic task interleaving (Hornof & Zhang, 2010; Neth et al., 2008), motor movement (Jarvstad et al., in press; Juni et al., 2011; Maloney & Mamassian, 2009; Maloney & Zhang, 2010; Trommershauser et al., 2003a, 2003b, 2008; Warren et al., 2012), and vision (Ballard & Sprague, 2007; Reichle & Laurent, 2006; Tatler et al., 2011). See also a discussion in Chapter 2.

4.1.2. Task paradigm: tracking-while-typing

The task environment that is used here captures relevant aspects of driver distraction, but at a different level of application (Salvucci & Taatgen, 2011) than Chapter 3. This allowed for more experimental control (e.g., for identifying when participants interleaved attention between tasks). In the task environment participants had to keep a randomly moving cursor inside a circular area and type a string of digits. However, they could only see and control one task at a time. Participants' performance was captured in a single payoff score, which reflected the payment the participant received at the end of the study.

Tracking tasks have been used in several multitasking studies (e.g., Ballas, Heitmeyer, & Pérez-Quiñones, 1992; Chong, 1998; Chong & Laird, 1997; Gopher, 1993; Hornof et al., 2010; Kieras et al., 2000; Lallement & John, 1998; Martin-Emerson & Wickens, 1992; Salvucci & Taatgen, 2008; Strayer & Johnston, 2001). The work presented here builds on and extends this work by showing how a payoff function enables one to bind normative cognitive models with experimental observations of multitasking behavior, and specifically, to show how strategy choice in dynamic discretionary task interleaving paradigms can be better understood by comparing observed performance to a prediction of optimal performance for maximizing payoff.

4.1.3. Structure of this Chapter

The remainder of this Chapter is structured as follows. First, I will report on an experiment in which people had to adapt to several manipulations of task characteristics, given a fixed payoff function. This is followed by development and evaluation of a model of average performance in this task that suggests that people adapt their performance to maximize payoff. I then report on an experiment in which not only task characteristics were manipulated, but also the nature of the payoff function. The experimental results showed that there were individual differences in performance. As a consequence, the model was revised to capture these differences. The modeling demonstrates how the strategy and performance space change due to differences in (1) task characteristics, (2) individual differences in skill, and (3) payoff function. In the general discussion the contribution to the literature, and limitations of the studies and models are discussed.

4.2. Experiment 4A: Adaptation to task conditions, given a payoff function

In this first experiment it will be investigated whether people adapt to different task conditions, given a specific payoff function. It is predicted that as the difficulty of the task changes, the type of strategies that are optimum will change as well. For example, when the tracking task is made harder it is expected that participants pay more frequent attention to this task. In this experiment section, I will investigate whether performance and applied strategies indeed differ between conditions. This will be followed by development of a cognitive model to see if participants applied optimal strategies, given their local context.

4.2.1. Method

Participants

Eight participants (4 female) between 20 and 35 years of age (M = 23 years) from the subject pool at UCL participated for monetary compensation. Payment was based on performance (details are provided in the Materials section). The total payment achieved by participants ranged between £7.13 and £11.45 (M = £9.14).

Materials

The dual-task setup required participants to perform a continuous tracking task and a discrete typing task, presented on a single 19 inch monitor with a resolution of 1280 x 1024 pixels. Figure 4.1 shows the layout of the tasks on the display, and Figure 4.2 shows the physical set-up of the experiment. The typing task was presented on the left side and the tracking task on the right. Each task was presented within a 450 x 450 pixels area, with a vertical separation of 127 pixels between the tasks.

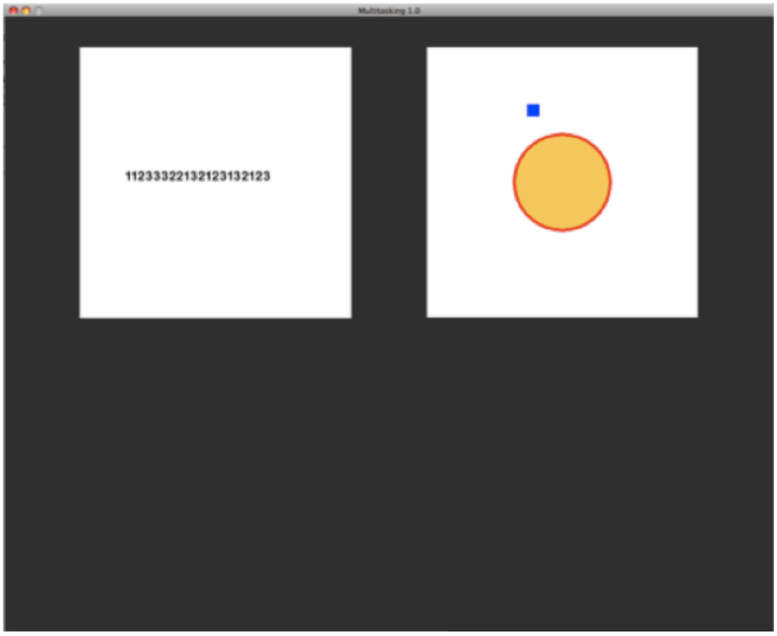

Figure 4.1: Position of the two tasks in the interface

Figure 4.2: Physical set-up of the experiment.
The tracking task is controlled with a joystick, and the typing task with the numeric keyboard.

The tracking task required participants to keep a square cursor that drifted about the display in a random fashion inside a target circle (see Figure 4.1). The cursor was 10 x 10 pixels and the target had a radius of either 80 (small radius, or target) or 120 pixels (large radius/target). A random walk function was used to vary the position of the cursor in the display. The rate at which the cursor drifted about the display was varied between different experimental conditions. In a low noise condition, the random walk had a mean of zero and standard deviation of 3 pixels per update, while in a high noise condition the random walk had a mean of zero and standard deviation of 5 pixels per update. Updates occurred approximately once every 25 milliseconds. To control the position of the cursor in the tracking display, participants used a Logitech Extreme 3D Pro joystick with their right-hand. The drift function of the cursor was suspended whenever the joystick angle was greater than + / - 0.08 (the maximum angle was + / - 1). The speed at which the cursor could be moved was scaled by the angle, with a maximum of 5 pixels per 25 milliseconds.

The typing task required participants to enter a string of twenty digits using a numeric keypad with their left-hand. The string was made up of the digits 1 to 3, where each digit occurred at least six times in a given sequence. Digits were presented in a random order with the constraint that no single digit was presented more than three times in a row in the sequence (e.g., “11233322132123132123” as in Figure 4.1). When a digit was entered correctly, it was removed from the to-be-entered sequence. In this way, the left-most digit on the display was always the next one to be entered. When an incorrect digit was typed, the string would not progress. No additional signal was given to indicate this error.

The study used a forced interleaving paradigm, in which only one of the two tasks was visible and could be worked on at any moment in time. By default the typing task was visible and the tracking task was covered by a gray square. Holding down the trigger of the joystick made the tracking task visible and covered the typing task. Releasing the trigger covered the tracking task and made the typing task visible once more. Input was only for the visible task and any input for the covered task was ignored (recall that the tracking task only received input from the joystick while the typing task only received input from the keyboard).

Design

The study manipulated aspects of the tracking task using a 2 (cursor noise: low vs. high) x 2 (target size: small vs. large) within-subjects design. The main dependent variables were the time required to complete the typing task, the maximum distance of the cursor from the center of the target circle, and the total time the cursor was outside of the target circle.

Participants were remunerated based on their performance using an objective payoff function. Within the payoff function, participants could gain points on the typing task (the faster the better). A penalty of 1 pence was applied for every digit that was typed incorrectly. A penalty was also applied when the cursor moved outside of the target area. These three functions were added together to calculate total payoff (in pounds), as follows:

Payoff = Gain + Tracking Penalty + Digit Penalty (Equation 4.1)

The general equation for the gain function on the typing task was as follows:

$$\textbf{Gain} = 0.15 \times e^{\text{severityOfTrialTime} \times (\text{TotalTrialTimeInSeconds} / 20) + \text{startValue}_{\text{gain}}}$$

(Equation 4.2)

In general, the gain decreased as the total trial time increased. However, the impact this had on the overall payoff depended on the scaling of the scores. In this experiment "severityOfTrialTime" was set to -1 and "$\text{startValue}_{\text{gain}}$" was set to 0.25. This function was determined using pilot studies, to make sure participants mostly gained money.

The general function for updating the tracking penalty function was as follows:

$$\textbf{Tracking Penalty} = \text{compensation} - 0.10 \times e^{\text{SecOutside} \times \text{severityOfBeingOutside} - \text{startValue}_{\text{tracking}}}$$

(Equation 4.3)

In this experiment "compensation" was set to 0, "severityOfBeingOutside" to 1.1090, and "$\text{startValue}_{\text{tracking}}$" to 0.6931. With this function £0.10 was lost when the cursor was outside of the radius for 0.625 s, and £0.20 was lost when it was outside of the radius for 1.25 s.

To encourage accurate typing, a digit penalty deducted £0.01 from the total payoff whenever an incorrect digit was entered. In experiment 4A and model 4A the effect of digit penalty on payoff will not be considered, as the total number of errors was relatively low, and in most trials no errors were made (see results). This issue will be redressed in experiment 4B and the subsequent modeling effort (model 4B).

To avoid participants from losing all their money on a given trial, the payoff function had a minimum score of – £0.20.

Procedure

Participants were informed that they would be required to perform a series of dual-task trials and that they would be paid based on their performance. A participant's payment was based on the cumulative payoff over the course of the study, in addition to a base payment of £3. Participants were told that they would gain more points by completing the typing task as quickly as possible, but that they would lose points if they made a typing error or if the cursor drifted outside of the target area in the tracking task. I did not give participants a formal description of the payoff function, but instead provided explicit feedback after every dual-task trial with the payoff score achieved.

After explaining how to perform each of the tasks, participants performed two single-task training trials for each task and two dual-task practice trials. Participants were instructed that for dual-task trials only one of the two tasks would be visible and controllable at any moment in time, and they were instructed how to switch between tasks using the trigger button on the joystick.

Participants then completed four blocks of experimental trials (one for each experimental condition). In the first two blocks, participants experienced a single noise level, either low or high noise. The noise level was randomly assigned to participants, and balanced across participants. On the first block a radius size (small or large) was also randomly assigned, on the second block the other radius level was assigned. For the third and fourth block this order was repeated, but with another level for noise. For each block, participants completed five single-task tracking trials, five single-task typing trials, and twenty dual-task trials. The dual-task trials were further grouped into sets of five trials, with a short pause between each set. The total procedure took about one hour to complete.

4.2.2. Results

Across all keystrokes in single-task typing trials, participants on average typed 2.5 % (range 0.5 – 5.2 %) of their keystrokes incorrectly (81 out of 3,281 keystrokes). At the trial level, 0, 1, 2, or more than 2 errors were made on respectively 61.9%, 29.4%, 5%, and 3.8% of the trials.

In the dual-task trials, the number of typing errors was also low. Participants on average typed 3.6% (range 1.2 – 5.4 %) of their keystrokes incorrectly (481 out of 13,281 keystrokes). At the trial level, 0, 1, 2, or more than 2 errors were made on respectively 52.5%, 29.4%, 12.0%, and 6.1% of the trials.

The occurrence of errors is interesting, but their occurrence is too low to draw any conclusions from (i.e., on most trials there were no errors). The effect that errors had on performance is therefore not further investigated here, but will be addressed in experiment 4B and model 4B.

For the further analyses of the data, I focused on the last five dual-task trials of each experimental condition. These trials reflect a period during which the participant had had time to adapt their behavior to the payoff function, based on the feedback received. A 2 (cursor noise) x 2 (target size) analysis of variance (ANOVA) was used for all statistical analysis with a significance level of .05.

Overall performance

To start, the effect was investigated that varying aspects of the tracking task had on the time required to complete the typing task, the maximum distance of the cursor from the center of the target circle in the tracking task, and the mean time the cursor was outside the target area. Figure 4.3 also plots the total trial time versus the maximum distance that the cursor moved away from the center of the target in one plot for all four conditions. It can be seen that all four conditions take up a unique point in this performance space. Statistical analysis confirms these findings.

It was found that trial time was significantly longer when there was greater noise in the tracking task (M = 11.17 s, SD = 4.32 s) than when there was a lower level of noise in the tracking task (M = 7.51 s, SD = 2.00 s), $F(1, 7) = 15.07$, $p < .01$. Trials were also longer when the radius of the target in the tracking task was smaller (M = 10.59 s, SD = 4.01 s) than when it was larger (M = 8.09 s, SD = 3.22 s), $F(1, 7) = 11.84$, $p = .01$. There was no significant interaction, $F(1, 7) < 1$.

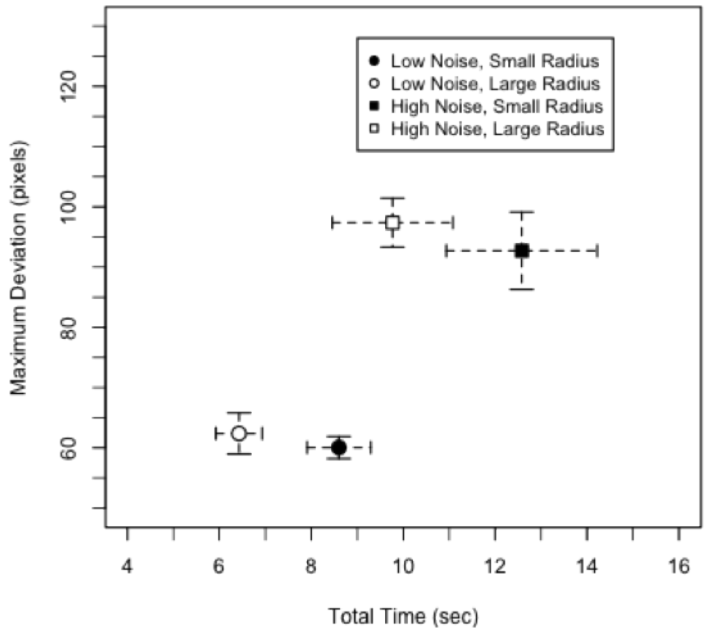

Figure 4.3: Plot of the performance space.
Total time versus maximum deviation. Error bars depict standard errors.

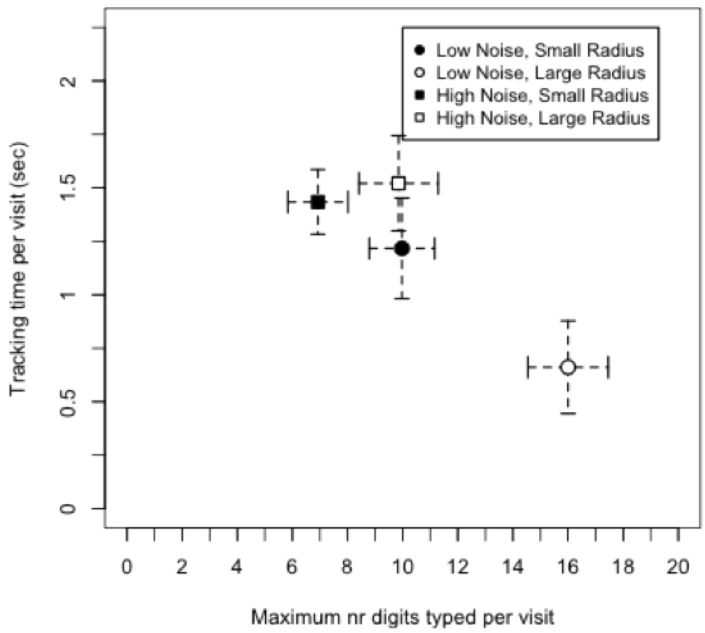

Figure 4.4: Plot of the maximum number of digits typed and time spent tracking, both per visit.
Error bars depict standard errors.

In the tracking task, the maximum distance of the cursor from the center of the target over the course of a trial was analyzed. It was found that the cursor drifted more when there was a higher level of noise (M = 95 pixels, SD = 15 pixels) than when there was a lower level of noise (M = 61 pixels, SD = 8 pixels), $F(1, 7) = 33.42$, $p < .001$. There was no effect of the target radius size on the maximum distance that the cursor drifted over a trial, $F(1, 7) = 1.19$, $p = .31$, nor was the interaction effect significant, $F(1, 7) < 1$.

Another measure of performance in the tracking task was the average time the cursor was outside of the target area per trial. Participants let the cursor remain outside of the target area for longer when there was high noise (M = 0.36 s, SD = 0.45 s), compared to when there was low noise (M = 0.04 s, SD = 0.10 s), $F(1, 7) = 7.28$, $p = .03$. The cursor also spent more time outside of the target area when the radius of the target area was small (M = 0.34 s, SD = 0.05 s), compared to when it was large (M = 0.05 s, SD = 0.11 s), $F(1, 7) = 13.26$, $p < .01$. The interaction was not significant, but there was evidence of a trend, $F(1, 7) = 4.58$, $p = .07$. This trend reflects that in the low noise, large radius condition the cursor never crossed the target area, whereas in the high noise, small radius condition the cursor crossed the target area for over half a second.

These differences in overall task performance between conditions were somewhat expected and unsurprising because they partly reflect differences in the difficulty of the tracking task. A more interesting question was how this performance was achieved. To this end, the dual-task interleaving strategies that were adopted in each condition were investigated.

Strategies

Two aspects determined a strategy: (1) the maximum number of digits typed during a visit to the typing window and (2) the amount of time spent in the tracking window per visit to this window. Figure 4.6 shows these two basic strategy dimensions for each of the four conditions. For the number of digits typed only correct digits were considered. It can be seen that for each experimental condition there was a relatively unique point in this strategy space – strategies differed between conditions.

The maximum number of digits entered per visit (see Figure 4.4) increased with an increase in size of the radius of the target , $F(1,7) = 23.43$, $p < .01$, and it also increased with a decrease in cursor noise, $F(1, 7) = 11.16$, $p = .01$. That is, more digits were typed when it took longer for the cursor to cross the boundary. There was no significant interaction, $F(1,7) = 2.53$, $p = .16$.

It can also be seen in Figure 4.4 that the time spent in the tracking window per visit increased with an increase in the noise associated with the cursor's movement[12], $F(1, 7)=13.55$, $p < .01$. An interaction effect was present as visit time was particularly short in the low noise, large radius condition, $F(1, 7) =8.98$, $p = .02$. There was no significant effect of radius, $F(1, 7) = 1.42$, $p = .27$.

[12] The algorithm for extracting the tracking visits is an improved version of the one used for the data reported in Janssen et al. (2011). This new algorithm takes better account of trials in which no visits to the tracking window were made. Although both methods result in slightly different means and standard error values, the statistical effects are the same for data produced by both methods.

4.3. A model of average performance

The results showed that participants adapted their dual-task behavior to changes in the difficulty of the tracking task by varying the amount of time that was given to each task before switching to the other task. However, what these results did not show was whether participants were adopting a strategy that was *optimal* in terms of maximizing the expected payoff that could be achieved in each condition, both for the individual task (tracking and typing) and the combination of tasks. To answer this question a Cognitively Bounded Rational Analysis model (Howes et al., 2009) was developed of aggregate human performance.

This framework was particularly useful for comparing the performance of alternative strategies, allowing strategies to be discriminated based on the payoff achieved. The model was inspired by the models that were developed for the dialing-while-steering tasks (see Chapter 3). Both dual-task environments share core characteristics, but the current work differs in that it incorporates an explicit payoff function against which various dual-task interleaving strategies can be evaluated. In the next section, the model is described that was used to determine whether people were selecting strategies that would maximize the financial payout that could be achieved in each condition.

4.3.1. Model development

Tracking model

What did people do when visiting the tracking window? The crucial question for developing a model of the tracking task is to understand how people set the angle of the joystick based on the position of the

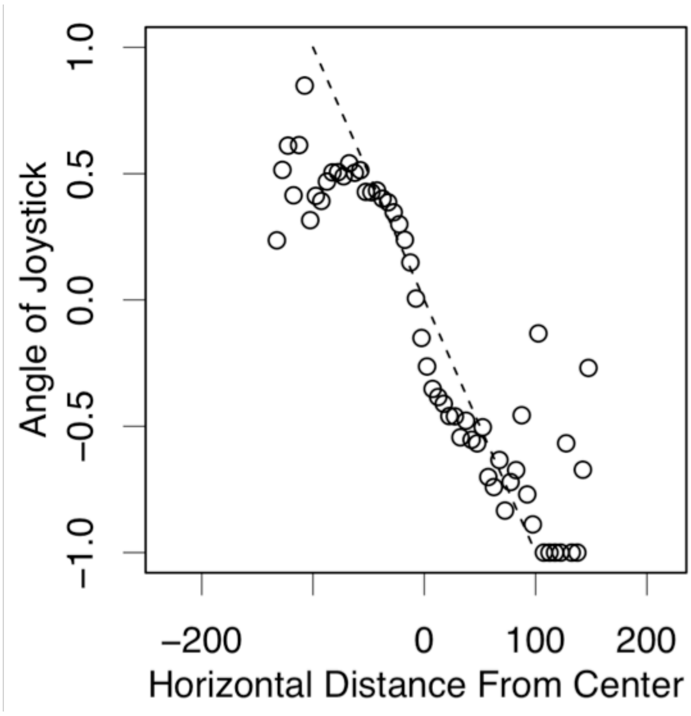

Figure 4.5: Plot of the angle of the joystick as a function of distance from the target. The dashed line shows a fitted function.

cursor in the display. Figure 4.5 shows the mean values for discrete bins of 5 pixels for the horizontal axes (vertical data is similar). A linear function was fitted (shown as a dashed line):

Angle = -0.01 * current distance from target

(Equation 4.4)

The joystick had a maximum angle of (-)1. This shows that participants' behavior in the tracking task can be captured by a simple linear function that sets the angle of the joystick based on the position of the target within the display. To implement this model, as in the experiment, the speed of the cursor was calculated by multiplying the angle of the joystick with a value of 5 pixels. Speed was calculated once every 250 milliseconds of tracking, and the cursor position was updated every 25 milliseconds based on this speed value. As in the experiment, the cursor can only be controlled when the tracking window is open. The total time spent tracking in dual-task was varied as part of the strategy (see below).

Typing model

To model the typing task, model performance was fitted to human single-task typing performance data. To get a measure of how long it took participants to enter a digit in the typing task, the mean single-task interkeypress interval time was taken, which was 260 milliseconds. This value was calculated by taking the mean value of participants' total typing time, and dividing this by the number of to-be-entered digits (20). In this way, errors were taken only indirectly into account. This time estimate was used to model the time it takes to enter a single digit in the typing task.

Dual-task model

The dual-task model worked as follows. The model started off with typing a series of digits (the length of which was varied as a strategy). The time to type each digit was taken from the single-task model (260 milliseconds). For switching between typing and tracking, a switch cost of 250 milliseconds was incurred, based on experimental data (time between last key press and pressing the trigger on the joystick: 247 milliseconds). The model then tracked the cursor for a designated amount of time (varied between runs as a strategy aspect). When it switched back to typing, a switch cost of 180 milliseconds was incurred (time between releasing the trigger and pressing the first key, corrected for the single task typing time: 185 milliseconds).

Strategies

The basic model described above was used to explore performance of a variety of different dual-task strategies. A strategy was determined by the number of digits that was typed in one sequence during a visit to the target window. Only a subset of twenty simple strategies was considered, that placed a consistent number of digits during each visit to the typing task (between 1 and 20 digits), with the exception of the last visit during which the remaining digits were placed (e.g., strategy 6-track-6-track-6-track-2). In addition, for each visit to the tracking task, more or less time can be spent on tracking. The model was used to systematically explore performance for different tracking times per visit to the tracking window in the range of 250 to 3,000 millisecond, using steps of 250 milliseconds (i.e., 12 alternatives). This gave a total of 229 (19 x 12 + 1) strategy alternatives.

The objective function for rating performance was similar as in the experiment (see equations 4.1 - 4.3) with the exception that the model did not make typing errors. For each strategy alternative, 100 runs were performed. Mean performance is reported.

4.3.2. Model results

The first question of interest was whether the model would fit the experimental data. To do this, a strategy was hardcoded for each condition that typed the same maximum number of correct digits per visit and spent about the same amount of time tracking as participants did. These values were set within two standard errors of the human means, as the model's strategy alternatives were more discrete than the human data (e.g., the model's tracking time was explored in discrete steps of 250 milliseconds). With these values set, an exploration was made to see whether the model's performance fitted the total trial time, maximum deviation, and time outside the target area in each experimental condition observed in the human data. This is important so as to know that a reasonable calibration of the model's performance relative to the human data was achieved. Model performance was within two standard errors of the human data for these variables.

Given that the model was reasonably calibrated to the observed strategies, the model could be used to evaluate the payoff achieved by different (unobserved) dual-task interleaving strategies. Figure 4.6 shows a plot of the maximum number of digits typed per visit to the typing window versus payoff. In this Figure (and Figure 4.7 and 4.8), the performance predictions of the model for each strategy alternative are represented by colored circles. The color of the circle reflects the average payoff that the model gained over 100 simulations when this strategy was applied. The warmer the color, the higher the score (i.e., the higher on the vertical axis in Figure 4.6). The maximum score is £0.16 , and each change in color reflects a change in payoff of £0.02 .

The horizontal axis of Figure 4.6 shows the maximum number of digits that the model typed per visit to the target window. Each of the twenty simple strategies takes a unique value on this axis (e.g., strategy 6-track-6-track-6-track-2 is plotted at value 6 on the horizontal axis). For each of these simple strategies, performance was explored for multiple strategy alternatives based on how much time the model spent on tracking. This resulted in multiple points for each value on the horizontal axis. In three out of the four conditions the human data (black point, with standard error bars) lie within the region of the highest payoff. That is, on average participants typed the optimum

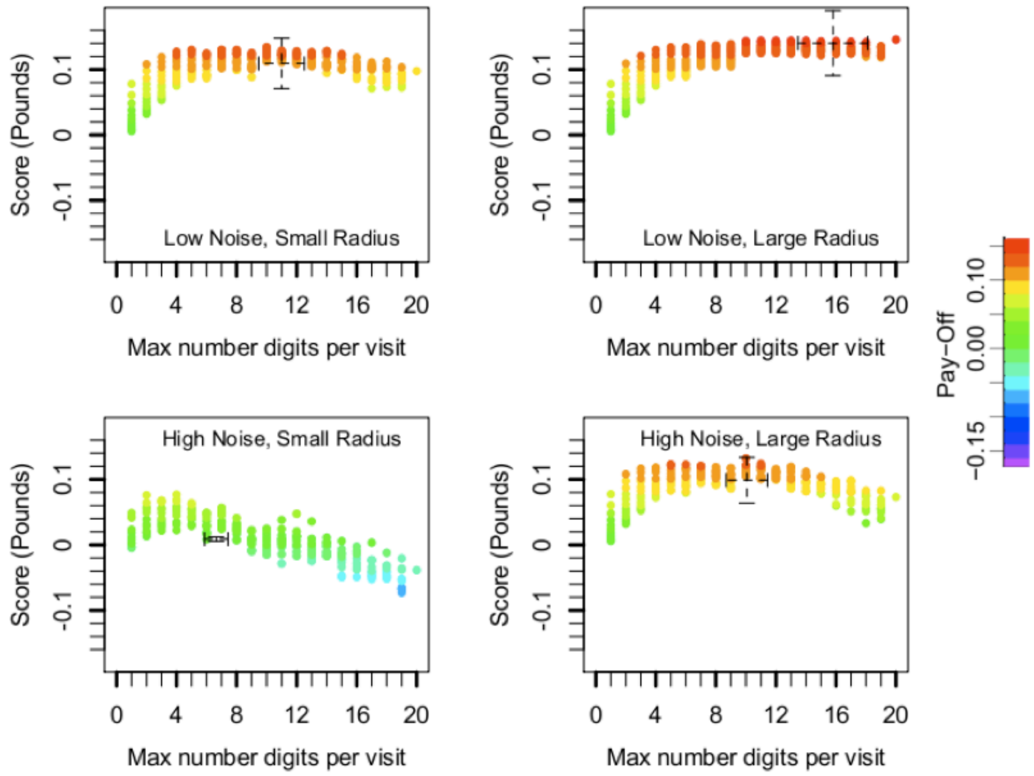

Figure 4.6: Plot of the average maximum number of digits typed per visit to the typing window versus predicted payoff per trial
for the modeled strategies per condition. Color represents the average payoff achieved by the model using that strategy. Human results are shown as black points with standard error.

number of digits per visit to the target window so as to achieve the highest payoff. Note that in the small radius, large noise condition participants did not achieve the highest score – they should have typed less digits per visit to the typing window.

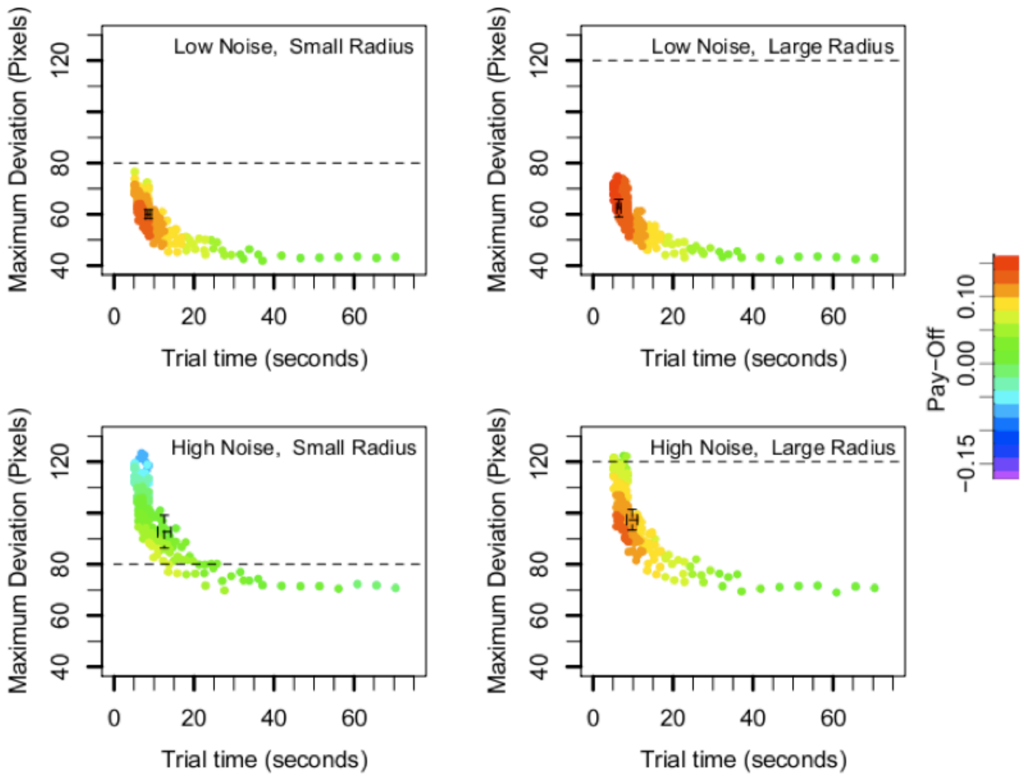

Figure 4.7: POCs of trial time versus maximum deviation for the modeled strategy alternatives per condition.
Color represents the average payoff achieved by the model using that strategy. Human results are shown as black points with standard error. The dashed line shows the target boundary.

The analysis above suggests that participants selected appropriate strategies in each condition. To investigate whether this strategy also resulted in good overall performance, I plotted Performance Operating Characteristics (POCs). Recall that POCs display performance on one task against performance on the other task (Navon & Gopher, 1979; Norman & Bobrow, 1975). Two types of POCs are given here. In Figure 4.7 the POCs are plotted for the total trial time and the maximum deviation of the cursor from the center of the target. In Figure 4.8, the POCs are plotted for the total trial time and the total time the cursor spent outside of the target area per trial. Again the color of the model data represents the average payoff achieved using this strategy.

Four general observations can be made of these POCs:

1. For each condition the shape of the POC differs.
2. The scores that can be achieved differ between conditions, as indicated by different color ranges in each condition.
3. The best performing strategies (i.e., the regions with the warmest colors) tend to cluster on the outer edge (left side, and bottom side) of the strategy space: the trade-off curve. That is, the best strategies make an optimal trade-off for performance on the combination of the two tasks.
4. In addition to point 3, for the current payoff function the optimum region is at different sections of the trade-off curve for some of the conditions. The biggest contrast is in Figure 4.7 between the low noise, large radius condition and the high noise, small radius condition. In the former, the best score is achieved by letting the cursor drift completely (i.e., the best performance is at the top left), whereas in the latter condition the optimum is at the inflection point (i.e., the middle of the curve, on the outside, where it crosses the dashed line). The model is essential for this assessment, as traditional POCs cannot predict optimal regions by themselves. Inherently, the exact location of the optimum region can also shift with a change in the payoff function.

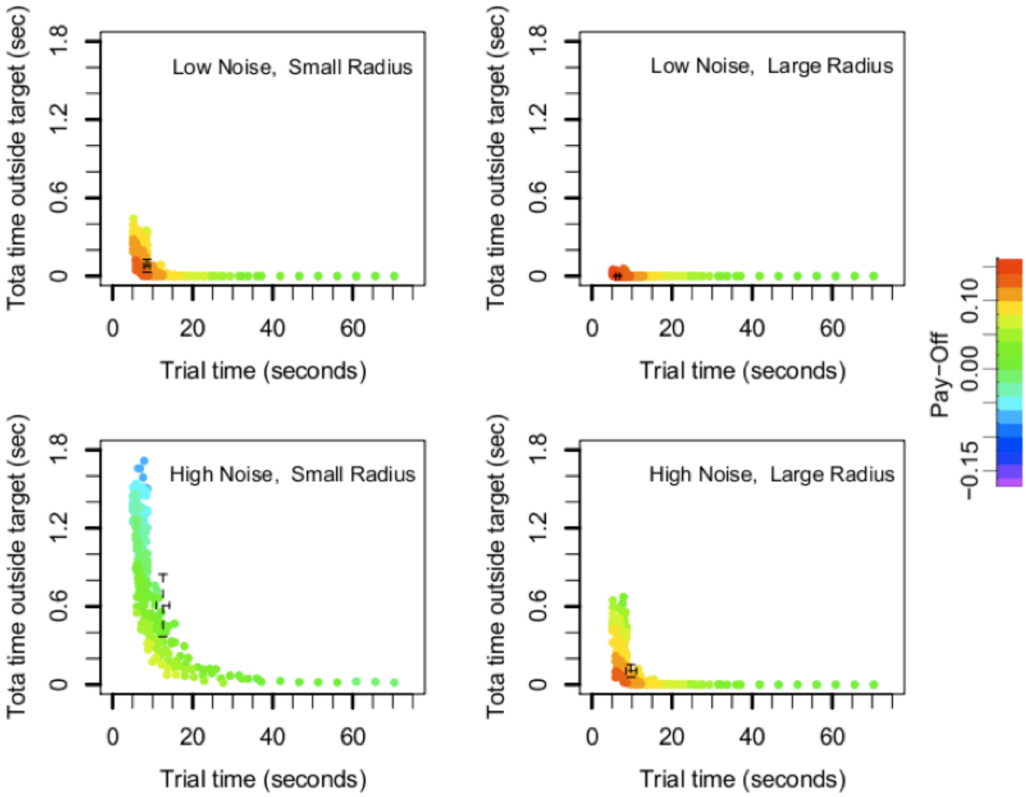

Figure 4.8: POCs of trial time versus time the cursor was outside of the target area for the modeled strategy alternatives per condition.
Color represents the average payoff achieved by the model using that strategy. Human results are shown as black points with standard error.

Leaning on this fourth point, the analysis introduced here helps to bracket *optimal* performance. For each of the measures of total trial time, maximum deviation of the cursor (see Figure 4.7), and time spent outside of the target area (see Figure 4.8), the model predicts that the optimal region lies in a different range of values. This is consistent with the finding in the human data for these measures that showed main effects of radius, or noise, or significant interactions of these factors. Note that this way of bracketing differs from bracketing methodologies that identify the fastest and slowest strategies for performance based on performance time (e.g., Kieras & Meyer, 2000). The model introduced here can be used to bracket performance for the *best* strategy alternatives (and others if necessary), based on the predicted payoff of those strategies. Similar to for example the work by Kieras and Meyer (2000), performance of these strategies can then be expressed in multiple dimensions of performance (e.g., in this case trial time, maximum deviation of the cursor and time spent outside of the target area).

Figure 4.7 suggests that for the different performance measures, human performance was around or at the optimum. In all four conditions, human performance overlapped with the optimum *range* of values for total trial time and maximum deviation of the cursor from the center of the target (see Figure 4.7). Figure 4.8 suggests that for three conditions, human performance also overlapped with the optimum range of values for the total time the cursor was outside of the target area. It did not overlap in the high noise, small radius condition (see the bottom left graph in Figure 4.8).

To close, the correspondence between performance predictions of the best model strategy alternatives and human data can also be assessed using mean performance. Figure 4.9 shows bar plots of mean human performance and mean model performance for three measures of performance. Model data is the mean performance for strategy alternatives that fall in the highest scoring region (i.e., in the region with the warmest color in Figures 4.6 - 4.8). The correspondence between the human and model bars is surprisingly high - error bars between model and human data overlap in most instances. This also reflects in R-squared (R^2) and Root Mean Squared Error (RMSE) values. The fit values are as follows for trial time (R^2 = 0.88, RMSE = 1.63 seconds), maximum deviation (R^2 = 0.95, RMSE = 5.27 pixels), and time

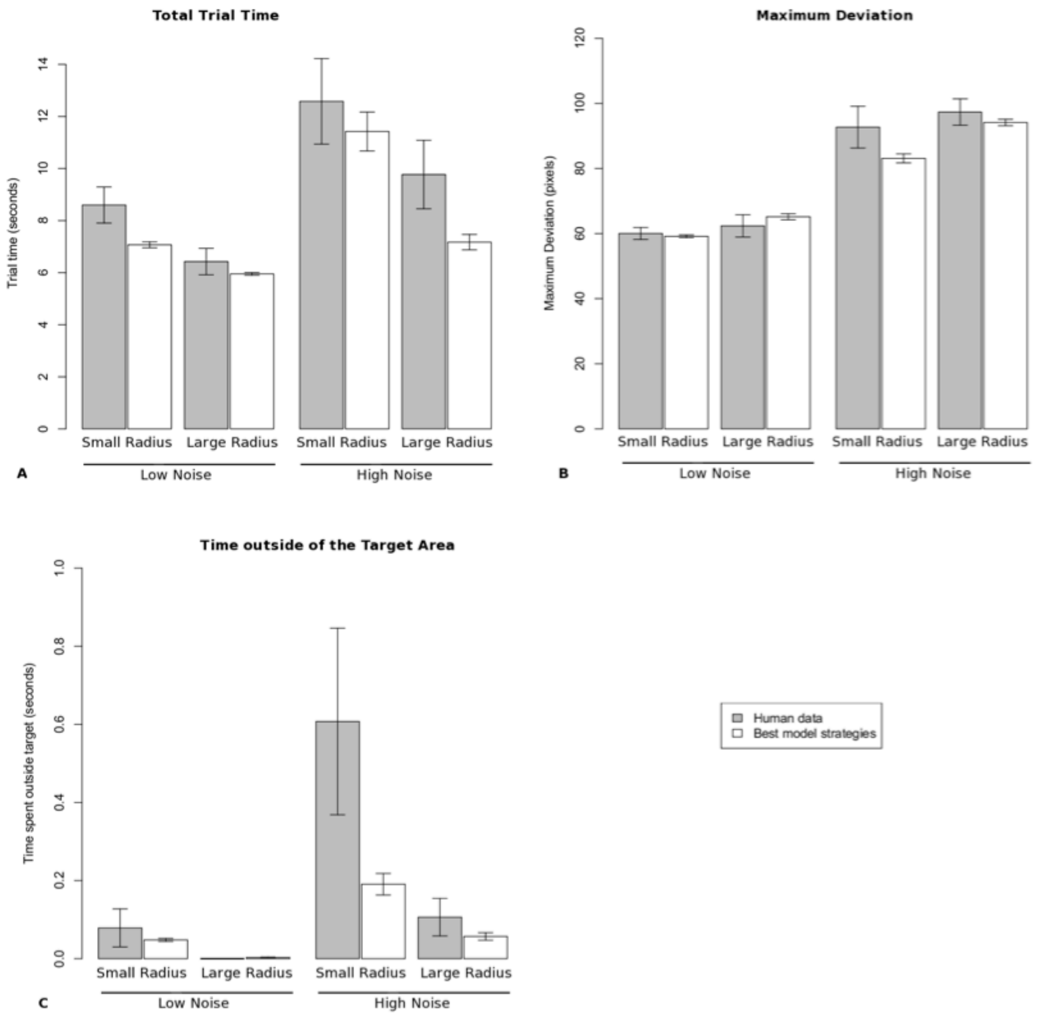

Figure 4.9: Bar plot of human and model performance in each condition.
Model performance is the mean performance for strategy alternatives that fall in the highest scoring region (i.e., in the region with the warmest color in Figure 4.6 - 4.8). Error bars show standard error. The three plots show (A) total trial time (in seconds), (B) maximum deviation of the cursor from the center of the target (in pixels), and (C) total time the cursor was outside of the target area (in seconds).

spent outside of the target area (R^2 = 0.98, RMSE = 0.21 seconds). These values can be considered to be high, given that the model was not fitted to optimize R^2 or RMSE values. Rather, the prediction of the model was based on the set of strategy alternatives that would achieve the highest payoff in each condition.

4.4. Discussion of results

The goal of experiment and model 4A was to get a stronger control on how people prioritize their interleaving of two tasks. To this end, a payoff function was introduced that captures performance on both tasks into a single score. In an experiment, it was investigated how people adapted to different task environment characteristics, given a fixed payoff function. This was followed by development of a model of aggregate human performance. This model was used to explore whether people applied the best possible strategy (i.e., the one that would achieve the highest payoff), given the local task characteristics and the payoff function at hand.

In this task environment, participants selected strategies that had the potential to achieve the maximum payoff in three out of four conditions (see Figure 4.6). In most conditions, participants optimized each of the individual performance measures (total trial time, maximum deviation of the cursor from center, and time spent outside of the target area). Their performance overlaps with the bracketed optimum performance of the model (see Figure 4.7 and 4.8) and the measures of fit between mean human and mean model performance of the best scoring strategy alternatives are high (see Figure 4.9).

The shape of the curves that plot payoff as a function of strategy (see Figure 4.7) is not ideal. This is because for the majority of conditions there is not a well defined global maximum, of which the value is very distinct from the other values. Rather, there seems to be a series of strategies that achieve a relatively similar high payoff value in most conditions (esp. in the low noise, large radius condition). This availability of many strategies that are close to the optimal value might lead to satisficing (Simon, 1956) of performance. In Chapter 5 I will discuss the characteristics of an ideal payoff function in more detail.

The model was developed with a minimal set of assumptions. This was already enough to demonstrate that people can adapt performance reasonably to an objective function in some situations. To test the generality of this hypothesis, participants' adaptation needs to be tested with different payoff functions, which, for instance, give greater weight to performance on one of the two tasks. This will be the objective of experiment 4B.

4.5. Experiment 4B: Adaptation to different pay-off functions

In this experiment, the generality of the findings from experiment 4A will be tested. To this end, the same experimental paradigm will be used, but two slightly different payoff functions will be used in a between-subjects manipulation. More details are provided in the method section, but roughly, the payoff manipulation is as follows. In one condition ("B", "speed"), fast completion of the typing task is highly encouraged (and tracking accuracy is only mildly penalized). In another condition ("C", "accuracy"), tracking accuracy is encouraged by heavily penalizing situations where the cursor moves outside of the target area; at the same time, typing faster or slower only makes a minor difference on achieved payoff. Based on the results of experiment 4A, the hypothesis is again that people will adapt to the payoff function at hand, and select strategies that can achieve an optimal score, as predicted by a model.

4.5.1. Method

Participants

Twenty-four students (9 female) from the UCL participant pool took part for monetary compensation. Participants were between 18 and 46 years (M = 24.3, SD = 6.6 years). Payment was based on performance (details are provided in the Design section). Total payment ranged between £5.00 and £13.03 (*M* = £8.72).

Materials

The set-up was identical to the set-up in experiment 4A.

Design

The experiment followed a 2 x 2 x 2 mixed factorial design. Within subjects, noise and radius levels were manipulated as in experiment 4A. Between subjects, the payoff function was manipulated with 2 levels. In both cases, the structure of the payoff functions followed the same equations as in experiment 4A (see equations 4.1-4.3). However, the parameters of the equations were manipulated between subjects to encourage different styles of interleaving. The exact parameter values are given in Table 4.1, and the gain and loss components are illustrated in Figure 4.10.

Descriptive, the manipulation was as follows. In payoff condition B ("speed"), fast completion of the typing task was encouraged as the gains to be won decayed rapidly as time progressed. At the same time, only a mild penalty was applied when the cursor moved outside of the target area (see Figure 4.10). In payoff condition C ("accuracy"), accuracy on the tracking task was encouraged through a severe penalty on the tracking task, and relatively little loss in gain was given when typing time progressed (see Figure 4.10). Participants were randomly assigned to a payoff condition.

Table 4.1: Parameter values for the Payoff functions

	Payoff function		
	A	**B ('speed')**	**C ('accuracy')**
severityOfTrialTime	-1	-4.6209812	-0.0854888
StartValue$_{gain}$	0.25	1.1552453	0.0170978
compensation	0	0.02294	0
severityOfBeingOutside	1.1090	1.1090	2.2180
startValue$_{tracking}$	0.6931	1.5	0.6931

Procedure

The procedure was the same as reported in the preceding experiment. As before, a participant's payment was based on the cumulative payoff over all dual-task trials. In payoff condition C a base payment of £3 was given. It was observed that participants in condition B ("encourage fast completion of the typing task") earned less than other participants. To compensate for this, the base payment for these participants was changed from £3 to £5. This guaranteed a minimum payment of £5 for all participants.

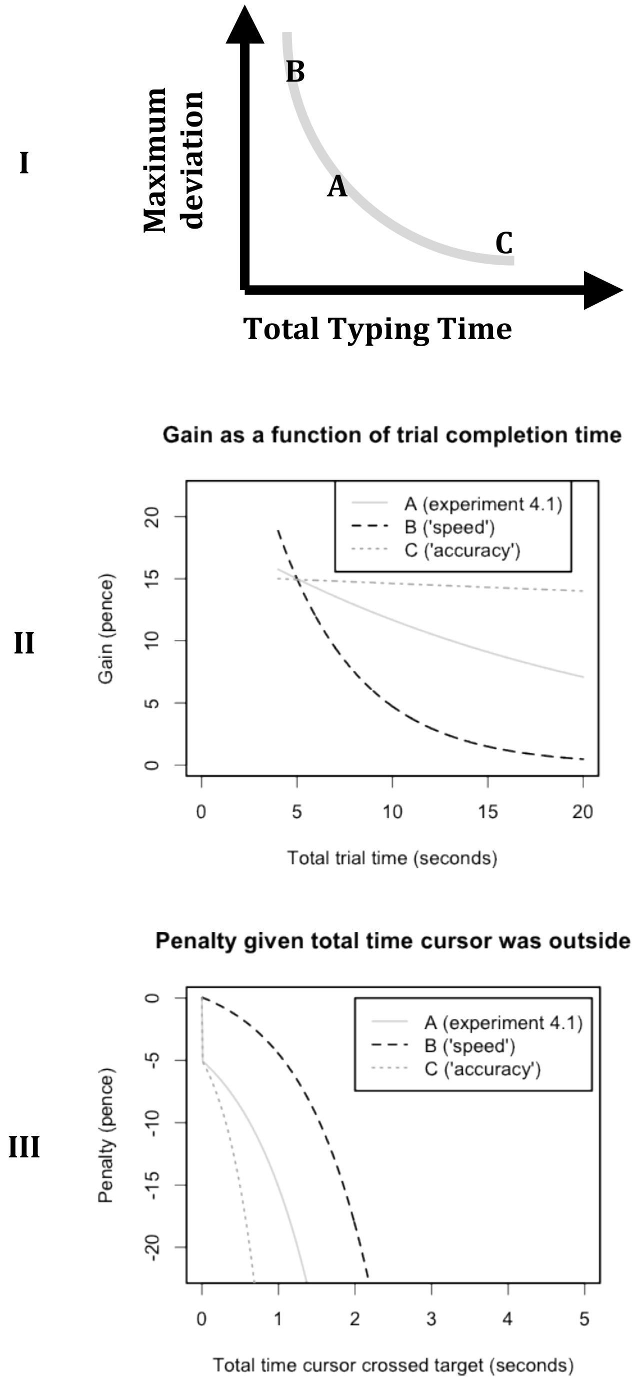

Figure 4.10: Representation of the three payoff functions used in experiments 4A ("A") and 4B ("B" and "C").
Plot I shows a schematic positioning of where behavior is roughly expected on the trade-off curve between total trial time and maximum deviation of the cursor. Plot II shows the gain component of the three payoff functions, as a function of trial time. Plot III shows the tracking penalty of the three payoff functions, as a function of time the cursor was outside of the target area. Note that total payoff could never reach values lower than -20 pence.

In all conditions, participants were told that they could gain points by completing the typing task faster. They were also instructed that they lost points when the cursor went outside of the target area, and for each typing error they made. Importantly, participants did not know the exact equation of the payoff function, and the relative impact of each component (gains and penalties) on performance. Rather, they had to learn the impact of each component based on the feedback given by the payoff function (a single score) at the end of each trial.

4.5.2. Results

Across all keystrokes in single-task typing trials, participants on average typed 2.92% (range 0.64 – 10.52 %) of their keystrokes incorrectly (343 out of 11,498 keystrokes). An independent-samples t-test was conducted to compare the percentage of errors between the two sets of participants. There was no significant difference in percentage between participants in payoff condition B ('speed', M = 2.80% , SD = 2.18%) and those in payoff condition C ('accuracy', M = 3.04% , SD = 2.82%), $t(22) = -0.24$, $p = .81$. At the trial level, 0, 1, 2, or more than 2 errors were made on respectively 60.42%, 23.33%, 9.58%, and 6.67% of the trials.

In the dual-task trials, the number of typing errors was also low. Participants on average typed 0.28% (range 0.04 – 1.12 %) of their keystrokes incorrectly (1,745 out of 744,315 keystrokes). Again, an independent-samples t-test was conducted to compare the percentage of errors between the two sets of participants. There was no significant difference between participants in payoff condition B ('speed', M = 0.25%, SD = 0.20%) and those in payoff condition C ('accuracy', M =

0.31 %, SD = 0.29%) , $t(22) = -0.54$, $p = .59$. At the trial level, 0, 1, 2, or more than 2 errors were made on respectively 49.32%, 28.13%, 12.97% and 9.58% of the trials. In the later model development, error occurrence will be taken into account.

In the remainder of this analysis, I again focus on performance during the last five dual-task trials of each experimental condition, as these reflect a period during which the participant had had time to adapt their behavior to the payoff function, based on the feedback received. While running the experiment, individual differences in typing speed were observed (see also Table A.4.1 and A.4.2: the spread of interkeypress interval times is wide, and was wider than in experiment 4A). It was expected that this could significantly influence performance. For example, a slow typist might take longer to complete a trial. This might mask effects of the payoff manipulation which (for condition B) was intended to encourage participants to complete the typing task fast.

To test whether interkeypress interval contributed to performance, I systematically compared performance of different linear regression models. The basic ANOVA had a 2 (payoff function) x 2 (cursor noise) x 2 (target size) form. The addition of mean interkeypress interval (as measured in single-task trials) as an additive factor in the ANOVA provided significant explanatory power for all measures except mean visit time to the typing window (there the additive effect of interkeypress interval did not provide significant explanatory power). Further addition of interactions between interkeypress interval and the other independent variables also led to significant improvement of the explanatory power for total trial time and total time the cursor was outside of the target window. For consistency, this later model was used for all analyses.[13,14] A significance level of .05 was applied throughout. Only main effects and two-way interactions were considered.

Overall performance

Figure 4.11 plots the performance space of total trial time versus the maximum distance that the cursor moved away from the center of the target in one plot for all eight conditions. It can be seen that the majority of the eight conditions roughly take up a unique point in this performance space.

In general, the cursor deviated more when the payoff function put less value on performance in the tracking task (payoff function B 'speed'; black points) compared to when it put strong value on

---

[13] This analysis was performed in the R statistical language using the LME4 package. The comparisons were made as follows. Three regression models were used to estimate the effects of different sets of variables on a dependent variable y. Model0 was the basic model without interkeypress interval:
model0 = y~Payoff x Radius x Noise
Model 1 contained interkeypress as an additive factor ("+"):
model1 = y~Payoff x Radius x Noise + Interkeypress
Model2 contained interkeypress as an interacting factor ("x"):
model2 = y~Payoff x Radius x Noise x Interkeypress
All three models also contained a random effect components for participants.
If a comparison of model0 and model1 (code: "anova(model0,model1)") resulted in significant effects, the addition of interkeypress interval lead to significant explanatory power. In these cases a further comparison was made between model2 and model1, to see if adding the interactions was also valuable (code: "anova(model1,model2)").
Note that I did not consider 3-way and 4-way interactions in the regression models.

[14] The appendix of this Chapter contains a table (Table 4.A.3) that reports statistical effects for a model that does not include interkeypress interval as a factor. In this model, payoff function did not have a significant effect on 3 dependent variables (total trial time, mean visit time in the tracking window, and total time the cursor spent outside of the target window). It might be that the effect that interkeypress interval had on performance masked the manipulation of payoff. For example, slow typists in payoff condition B had a relatively longer trial time, which might cancel out the intended manipulation of achieving a faster trial time in this condition.

performance in the tracking task (payoff function C ‘accuracy’; grey points). In line with findings in experiment 4A, the cursor also deviated more when the noise was high (Figure 4.11: squares) compared to low (circles), and when the radius was large (open points) compared to small (closed points). Similar main effects occurred for trial time: participants were faster in payoff condition B (‘speed’; black), when noise was low (circles), or when the radius was large (open points).

Statistical analysis confirmed these findings. The effects are summarized in Table 4.2, and discussed in more detail below. Interestingly, all measures of overall performance had significant main effects of the three constraints of interest (see also Chapter 2): payoff, task characteristics (noise and radius), and individual differences in skill (interkeypress interval).

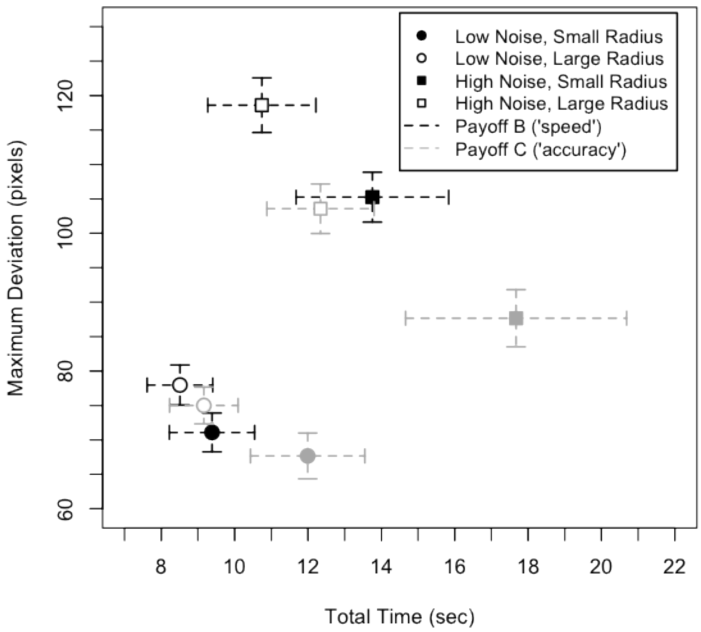

Figure 4.11: Plot of the performance space for all eight conditions.
Total time against maximum deviation of the cursor. Error bars show standard errors.

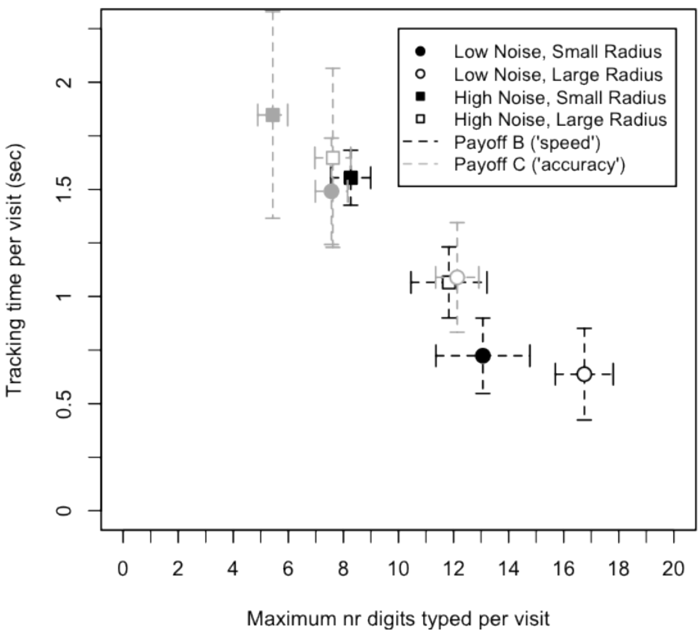

Figure 4.12: Plot of the maximum number of digits typed and time spent tracking.
Both are per visit, for all eight conditions. Error bars depict standard errors.

Table 4.2: Summary of statistical effects in experiment 4B.

		Dependent variable					
		Total trial time	Maximum cursor deviation	Total time cursor outside target	Maximum nr of digits per visit	Mean visit time, typing window	Mean visit time, tracking window
Independent variable	**Payoff function (P)**	**	***	*	***	***	**
	Noise (N)	***	***	***	***	***	**
	Radius (R)	***	***	***	***	***	
	Interkeypr. interval (I)	***	***	***	***		***
	P x N		***	*		*	
	P x R						
	N x R		.	**			
	P x I				***		
	N x I	**		*			
	R x I	*		**		.	

.: .05 < p <= .10;
*: .01 < p < .05;
**: .001 < p < .01;
***: p <= .001

For trial time, the trial time was longer in payoff condition C ("accuracy"; M = 12.79 sec; SD = 5.58 sec) compared to payoff condition B ("speed"; M = 10.60 sec, SD = 4.59 sec), $F(1, 85) = 9.45$, $p = .003$. That is, people were faster when the payoff condition encouraged them to be faster. As expected, given the results of experiment 4A, trial time was affected by the difficulty of the task. Specifically, the trial time was longer when there was high noise (M = 13.62 sec, SD = 6.71 sec) compared to low noise (M = 9.76 sec, SD = 3.89 sec), $F(1, 85) = 29.33$, $p < .001$. Similarly, trial time was longer when the radius of the target was small (M = 13.20 sec, SD = 6.54 sec) compared to large (M = 10.19 sec, SD = 3.99 sec), $F(1, 85) = 17.80$, $p < .001$. That is, people were slower when the task conditions were more difficult. There was also a main effect of interkeypress interval, $F(1, 85) = 17.81$, $p < .001$. The corresponding linear model predicted that for every 0.100 second increase in interkeypress interval time, the total trial time increased with 4.88 seconds. That is, faster typists completed the task faster. Finally, interkeypress interval interacted with noise, $F(1, 85) = 11.31$, $p = .001$, such that for every 0.100 seconds increase in interkeypress interval, trial time increased more in the high noise condition (slope of 6.24 sec) compared to the low noise condition (slope of 3.52 sec). Similarly, interkeypress interval interacted with radius, $F(1, 85) = 6.83$, $p = .010$, such that trial time increased more with an increase of 0.100 seconds in interkeypress interval in the small radius condition (slope of 5.95 sec) compared to the large radius condition (slope of 3.81 sec). There were no other interaction effects.

For maximum deviation, the cursor deviated more in payoff condition B ('speed', M = 93.23 pixels, SD = 7.84 pixels) compared to payoff condition C ('accuracy', M = 83.48 pixels, SD = 8.77), $F(1, 85)$ = 19.74, $p < .001$. That is, when the penalty for being outside of the target area was harsher (condition C), the cursor was kept closer to center. Not surprisingly, the cursor also deviated more in the high noise condition (M = 103.77 pixels, SD = 13.76 pixels) compared to the low noise condition (M = 72.93 pixels, SD = 8.21 pixels), $F(1, 85)$ = 197.48, $p < .001$. The cursor also deviated more in the large radius condition (M = 103.77 pixels, SD = 10.53 pixels) compared to the small radius condition (M = 72.93 pixels, SD = 11.00 pixels), $F(1, 85)$ = 24.52, $p < .001$. The last main effect was of interkeypress interval, $F(1, 85)$ = 19.64, $p < .001$, such that with every 0.100 sec increase in interkeypress interval, the cursor deviated 5.31 pixels further. That is, the more slowly a participant typed, the more the cursor deviated (as per digit typed the participant was longer away from the tracking task). There was one significant interaction effect, of payoff function and noise, $F(1, 85)$ = 8.94, p = .004. There was a marginal interaction effect between payoff function and radius, $F(1, 85)$ = 2.94, p = .09.

For total time that the cursor spent outside of the target area, there was again a main effect of payoff function, $F(1, 85)$ = 5.27, p = .024, such that the cursor was longer outside of the target area in payoff condition B (M = 0.57 sec, SD = 0.34 sec), compared to payoff condition C (M = 0.33 sec, SD = 0.32 sec). Again, this reflected the payoff function's nature of penalizing time outside more in condition C than condition B. The cursor was also longer outside of the target area in the high noise condition (M = 0.74 sec, SD = 0.63 sec) compared to the low noise condition (M = 0.16 sec, SD = 0.16), $F(1, 85)$ = 31.71, $p < .001$. The cursor was also longer outside of the target area when the radius was small (M = 0.71 sec, SD = 0.58 sec) compared to when it was large (M = 0.19 sec, SD = 0.19 sec), $F(1, 85)$ = 25.93, $p < .001$. There was also a main effect of interkeypress interval, $F(1, 85) = 11.70$, $p < .001$, such that for every 0.100 sec increase in interkeypress interval the cursor was 0.19 sec longer outside of the target area. There were also four significant interaction effects, between payoff function and noise, $F(1, 85) = 5.53$, $p = .020$; between noise and radius, $F(1, 85) = 7.79$, $p = .006$; between noise and interkeypress interval, $F(1, 85) = 4.88$, $p = 4.88$, $p = .030$; and between radius and interkeypress interval, $F(1, 85) = 10.22$, $p = .002$.

These results can be summarized as follows. First, in general, most of the effects of the preceding experiment were replicated. That is, performance depended on the task characteristics (noise, radius), and performance suffered when the task was harder (due to a small radius or high noise). Second, participants' performance depended on the payoff condition. In payoff condition B, where fast completion of the typing task was highly rewarded and tracking penalties were not severely penalized, participants completed the typing task faster, by letting the cursor drift for longer, leading to a larger distance from the center of the target area. Finally, there were individual differences in performance, as indicated by the significant effects of interkeypress interval time on all performance measures. Participants that in general took longer to type a single digit (i.e., that had a larger interkeypress interval time) also took longer to complete a trial, had the cursor deviate further away from center (as the cursor had more time to deviate while a digit was being typed), which lead to longer times during which the cursor was outside of the target area.

Strategies

How were these differences in performance achieved? To this end, the strategies that people used for interleaving between the two tasks were investigated, focusing on two aspects: (1) the maximum number of digits typed during a visit to the typing window and (2) the amount of time spent in the tracking window per visit to this window. Figure 4.12 shows these two basic strategy dimensions for all eight conditions. For the number of digits typed only correctly typed digits were considered. It can be seen that for each experimental condition there again is a relatively unique point in this strategy space, especially when comparing strategies between the two payoff conditions (i.e., comparing the black points with their grey equivalents in Figure 4.12). A summary of statistical effects is again given in Table 4.2, and discussed in more detail below.

For the maximum number of digits typed per visit to the typing window, more digits were typed in payoff condition B (‘speed’, M = 12.48 digits, SD = 3.67 digits) compared to payoff condition C (‘accuracy’, M = 8.19 digits, SD = 1.33 digits), $F(1, 85) = 60.92$, $p < .001$. That is, more digits were typed when the payoff condition encouraged fast completion of the typing task (payoff B). More digits were typed when noise was low (M = 12.38 digits, SD = 4.03 digits), compared to when noise was high (M = 8.29 digits, SD = 3.25 digits), $F(1, 85) = 55.37$, $p < .001$. More digits were also typed when the radius was large (M = 12.08 digits, SD = 3.53 digits) compared to when the radius was small (M = 8.58 digits, SD = 3.69 digits), $F(1, 85) = 40.52$, $p < .001$. . Taken together, these results confirm that more digits were typed per visit to the typing window when the task environment conditions were easier (i.e., low noise, large radius). There was also a main effect of interkeypress interval, $F(1, 85) = 49.14$, $p < .001$. For every 0.100 sec increase in interkeypress interval, the maximum number of digits typed per visit decreased with 2.24 digits. That is, slow typists typed fewer digits per visit to the typing window. There was also a significant interaction effect between payoff function and interkeypress interval, $F(1, 85) = 14.57$, $p < .001$. There were no other significant interactions.

It was expected that the task environment (implicitly) posed a maximum time that people could spend in the typing window before the cursor would move outside of the target area. This time should be independent of people's typing speed, as it was completely determined by the tracking task (i.e., by how quickly the cursor on average would cross the target area boundary, given the noise of the cursor movement, and the size of the radius). To test this, an ANOVA was performed on the mean time spent in the typing window per visit. As can be seen in Table 4.2 there were again main effects of task constraints (noise, radius) and of the payoff function. However, as hypothesized, there was no effect of interkeypress interval. That is, all participants spent roughly the same amount of time in the typing window, given the specific task condition. Within this roughly similar time frame, different participants could type a different number of digits.

Were similar effects present in the time spent in the tracking window per visit? As Table 4.2 shows, there were indeed again effects of task conditions, the payoff function, and individual differences in interkeypress interval. More time was spent in the tracking window in payoff condition C ('accuracy', $M = 1.52$ sec, $SD = 1.19$ sec) than in payoff condition B ('speed'; $M = 1.00$ sec, $SD = 0.48$ sec), $F(1, 85) = 7.49$, $p = .008$. More time was also spent in the tracking window when noise was high ($M = 1.53$ sec, $SD = 1.14$ sec) compared to when noise was low ($M = 0.99$ sec, $SD = 0.80$ sec), $F(1, 85) = 8.07$, $p = .006$. Surprisingly, the time spent in the tracking window per visit also changed with interkeypress interval, $F(1, 85) = 13.91$, $p < .001$, such that with every 0.100 sec increase in interkeypress interval time, the time spent tracking increased by 0.41 seconds. This extra time was probably needed because the cursor drifted more per digit typed for higher interkeypress interval values. There was no significant main effect of radius, and there were no significant interactions.

4.5.3. Discussion of results

The statistical analysis demonstrated that participants' performance adapted to task constraints (noise, radius), individual differences in skill (interkeypress interval), and the payoff function. These differences arose, because people applied different strategies in each condition. For example, participants that were relatively slow typists typed fewer digits per visit to the typing window. As a result, they had to interleave more often between the typing and the tracking task, which resulted in longer trial times. Similarly, participants typed fewer digits in payoff condition C, to avoid the relatively strong penalty for the tracking task. Again, this required more visits to the typing window, and resulted in a longer trial time, but also resulted in a less severe deviation of the cursor, and a shorter total time during which the cursor was outside of the target area.

What these data do not reveal is whether participants adopted strategies that would maximize their payoff, as was predicted based on results in the preceding experiment. To this end, I again developed computational cognitive models to explore the performance and strategy space. As individual differences in performance arose in the experiment, a model of aggregate performance would not be sufficient to explain performance, as it has no means to capture these differences. The models were therefore adapted to capture some of the individual differences in skill (e.g., as measured in single-task), to see what the effect of these differences in skill was on dual-task performance.

4.6. Model 4B

4.6.1. Model development

To capture the individual differences in performance, it was suspected that the model had to be changed. To this end, components of the model were explored that could account for these individual differences. Below I will describe the components that could be added. This will be followed by an analysis that investigates the fit of the model in relationship to the added complexity of the new parameters and calibration sets. All models had a component for tracking, typing, and dual-tasking.

Typing model

The basic typing model was the same as before, with the time needed to type a digit set to the mean interkeypress interval time as measured in single-task trials. In addition, one optional refinement was to include typing errors. The model of typing errors was kept simple, as a detailed understanding of typing errors was not the focus of this thesis.

The mean number of errors that the model made per trial was set to the observed value in single-task trials. The value was calculated up to one decimal accurate. Within a single simulation run, the model could only make an exact (integer) number of errors. To achieve the exact mean value, the model sometimes made fewer errors than the mean human value, and at other times it made more errors than the mean human value, in such a way that the mean across simulations was similar to the mean in the human data. For example, take a participant that made an average of 2.3 errors per trial. In this case, the model made 2 typing error on 70% of the simulations, and 3 errors on 30% of the simulations, giving an average of 0.7 * 2 + 0.3 * 3 = 2.3 errors per trial. The position of the error within the string of digits was determined at random at the start of every simulation run.

It was assumed that the time needed to type an incorrect digit was similar to the time needed to type a correct digit. The model always typed the correct digit after making an error. It was assumed that this correct digit suffered from a post-error slowing effect (Rabbitt, 1966) (for more recent and more detailed theories on post-error slowing see e.g., Castellar, Kühn, Fias, & Notebaert, 2010; Danielmeier & Ullsperger, 2011; Notebaert et al., 2009). That is, it was assumed that more time was taken than usual for typing the next correct digit. This was consistent with observed performance by the majority of participants (see Tables 4.A.1 and 4.A.2 for post-error slowing values). The mean post-error slowing time was estimated by subtracting the normal interkeypress interval time (from the typing model) from the average time observed in the interval for the first correct digit after an error. [15]

---

[15] For some participants this resulted in a negative value (i.e., a speed increase). These values were kept as is, to better match model results with human results, despite that this was not a traditional "slowing" effect. Again, the focus here was to incorporate a basic model of errors to improve this model compared to model 4A. The focus was not to develop a detailed theory of typing errors.

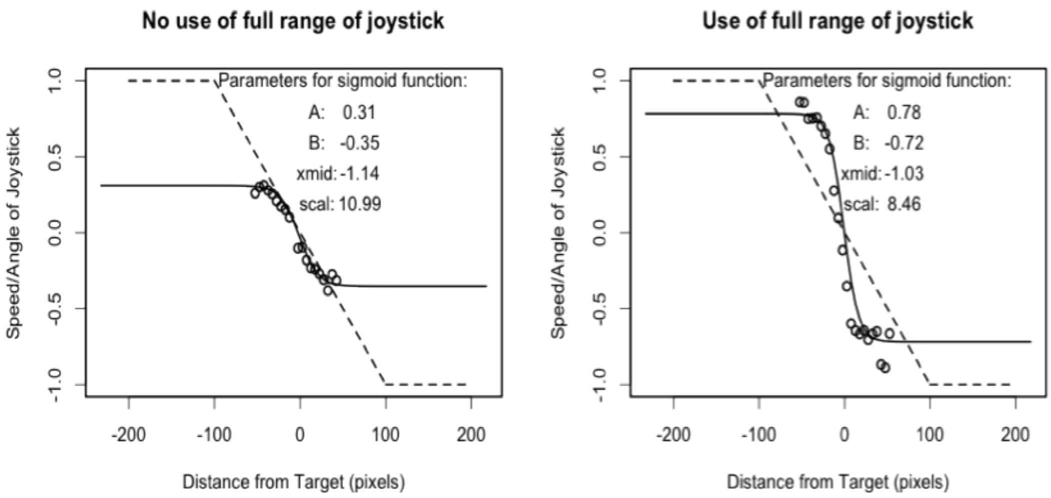

Figure 4.13: Binned data for the angle of the joystick as a function of distance from the target (dots)
for two participants who do not (left plot) or do (right) make use of the full range of the joystick angle. The model shows the line of the linear steering model, as in the initial dual-task model (dashed), and a fitted line for the sigmoid function (solid). Note how the solid lines overlap with more data points than the dashed lines. Parameter values of the fitted sigmoids are given in the Figure. A and B are values for the asymptotes, xmid is the inflection point and scal determines the steepness of the middle part of the curve.

Tracking model

As in the experiment, the cursor was not actively moved when the model was typing or switching attention. During these time intervals, movement followed the same random functions as used in the experiment.

For active tracking, two different models were explored. First, the tracking model from model 4A was explored, in which the angle of the joystick was calculated as a linear function based on the distance of the cursor from the center of the target area (see Figure 4.5 and Equation 4.4).

Given that this tracking model had three components (a slope, and two asymptotes), it was suspected that perhaps a Sigmoid function might better fit the data, as it captures the transition between the slope and the asymptotes more smoothly. The solid line in Figure 4.13 shows two examples of fitted Sigmoid functions for two participants. Formally, Sigmoids are characterized by four variables: two asymptotes for the maximum positive and negative angles of the joystick, an inflection point at which the angle has a value midway between the two asymptotes, and a scaling function for the steepness of the line (see Figure 4.13 for two examples). For each participant, the values of these parameters were derived using automated algorithms (SSfpl, and non-linear least squares function) in the R statistical language (www.r-project.org). These functions automatically determined the best fitting parameter values given a dataset. For each participant I provided one data set, which contained pairs of values for the average angle of the joystick given the distance from the center of the target.[16]

Dual-task model

The dual-task model incorporated two switch costs. One for switching attention from the typing task to the tracking task, and one for switching attention from the tracking task to the typing task. The switch costs were calculated in a similar manner as in model 4A (but see below for calibration aspects).

The dual-task model worked as follows. The model started off by typing a series of digits (the length of which was varied as a strategy,

---

[16] In other work we have used a model in which the moment-to-moment tracking movement was not estimated (Farmer, 2010; Farmer et al., 2011; Janssen, Brumby, & Howes, 2012). Rather, those models estimated the total tracking time and the end position after tracking. These models were not explored here, because they introduced a wider set of model parameters without providing additional insight into human performance.

see later). The time to type each digit, and the number of errors per trial was taken from the single-task typing model. For switching between typing and tracking, a switch cost was incurred.

The model then pursued active tracking of the cursor, based on the tracking model (another aspect of a strategy alternative, see below). After this time had passed, another switch cost was incurred.[17] The model then continued typing in digits, and followed the same structure as above until all twenty digits were typed in.

Summary of model manipulations and possible calibrations

To summarize, the general structure of the model was kept similar to that of model 4A. For the typing model an optional model for typing errors could be added. For the active tracking model either a linear or a sigmoid function could be used. Each model component (typing, tracking, dual-tasking) could be calibrated to (1) mean performance in experiment 4A (i.e., zero parameter), (2) mean performance in experiment 4B, or (3) performance of each individual.

In a series of models I investigated systematically how adding more complexity to the model improved the fit of the model. I only explored a subset of models, as summarized in Table 4.3. Starting point was the original model for experiment 4A (labeled model 1 in Table 4.3). New components were first systematically added to this model (models 2, 3). Then the effect of calibration was explored (models 4-8). Further calibration of the tracking and dual-task model was only done

---

[17] For some participants the tracking to typing switch cost had a negative value. That is, these participants typed the first digit of a series faster after tracking than they would do normally. This might have multiple causes, such as a rehearsal of this digit before starting the tracking, or mentally preparing to continue to type while still tracking. The human data was not fine-grained enough to give insights into which effect was present (e.g., eye-tracking data was not collected). Therefore, the model of individual switch costs was not refined to give a detailed account for this.

when the typing model was already calibrated to individual performance. It was suspected that a calibration of the typing model was the bare minimum that was required to explain individual differences in performance. This was suspected given the significant effects of typing speed on performance as found in the experiment.

Table 4.3: Summary of model manipulations.
The columns describe (1) the models number, (2) a description, (3) whether the model contained a component for typing errors, (4) which experiment dataset the typing model was calibrated to, (5) whether the tracking model used a linear or sigmoid function, (6) what dataset tracking and dual-task performance was calibrated to, and (7) a numerical estimate of the model's complexity.

Nr	Model description	Typing errors made?	Typing model calibration	Tracking model	Tracking & dual-task calibration	Com-plexity
1	Original (4A)	No	Mean 4A	Linear	Mean 4A	0
2	Original + typing errors	Yes	Mean 4A	Linear	Mean 4A	1
3	Sigmoid + errors	Yes	Mean 4A	Sigmoid	Mean 4A	2
4	Typing as in mean 4B, linear tracking	Yes	Mean 4B	Linear	Mean 4A	2
5	Individual typing, linear mean tracking 4A	Yes	Individuals 4B	Linear	Mean 4A	3
6	Individual typing, sigmoid mean tracking 4A	Yes	Individuals 4B	Sigmoid	Mean 4A	4
7	Individual typing, sigmoid mean tracking 4B	Yes	Individuals 4B	Sigmoid	Mean 4B	5
8	Fully calibrated	Yes	Individuals 4B	Sigmoid	Individual 4B	6

Table 4.3 also gives an estimate of what I call the 'complexity' of the model. Complexity was estimated as follows. For each component that was added, and for each further level of calibration, a number was assigned ranking the 'complexity' of that choice.

If the model was similar to the original (model 4A, labeled model 1 in Table 4.3), a complexity score of zero was assigned. For each additional layer of complexity, one point was added. For example, calibration to experiment 4A gave no complexity points, calibration to mean performance in experiment 4B gave one complexity point, and calibration to individuals in experiment 4B gave two complexity points. Note that the calibration of the typing model was ranked separately from the calibration of the tracking and dual-task model. In general, models of lower complexity should be preferred, as they are not calibrated to the dataset at hand.

Strategies

With these models, the strategies for interleaving were manipulated the same way as in model 4A. A strategy was determined by the number of digits that were typed in one sequence during a visit to the target window (ranging between 1 and 20).

For each strategy, multiple strategy alternatives were explored based on how much time was spent in the tracking window per visit (ranging between 250 and 3,000 milliseconds, in steps of 250 milliseconds). As in model 4A, a simple set of 229 strategy alternatives was explored, in which a consistent number of digits was typed and a consistent amount of time was spent on tracking per visit within a simulation (however, these values were alternated between simulations, see model 4A for details).

For all models, 50 simulations were run for each experimental condition (noise, radius, payoff), and for each strategy alternative. For models that were not calibrated to individual performance this gave 50 x 2 x 2 x 2 x 229 = 91,600 simulations per model. If the model was in any way calibrated to individual performance, this method was applied for every individual. This gave a total of 12 (participants) x 2 (payoff functions) x 50 x 2 x 2 x 229 = 1,099,200 simulations per model.

The objective function for rating performance was identical to the one used in the experiment (see Equations 4.1 - 4.3). Note that this time (in contrast to model 4A), typing errors were also taken into consideration when calculating the payoff function.

4.6.2. Model Results

Assessment of fit

The first task was to select the model that most accurately described the human data. For each of the models in Table 4.3, I compared observed mean performance of each individual in each condition (96 datapoints) with performance of the model when its strategy was fixed to these observed strategies. A strategy alternative had two components: the number of digits typed per visit and the time spent in the tracking window per visit. For number of digits typed, the observed mean value was rounded to the nearest integer. For tracking time the value was rounded to the nearest multiple of 250 milliseconds.

Eight human data points were not considered for this analysis, as they did not fit in this framework for one of the following reasons: (1) more than 3,000 milliseconds was spent on tracking, (2) some tracking occurred despite that all 20 digits were typed in one visit, or (3) no tracking occurred despite that less than 20 digits were typed in one visit. The first situation applied to one participant in all four conditions. The other two situations applied to different participants and most likely were a result of averaging performance across trials and then rounding performance to fit the framework of the model.

For each model, the fit was assessed on four performance metrics: trial time, maximum deviation of the cursor, total time the cursor was outside of the target area, and the payoff score. Five measures of fit were used for each performance metric. First and second, the percentage of model data points that fell within one standard error or two standard errors of the mean performance by the corresponding individual was calculated. In general, higher percentages are better. Note that a score of 100% was not expected, given that some participants performed very consistently, and therefore had small standard error bars (with a value of 0 for some measures).

Third, the R-squared (R2) value was calculated. A higher value indicated that the model was better in capturing the variance in the data. Fourth, the Root Mean Squared Error (RMSE) values were calculated for each measure. To allow for averaging across dependent variables, RMSEs were expressed as a percentage of the mean human score on that measure. Lower RMSE percentages are preferred.

Finally, a plot of human performance as a function of model performance was made, and a linear regression line was fitted through the data. Ideally, model data should be identical to the human data, with a slope of $y = x$. To compare slopes across measures, I investigated the absolute difference between the slope value and 1. For example, if two slopes had values 0.74 and 1.20, the average absolute difference of the slope would be 0.23 (the average of 0.26 and 0.20). In contrast, if not the absolute value, but the observed value would be taken, the average slope would be an unrepresentative value: 0.97 (i.e., with an absolute difference of only 0.03).

For each model, I calculated the average fit value across the four performance metrics, as reported in Table 4.4. To compare the fit across different measures of fit, the scores for each metric were ranked between models, from best fitting (scoring "1") to the worst fitting model (scoring "8"). I then calculated the mean rank of each model across measures (see column "mean rank" in Table 4.4).

Table 4.4: Comparison of model performance on several measures.
These measures produce a ranking of the overall goodness of fit, which should be balanced with the model complexity.

Nr	Model description	R2	RMSE %	% within 1 St Err	% within 2 St Err	Diff. of slope from 1	Mean rank	Com-plexity
1	Original (4A)	0.54	79	34	52	0.48	7	0
2	Original + typing errors	0.57	78	33	51	0.56	7	1
3	Sigmoid + errors	0.52	78	35	50	0.41	6	2
4	Typing as in mean 4B, linear tracking	0.47	76	36	54	0.45	6	2
5	Individual typing, linear mean tracking 4A	0.67	61	41	64	0.28	3	3
6	Individual typing, sigmoid mean tracking 4A	0.67	61	41	65	0.30	3	4
7	Individual typing, sigmoid mean tracking 4B	0.67	59	43	63	0.26	2	5
8	Fully calibrated	0.71	56	45	65	0.26	1	6

Table 4.5 Mean performance of model number 5 on four dependent variables, expressed in seven measures of fit.

	R2	**RMSE**	**RMSE %**	**% within 1 St Err**	**% within 2 St Err**	**Slope**	**Diff. of slope from 1**
Total time	0.86	3.42	30	26	52	1.55	0.55
Maximum deviation of the cursor	0.67	12.39	14	56	82	0.81	0.19
Total time cursor was outside of target	0.53	0.47	111	35	55	1.16	0.16
Score	0.62	0.05	87	47	66	0.78	0.22
Mean	0.67	N.A.	61	41	64	N.A.	0.28

In general, the model's performance (mean rank) increased with an increase in the complexity of the model. As expected given the observed individual differences in performance (see experimental results), the models that were not calibrated in any way to individual differences in skill (i.e., models 1-4) performed poorly. That is, the model needed at least some components for individual differences to explain the observed patterns in the human data.

Based on the ranking and the model complexity, I decided to perform further analyses on model 5. The performance of this model (when judged by average fit) was close to the performance of the best fitting model (model 8). However, its complexity value was a lot lower. That is, this model could explain a relatively similar proportion of human performance with a relatively smaller set of parameters. The model was identical to the model for experiment 4A, except that a component for typing errors was added and that the typing model was calibrated to individual performance in single-task trials. No further calibration to individual performance (i.e., of the tracking and dual-task components) was part of this model.

Table 4.5 provides the measures of fit for each dependent variable (or performance metric) for model 5. Note that the goodness of fit varied between different performance metrics (e.g., trial time, maximum deviation) and between different measures of fit (e.g., R2, RMSE, slope). For example, the measure "total time" had the best R2 score, and a good RMSE percentage, but performed poorly on the "slope" measure. In contrast, for the measure of "total time outside" this pattern was reversed. This diversity supports the choice for the analysis method in Table 4.4, in which average fit was compared across different performance metrics and for multiple measures of fit.

Assessment of aggregate performance: Did participants perform in line with the predicted optimal strategies?

With the selected model (model 5 in Table 4.4), I investigated how well human performance corresponded with the model's predictions for performance by optimal strategies. For the human data, the same data was used as described in the experimental results section (i.e., average performance for each individual during the last five trials of each condition). For the model data, for every individual and every condition, the strategy alternative was selected that on average achieved the highest payoff score.

For some individuals and some conditions, the model predicted that multiple strategy alternatives could achieve a score that was highest. In these cases, performance for all measures of interest (e.g., trial time, maximum deviation of the cursor, number of digits typed per visit) was averaged across these optimal strategy alternatives.[18]

Performance for the model (white bars) and human data (grey bars) is plotted side by side in Figures 4.14-4.19 for the following dependent variables: trial time, (Figure 4.14) maximum deviation of the cursor (4.15), total time the cursor was outside of the target (4.16), payoff score (4.17), average maximum number of digits typed per visit (4.18), and mean time spent in the tracking window per visit (4.19). Table 4.6 summarizes the fit of these models on each variable for the following measures of fit: R2, RMSE (and RMSE%), and the number of conditions for which the error bars of model and human data overlap. In addition, an ANOVA was applied to the model data to see if similar effects occur for the model data as observed in the human data (similar to the proposal by Taatgen & Van Rijn, 2010). The ANOVA used the same factors as the ANOVA that was applied to the human data. The ANOVA results are given in Table 4.7 (compare this with Table 4.2 that reports human performance).

Table 4.6 summarizes for the main effects and the interaction effects of the ANOVA how many effects were correctly predicted by the

[18] Of course the *average* value for performance metrics (e.g., trial time, maximum deviation) might differ across the various strategy alternatives, and their mean value might not be representative of any individual strategy alternative. However, this method was the simplest way of making performance comparable across different strategy alternatives. Alternative selection methods might for example be to select the strategy that achieved the best mean value on some other measure of performance (e.g., trial time, or maximum deviation of the cursor). These selection methods underlie stronger assumptions (e.g., that some measures of performance are more important than others) than the approach taken here.

model and how many effects were incorrectly predicted (i.e., were predicted not to be significant but were significant in human data, or the inverse). In cases where one data set (i.e., model or human data) predicted a marginal effect and the other dataset predicted no effect or a significant effect, this was counted as explaining "half" of the effect. ANOVAs were not applied to the payoff score data, as the payoff function was an independent variable.

Table 4.6: Measures of fit that express the correspondence between human performance and the model's predictions for optimal strategies for six dependent variables.

	R2	**RMSE**	**RMSE %**	**Nr error bars**	**ANOVA main effect correct**	**ANOVA main effect wrong**	**ANOVA inter-action correct**	**ANOVA inter-action wrong**
Total time	0.89	5.2	44	4/8	4/4	0	2/2	3
Maximum deviation of cursor	0.90	11.11	13	2/8	4/4	0	0.5/1.5	1.5
Time outside of target	0.63	0.34	101	3/8	3/4	1	1/4	4
Maximum nr of digits typed	0.97	2.83	27	4/8	4/4	0	1/1	0
Tracking time	0.90	0.79	63	0/8	3/3	1	0/0	0
Payoff score	0.76	5.3	93	2/8	NA	NA	NA	NA
Mean	0.84		57	2.5/8	95%	0.4 effects	72%	1.7 effects

Table 4.7: Summary of statistical effects in model 4B when the mean performance of the best strategy alternatives are used as data points (i.e., one datapoint per participant, per condition).

		Dependent variable				
		Total trial time	**Maximum cursor deviation**	**Total time cursor outside target**	**Maximum nr of digits per visit**	**Mean visit time, tracking window**
Independent variable	**Payoff function (P)**	***	***	***	***	***
	Noise (N)	***	***	***	***	***
	Radius (R)	***	***	***	***	***
	Interkeypr. interval (I)	***	***		***	***
	P x N	***	***	***		
	P x R	***		***		
	N x R	***	*			
	P x I				***	
	N x I	**				
	R x I	***	**			

.: .05 < p <= .10;
*: .01 < p < .05;
**: .001 < p < .01;
***: p <= .001

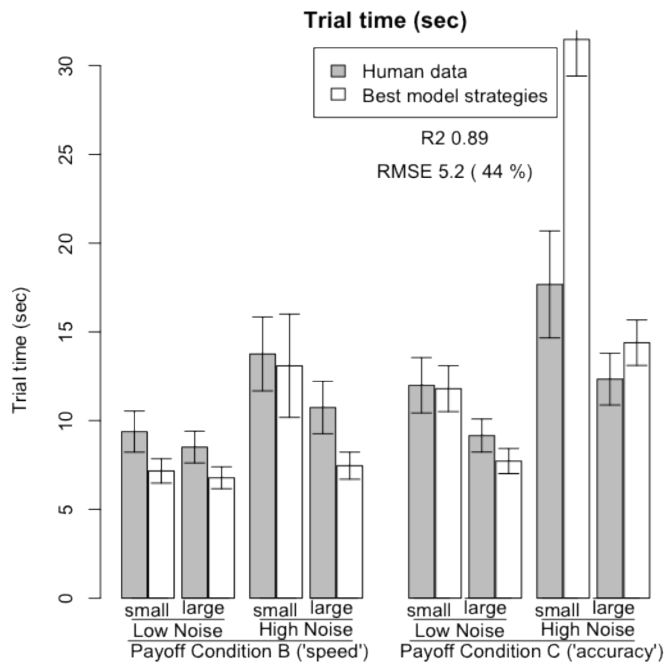

Figure 4.14: Correspondence between human mean performance and model predictions by optimal strategies for total trial time.

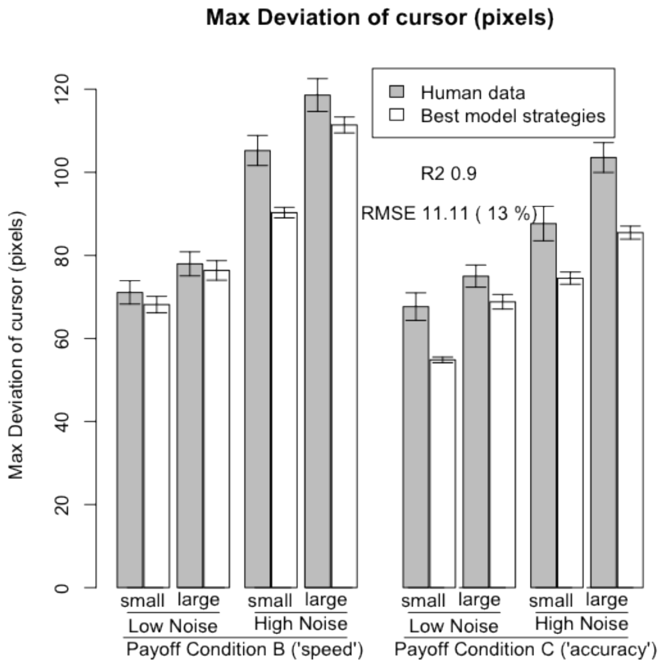

Figure 4.15: Correspondence between human mean performance and model predictions by optimal strategies for maximum deviation of the cursor.

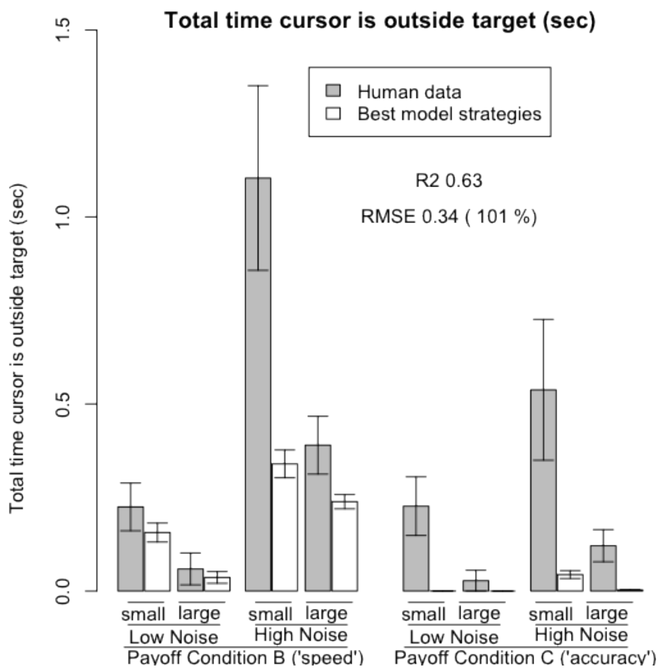

Figure 4.16: Correspondence between human mean performance and model predictions by optimal strategies for total time that the cursor is outside of the target area.

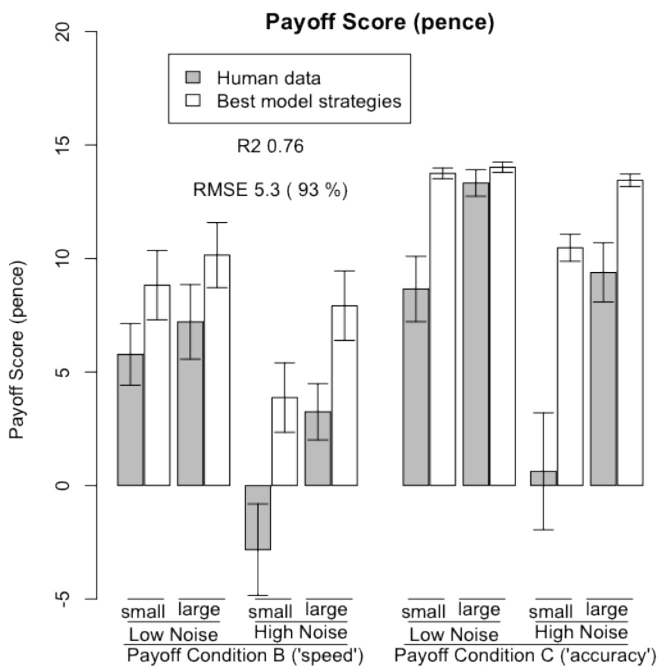

Figure 4.17: Correspondence between human mean performance and model predictions by optimal strategies for payoff score.

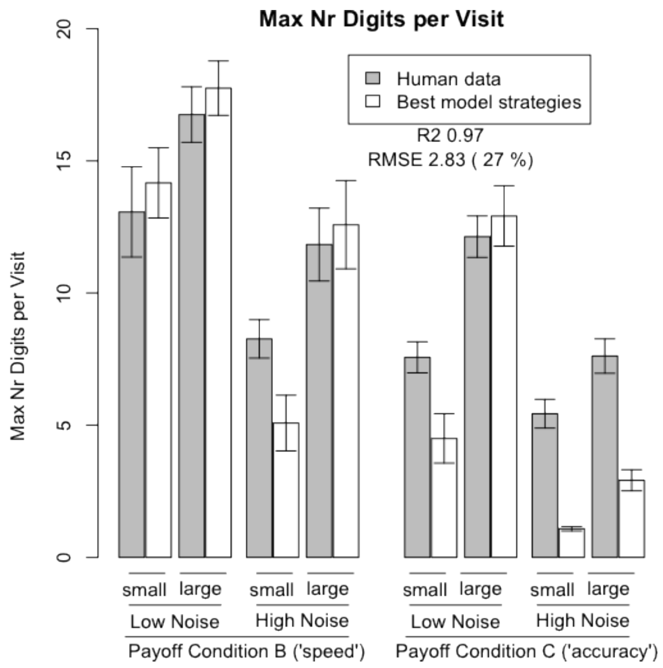

Figure 4.18: Correspondence between human mean performance and model predictions by optimal strategies for maximum number of digits typed per visit to the typing window (“strategy”).

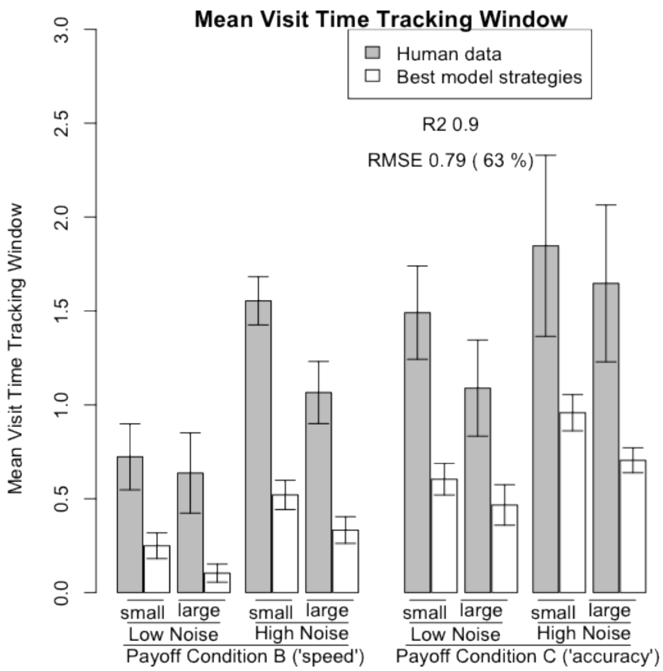

Figure 4.19: Correspondence between human mean performance and model predictions by optimal strategies for average time spent in the tracking window per visit.

In general, human performance followed the same trend as the optimal model strategies predicted. This is reflected in two ways in the results. First, the R2 values were reasonable to good (i.e., three out of five measures were above 0.89). Second, the model captured all main effects correctly for four dependent variables, and the model also captured most interaction effects correctly. At the same time, few effects were incorrect (on average 0.4 main effects and 1.7 interaction effects).

Note also, how both the model and the human data predict that on almost all dependent variables there should be main effects of the three variables of interest: payoff function, task difficulty (noise and radius), and individual differences in skill (interkeypress interval). This correspondence in trend suggests that participants adapted their performance in such a way as to perform optimal, given the payoff function, the task difficulty (noise and radius), and their individual differences in interkeypress interval typing speed.

At the same time, the model predicted that, on average, human performance was not yet identical to predicted optimal performance. This was evident in two measures of performance. First, in most conditions (5.5 out of 8) the error bars surrounding the model data and surrounding the human data did not overlap. This hints at a significant difference between human and model performance. A second indicator of the discrepancy between model and human data was that the RMSE percentage scores were relatively high. On average, predicted optimal performance was more than 50% away from the mean human score. Figures 4.14 - 4.19 help to identify where these discrepancies occurred. For most measures, the largest discrepancy was in the high noise, small radius condition. Note that this was also the hardest task condition.

Other discrepancies also occurred in the high noise, large radius condition.

The RMSE values of the various dependent variables provide insight into why participants might not have achieved the predicted optimal payoff scores. The number of digits that they typed per visit to the typing window (see Figure 4.18) was on average 2 to 3 digits away from predicted optimum performance. In payoff condition B ("speed") the number of digits typed was too low in three conditions, whereas in payoff condition C ("accuracy") it was too high in three conditions.

For the measure of mean time spent tracking per visit (Figure 4.19), the model systematically predicted that participants could have spent less time in the tracking window than was observed for the human data. Most likely, this can be attributed to a simplification in the model: the model immediately started tracking when the tracking window opened. In contrast, human participants might have needed time to locate the position of the cursor first. Furthermore, if participants held the joystick in an incorrect angle when they opened the tracking window, this might first have moved the cursor away from center. As a result, participants then had to spend extra time compensating for this movement. The model can be considered a model of idealized tracking performance, as it does not take these effects into account. More fine-grained data of human performance (i.e., eye-tracking data) is needed to model these effects.

Non-surprisingly, these slight discrepancies in strategy between model and human data had their effect on all other measures of overall performance (Figures 4.14 - 4.16) and on the score (Figure 4.17). The score achieved by the human participants was on average lower than the predicted optimum (RMSE of 5.3 pence). I will now discuss the predicted performance data for various strategies in more detail to find explanations for these discrepancies.

Score predictions for other strategies

The preceding results demonstrated that participants' performance followed the same trend as predicted by the model for optimal performance. However, the strategies that the participants applied were not identical to the predicted optimum strategies. Might it be that the strategies that they applied also lead to good performance? To investigate this, I analyzed the predictions for the payoff score of various strategies.

Figure 4.20 plots the model's predictions for the mean score per strategy (vertical axis) as a function of strategy (horizontal axis). The different payoff conditions are plotted in two columns, and the different task difficulty conditions are plotted in different rows. To create these plots, I calculated the average score per strategy alternative for each participant in each condition. For each strategy, I then selected the strategy alternative that achieved the highest score. With this data, I then calculated the mean scores across participants and calculated the associated standardized errors.

In Figure 4.20 model mean values are plotted as circles, with their standard error bars plotted as crosses above and below the means. The range of strategies of which the standard error bars overlap with the standard error bars of the *best* model strategy are emphasized in two ways. First, these model points are drawn in black (compared to grey for other points). Second, the plot area behind these strategies is shaded in dark grey.

The Figure draws the attention to the fact that in seven out of the eight conditions there are multiple strategies of which the predicted score overlaps with the predicted optimum score. Especially in payoff condition B, the range of "optimal" strategies is large (i.e., more than 10 strategies).

Human mean performance and standard error bars are also drawn on the Figure (as a plus sign with error bars surrounding it). At the position of the human mean, a dashed grey line was drawn to make it easier to locate the point. In six conditions the error bars of the achieved score by the humans overlap with the error bars of the model's predictions for scores of this strategy. This highlights the accuracy of the model in predicting the score. Moreover, in five out of eight conditions, human performance is inside the grey rectangular area. That is, in these conditions participants applied strategies of which the score was predicted to be close to the highest possible score.

This analysis suggests that participants *did* achieve good scores. This is a good and a bad sign. It is a good sign in that participants were able to achieve good performance. At the same time, it is a bad sign, because it implies that there were multiple strategies that could be considered optimum. In particular, in the low noise, large radius condition for payoff condition C the set of "optimal" strategies was extremely under constrained. It seems like participants' score should not vary that much depending on the strategies that were applied.

This observation contrasts with the aim of this Chapter: to use a payoff function to identify a more narrow, constrained set of strategies that is optimal, and to see whether participants adopted those strategies. Ideally, a payoff function has a distinctive global maximum and no other local maxima. In Chapter 5, I will use a mathematical

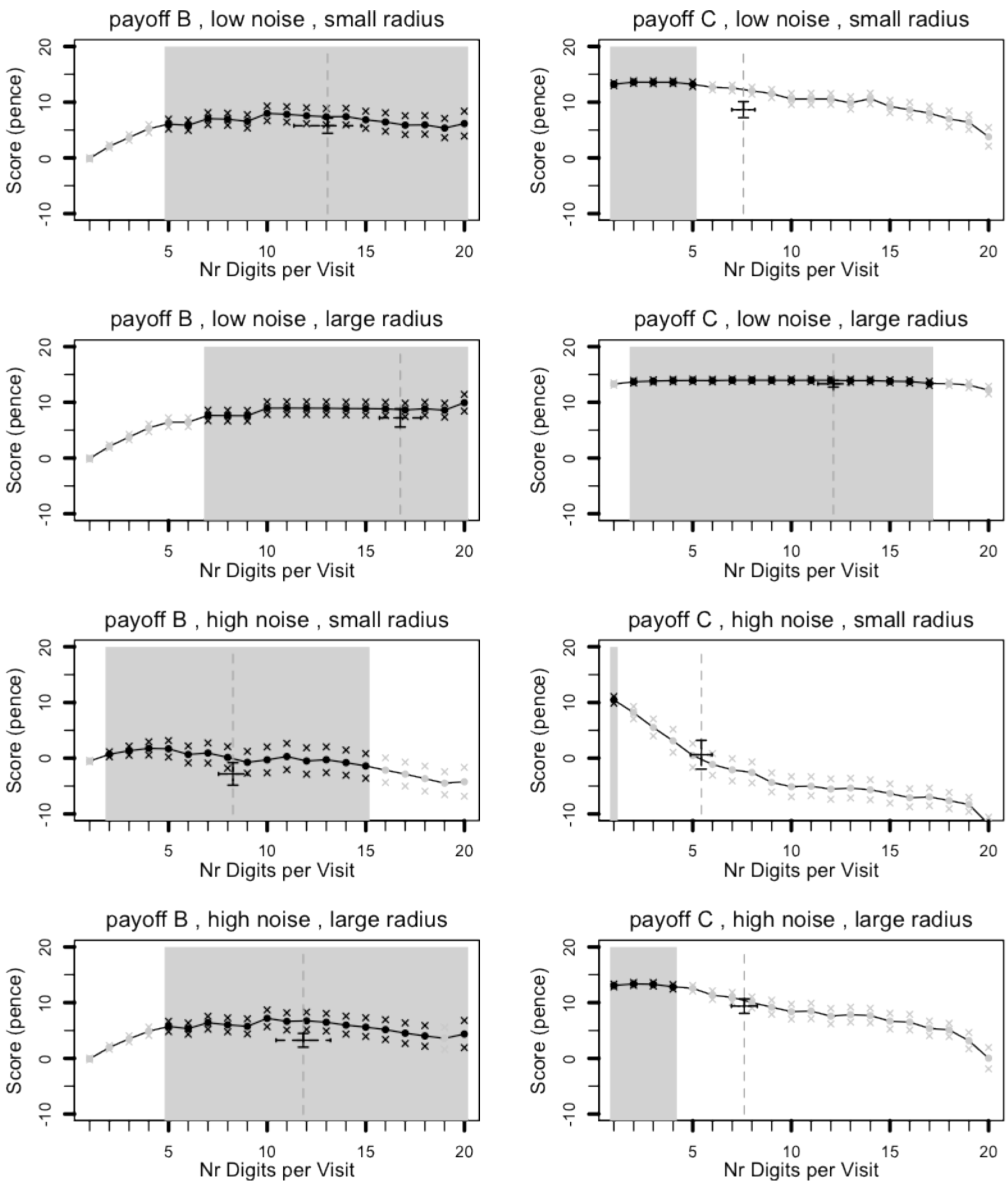

Figure 4.20: Predictions for the on average best achievable scores per strategy (circles) with human mean performance overlaid on top (use the grey dashed bar to locate this). The grey colored areas highlight strategies of which the error bars surrounding the mean model score of a strategy overlap with the error bars of the best scoring strategies.

model to explore whether such 'ideal payoff curves' (with a unique, distinctive global maximum) are even possible in dual-task environments like the one used here.

There are at least two caveats to note to the above conclusion on the absence of ideal payoff curves in this dataset. First, for the model analysis I selected the optimal strategy alternatives for each strategy (i.e., I selected the optimum tracking time for each strategy). There might be more variability in achievable scores between different strategy alternatives within a strategy.

Second, the plots report mean performance across participants. As there are individual differences in performance, it is likely that the shape of the payoff curves varied between individuals, and that the flat curves in Figure 4.20 are a result of averaging across participants.

To further look into this, the next analysis will investigate the shape of the payoff curves for each individual, and will investigate how people adapted to their individual curves over time. It might for example be that participants were still learning how to adapt their performance when I assessed their performance during the last five trials of each condition.

Learning

To investigate adaptation over trials, I produced a series of plots that shows the maximum number of digits that participants typed per visit ("strategy", grey dots, vertical axis) as a function of the dual-task trial number (horizontal axis). Figures 4.21-4.26 show the results. Per Figure, plots for four participants are shown. Each plot contains data on all four task conditions that the participant experienced, in the order in which they experienced those (see labels at the top of each plot). A red dashed line shows the predicted trend line in the human data, as predicted by a statistical linear model for each condition.

Behind the participant data, I plotted rectangular areas that show the relative success of each strategy as predicted by the model. The rectangles differ in their grey shade: the darker the shade, the better the score. The darkest grey rectangles highlight strategies that on average achieved a score that was less than 0.5 pence away from the predicted maximum score for that specific condition and participant. The second darkest grey rectangles highlight strategies that achieved a score that was between 0.5 and 1 pence away from the maximum score. The lightest grey shade highlights strategies that on average achieved a score that was between 1 and 2 pence away from the maximum score. Strategies that achieved a score that was 2 or more pence away from the predicted maximum were not highlighted. Note that a correspondence in "grey shade" between different conditions or different participants does *not* imply that the same scores were achieved. This is, because "grey shade" was always determined relative to the current condition and participant.

If participants adapted their performance optimally to each condition, then their performance should lie inside the grey rectangular areas, preferably inside the dark grey areas. The Figures suggest that this differs between participants and conditions. Some participants adapted their performance really well, and their performance was almost always inside the grey areas. For example, participants 101, 108, and 110 in payoff condition B mostly applied strategies that were predicted to lead to optimum scores. In payoff condition C, only few participants adapted strategies that were predicted to be optimum.

Participants 203 and 204 seem to be two of the better performing participants.

To quantify how well performance was when compared to the (relative) optimum score, I analyzed the average total number of trials during which the strategy was inside the grey bars (i.e., in an area where the score was predicted to be less than 2 pence away from the optimum score). For this analysis, an ANOVA was applied on the data for total number of strategies that fell inside the grey areas. The ANOVA had payoff function, noise, and radius as factors. The ANOVA found a significant main effect of noise, $F(1, 22) = 121.57$, $p < .001$. Performance was inside the optimal area in more than three times as many trials when noise was low ($M = 15.04$, $SD = 3.64$) compared to when noise was high ($M = 5.73$, $SD = 4.35$). Similarly, performance was better when the radius was large ($M = 14.04$, $SD = 3.36$), compared to when it was small ($M = 6.73$, $SD = 4.36$), $F(1, 22) = 73.07$, $p < .001$. There was no effect of payoff function, $F < 1$. There was a significant interaction effect between payoff function and noise, $F(1, 22) = 7.23$, $p = .013$. There were no other significant interaction effects.

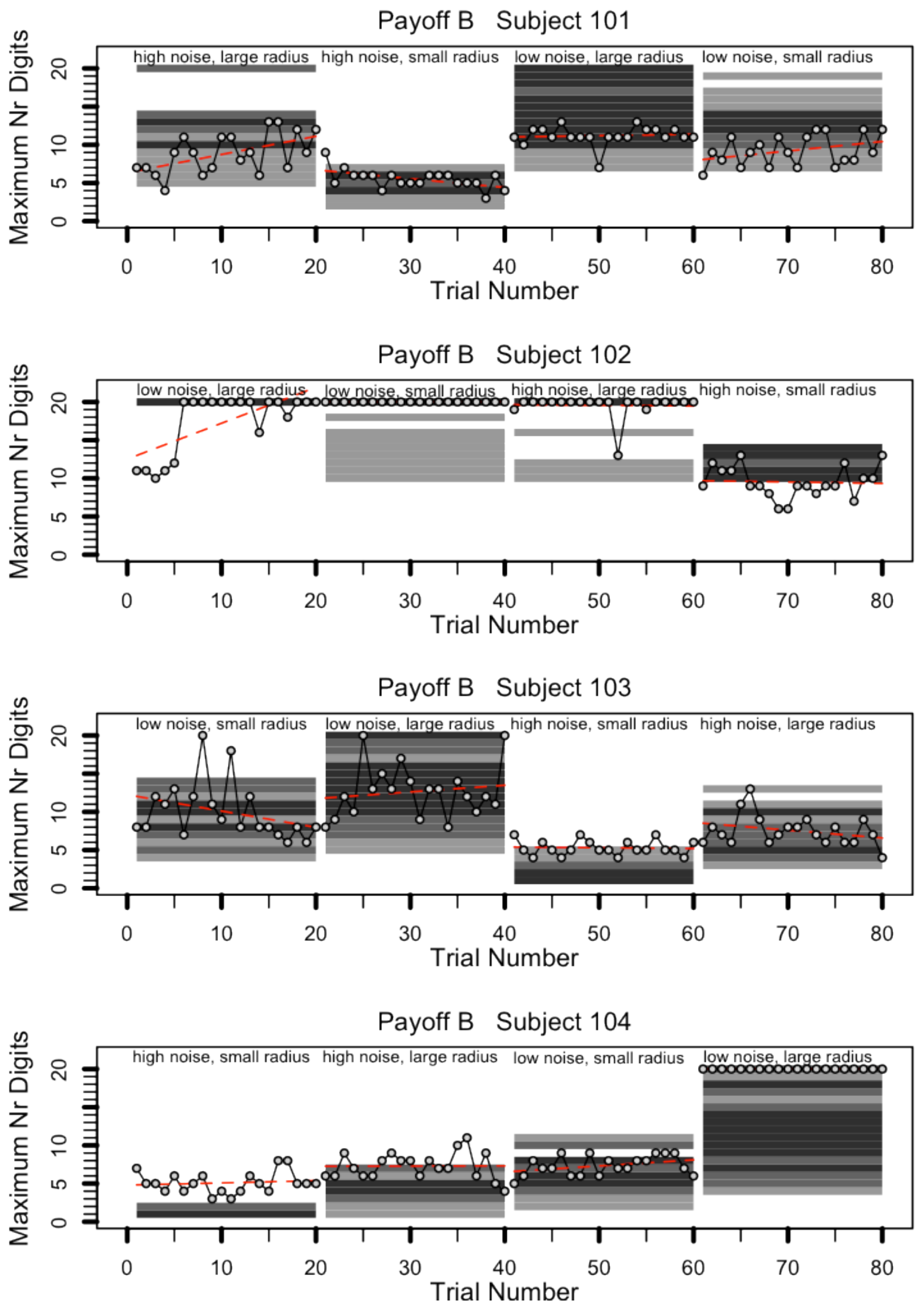

Figure 4.21: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition B, participants 101-104.

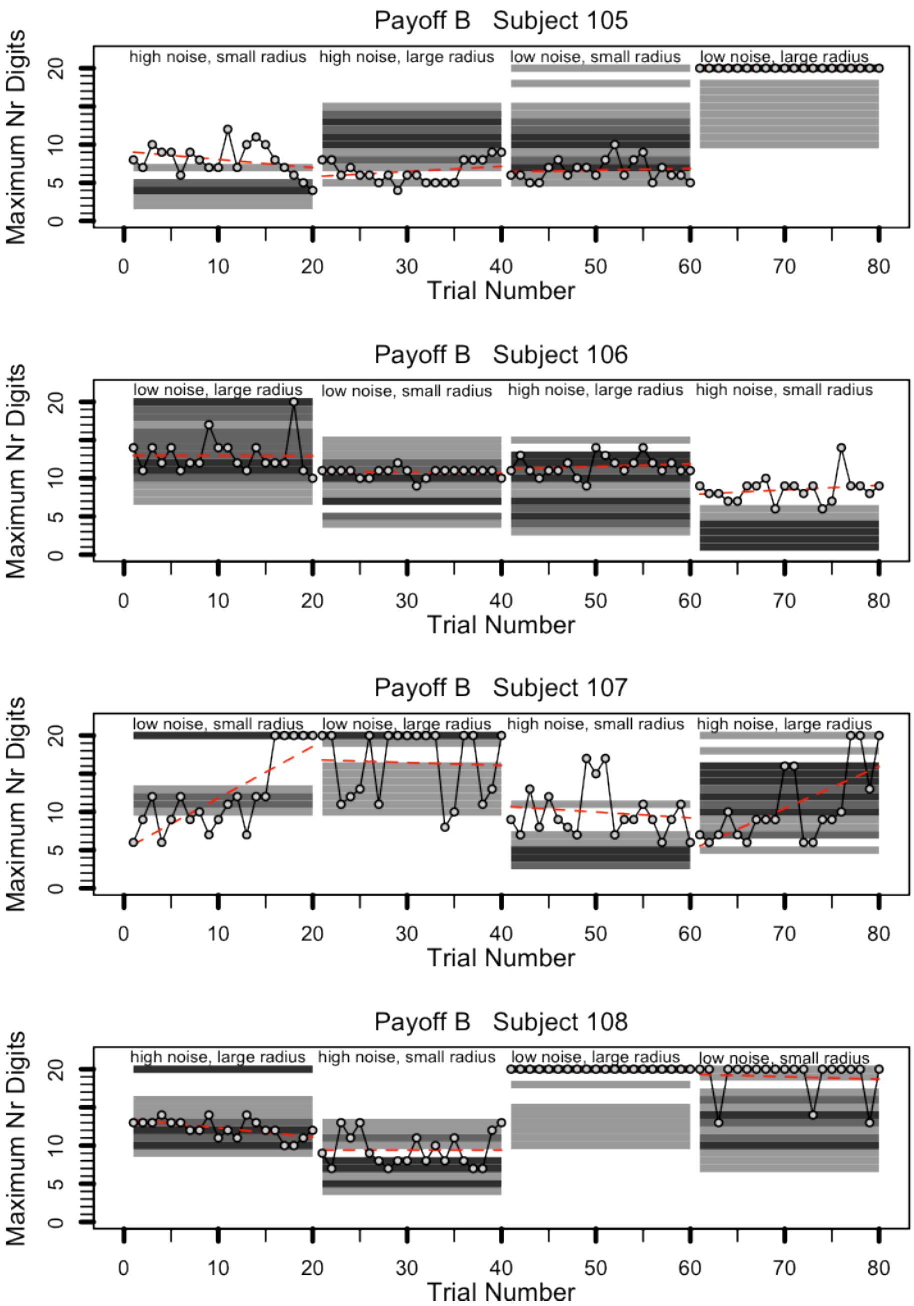

Figure 4.22: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition B, participants 105-108.

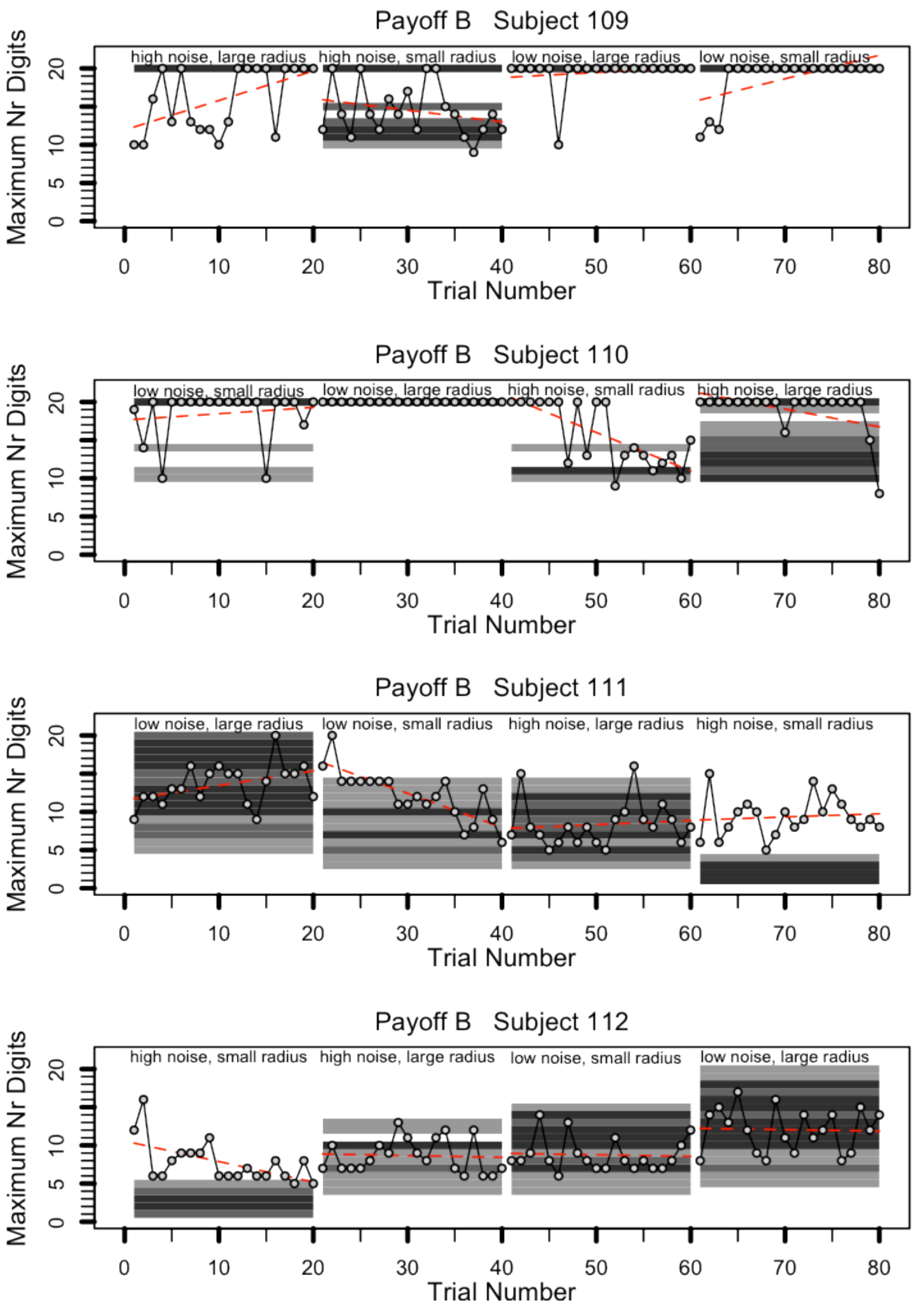

Figure 4.23: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition B, participants 109-112.

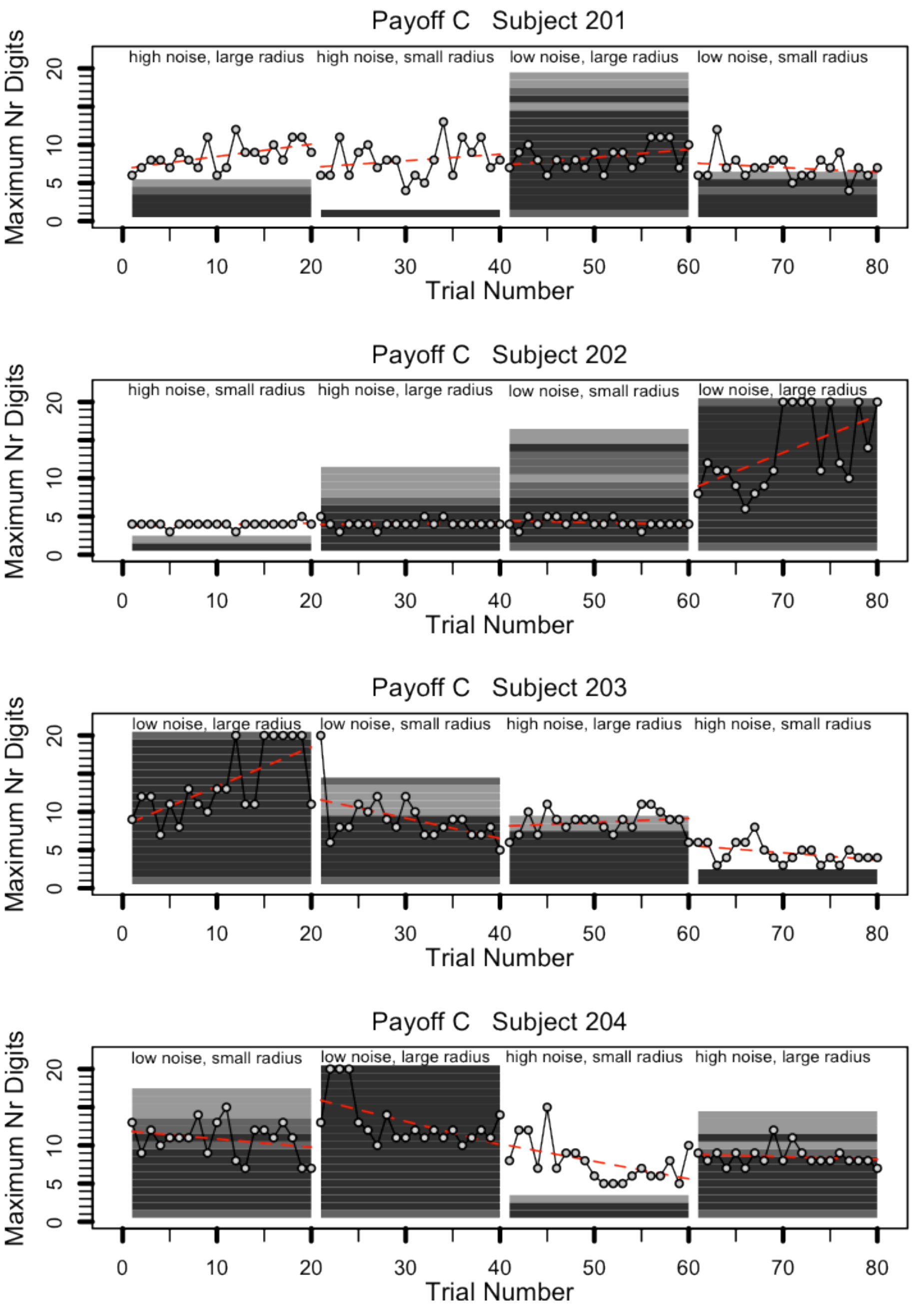

Figure 4.24: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition C, participants 201-204.

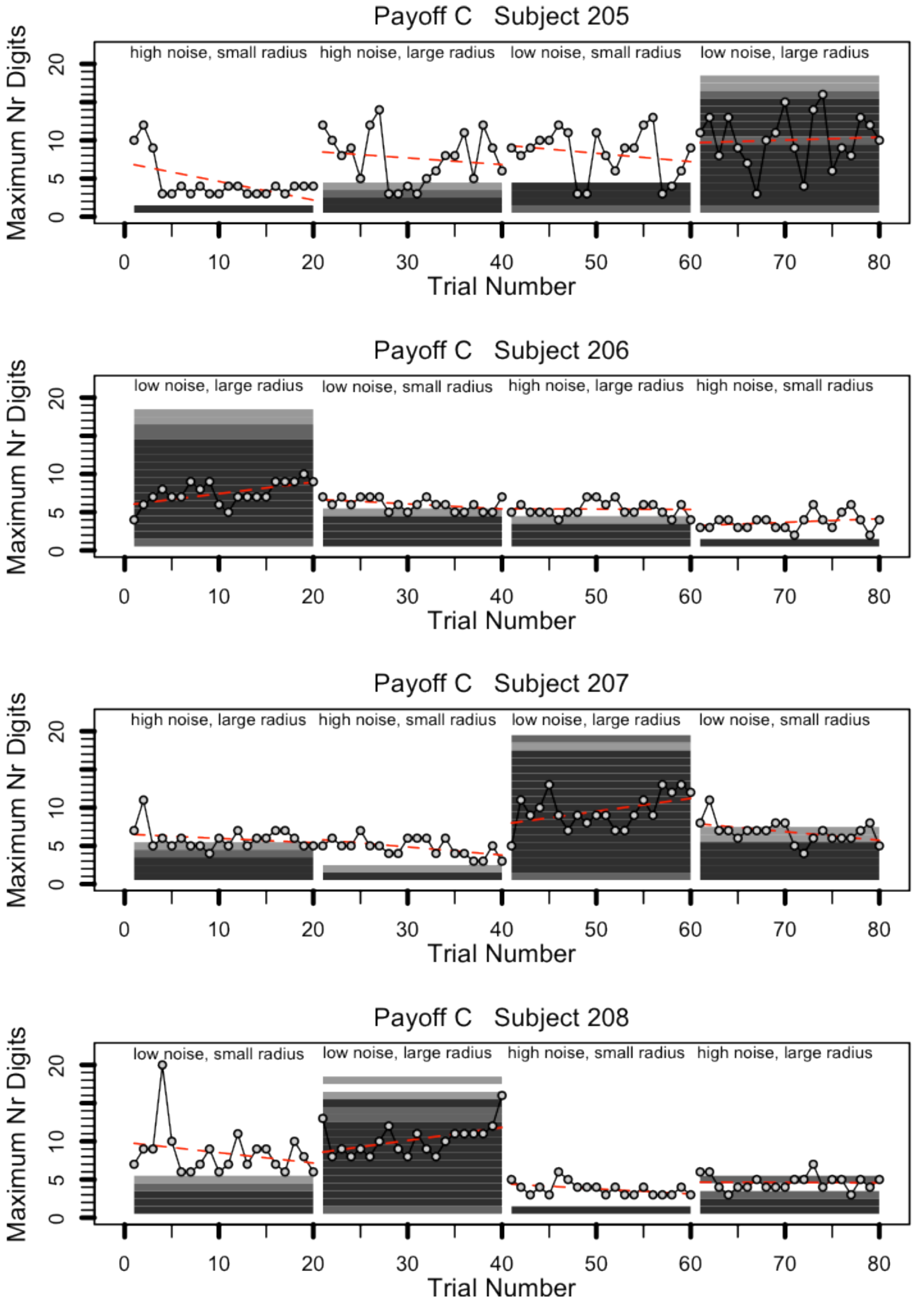

Figure 4.25: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition C, participants 205-208.

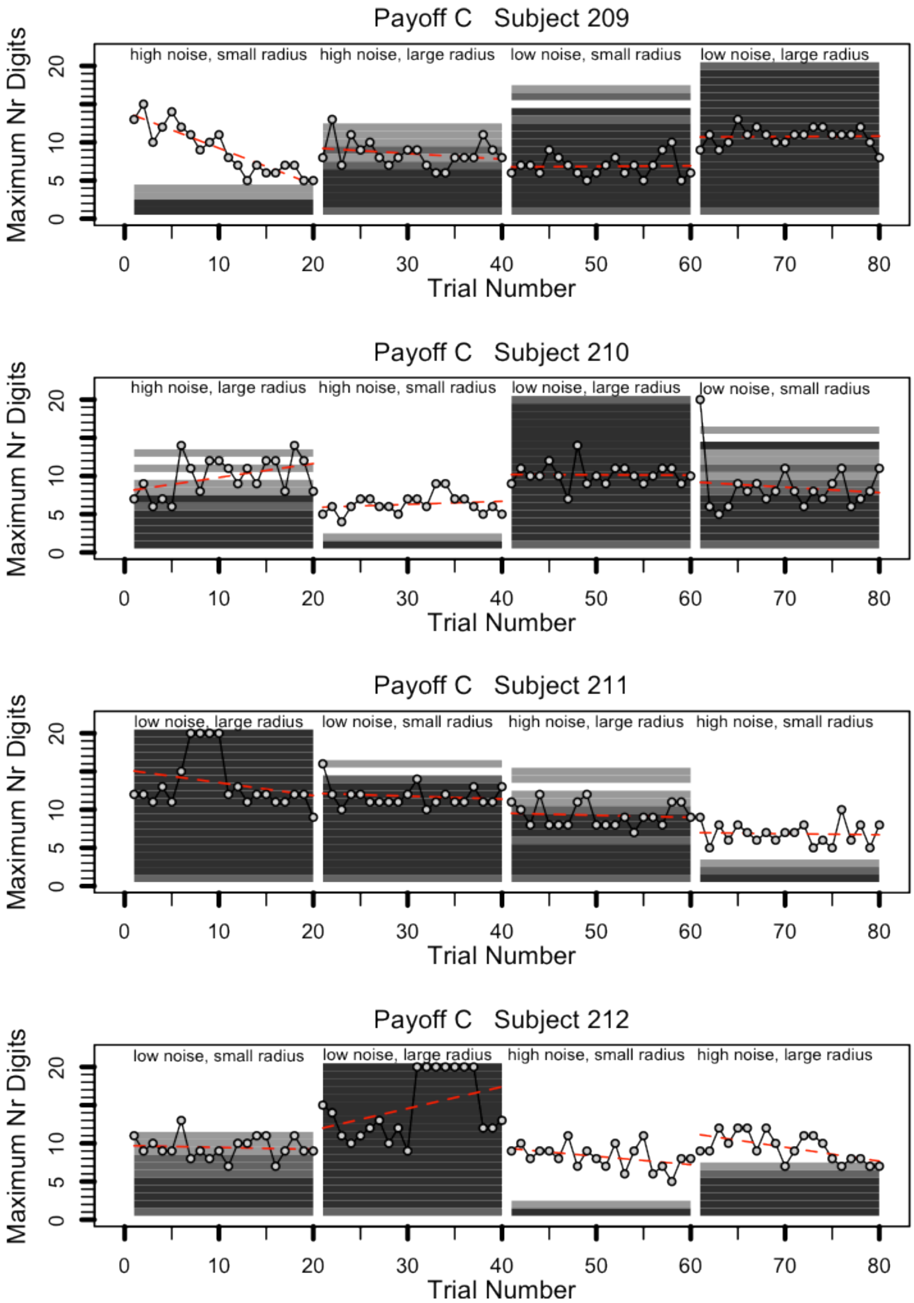

Figure 4.26: Applied strategy by humans (grey dots) over trials with a red dashed trend line fitted through them, per condition in the experienced order (see labeling at top). Bars behind the participant data show which strategies were within 0.5 pence of the best scoring strategy for that participant and condition (black), within 1 pence (dark grey), or within 2 pence (light grey). Data for payoff condition C, participants 209-212.

The above analysis was repeated, but this time only looking at the average number of strategies that could achieve a score that was within 0.5 pence of the maximum score (i.e., that was inside the dark grey bar). Mostly similar effects were found. Performance was better when noise was low (M = 12.33, SD = 4.05) compared to when it was high (M = 3.54, SD = 3.98), $F(1, 22) = 186.42$, $p < .001$. Performance was also better when the radius was large (M = 10.90, SD = 3.97) compared to when it was small (M = 4.98, SD = 4.07), $F(1, 22) = 63.12$, $p < .001$. There was again no main effect of payoff function, $F < 1$. There was a significant interaction effect between payoff function and noise, $F(1, 22) = 12.67$, $p = .002$. There also was a marginal interaction effect between payoff function and radius, $F(1, 22) = 4.06$, $p = .056$.

Taken together, these results suggest that how well participants performed in comparison with their own payoff curve (i.e., with the location of the maximum strategy) depended on the task difficulty, but not on the payoff function. When the task difficulty was relatively easy, due to low noise or a large radius, participants on average achieved a maximum score on more than half of the trials.

The absence of a significant effect of payoff function in this analysis is good. It implies that the manipulation of payoff function did not pose any limitations on participants' ability to adapt performance to the payoff function.

Even when participants did not yet apply optimum strategies, they might still be adapting their performance in an appropriate manner. In these cases, their performance should gravitate towards the optimum strategies over the course of the trials. This can be judged by investigating whether the trend line was nearing the area of the optimum strategy.

For some participants this was clearly not the case. For example, participant 201 (payoff condition C) applied very similar strategies across the different task difficulty conditions, and in three of the four conditions, these strategies were predicted to *not* lead to the highest possible payoff scores. Participants also seemed to transfer strategies that proved successful in the past to other conditions. This can be seen in the Figures when the trend lines are relatively flat, and located around similar positions across conditions. This was for example the case for participants 106 and 202.

Did some conditions require more adaptation than others? To quantify this, I analyzed the slope of the regression lines. It was assumed that steeper lines indicated faster adaptation, irrespective of whether the slope was positive (i.e., an increase in the number of digits typed over trials) or negative (i.e., a decrease in the number of digits typed over trials). In contrast, shallow slopes were assumed to involve less adaptation. I therefore calculated the absolute value of each trend line and compared the values of these trendlines in an ANOVA that had payoff, noise, and radius as factors. It was expected that the amount of required adaptation also differed depending on the total experience that the participant had with the task. To accommodate for this, the block in which a condition was experienced (first, second, third, or fourth) was added as an additive factor.[19]

The analysis found a significant main effect of block, $F(1, 87) = 6.49$, $p = .012$. A bonferroni-corrected post-hoc test found a significant difference between the absolute slope value of the first ($M = 0.21$, $SD = 0.18$) and the second block ($M = 0.10$, $SD = 0.12$), $p = .046$, between the

[19] That is, the LME4 regression equation in R was: x~PayOffFunction x Radius x Noise + Block + randomEffect(Participant)

first block and third block (*M* = 0.08, *SD* = 0.11), *p* = .007, and a marginal significant effect between the first block and the fourth block (*M* = 0.11, *SD* = 0.14), *p* = .059. There was no significant difference between any other combination of blocks. The higher absolute value of the slope in block 1 indicated that participants adapted their performance in this block at a faster rate than in other blocks. This is perhaps non-surprising, given that it was the first block during which they experienced the dual-task scenario.

The ANOVA did not produce any other main effects (all *F*s < 1). There were no interactions (*F*s < 2) except for a three-way interaction between payoff function, noise, and radius, $F(1, 87) = 6.79$, $p = .011$. A follow-up test found only a significant interaction between noise and radius when the payoff condition was C. Further analysis suggested that this was because participants had a steeper learning curve in the low noise, large radius condition. Note that this was the only condition (in payoff condition C) in which typing many digits per visit was optimal for most participants. In the other conditions, it was best to only type a small set of digits. The steep learning curve might therefore be a result of this large contrast between conditions.

The above quantitative analysis is surprising. It suggests that the amount of learning mostly depended on the position at which a condition was experienced. If it was the first condition, more learning was required. Despite that some of the conditions were more difficult than others (e.g., the small noise, large radius condition), this was not reflected in the ANOVA results: the steepness of the learning curve did not depend on the payoff, noise, or radius level.

4.6.3. Discussion of Results

The model analysis suggested that participants adapted their performance in such a way as to achieve an optimum score, given the payoff function, the task difficulty (noise and radius), and individual differences in interkeypress interval. There was a high correspondence between the trend in the model data and the trend in the human data, as expressed using R^2 and correspondence in ANOVA results. However, the exact strategies that participants applied were not yet the ones that (on average) achieved the highest mean score, as evident in for example high RMSE values.

In additional analyses, I explored two aspects that were thought to contribute to these discrepancies. First, an exploration of the mean score across different strategies found that the payoff curves were not ideal. In most conditions, there were multiple strategies that achieved scores that were close to the predicted optimum score. The participants mostly applied such strategies.

Second, I explored how participants adapted their strategies over time. It was found that there were qualitative differences between participants in how well they adapted their performance to the payoff conditions. For some participants, there were hints at strategy transfer effects between conditions. Compared to their own payoff curve, participants applied more appropriate strategies (which were predicted to on average achieve a score within 0.5 pence of the predicted maximum) when the tracking task was easier due to smaller noise or a larger radius of the target.

A quantitative analysis of the learning speed between conditions and blocks showed that participants had the steepest learning curve during the first block they experienced. The curves did not differ between conditions, except for one three-way interaction between payoff function, noise, and radius.

In the model development section, I described how multiple models could be developed to explore the above issues. I selected a model based on its goodness of fit with the human data (when comparing performance for similar strategies as observed in the human data) and based on the complexity of the model. It was found that the model needed at the least a calibration of the typing model to capture the trend in the data. Further refinements required additional model complexity, but only lead to minor improvements in model fit.

4.7. General discussion

4.7.1. Summary of this Chapter

This Chapter contributed to the overall objective of my thesis by demonstrating how people adapt their interleaving strategy in a dual-task setting to three factors: (1) task difficulty characteristics, (2) individual differences in skills (i.e., typing speed), and (3) payoff (a formal way of capturing objective). Subsequent modeling analyses suggested that people also were reasonably good in adapting in such a way as to achieve a good payoff value. Human performance followed the same trend as the predictions of a model that achieved optimum scores in each condition. However, participants did not apply the exact same strategies as the model predicted to be optimal. A further analysis investigated two potential causes for this. First, there were multiple strategies that achieved average scores that were similar to the score of the best strategy (i.e., the error bars overlapped). In most conditions, participants applied such strategies. This can be taken as an argument that participants were satisficing, rather than optimizing performance (Simon, 1956). Second, I investigated the learning trajectory of participants and found that some participants transferred strategies from one condition to the next. More detailed explanations of these effects and ways to reduce their impact are discussed below.

4.7.2. Contribution of this Chapter to the literature

Contribution 4.1: Introducing a formal method to identify 'optimum' strategies in dynamic concurrent dual-task scenarios

The first contribution of this Chapter to the literature is the introduction of a method to make more precise predictions about what "optimal" behavior is in a concurrent dual-task setting. The majority of previous studies, including those reported in Chapter 3, used instructions to get participants to prioritize one task over another. However, such instructions are open to subjective interpretation.

The modeling approach from Chapter 3 can partly deal with this, by identifying an optimal trade-off curve (or performance operating characteristic, Navon & Gopher, 1979; Norman & Bobrow, 1975). For all points on this curve, performance on the secondary task (e.g., dialing) is optimal given the criterion for a primary task (e.g., driving). However, as the curve is relatively long, the prediction of optimality is under constrained. Due to the use of a payoff function in this Chapter, performance on both tasks could be combined into a single objective feedback function. This allowed the identification of an *exact* optimum strategy, or series of optimum strategies, namely those that on average achieved the highest payoff value. Such a method has not been applied before to dynamic, concurrent multitasking settings, although similar methods have been used in PRP studies (Howes et al., 2009; Schumacher et al., 1999), and in discrete dual-task scenarios (Payne et al., 2007). In addition, incentives have previously been used to influence people's behavior in multitasking environments, without making predictions about optimal performance (Wang et al., 2007, 2009).

The use of a payoff function in this Chapter is not meant to be taken as an argument that payoff (or monetary incentives) is a prevalent feature in the real-world. Rather, payoff helps to quantify, for both the model and the participant, the objectives that people might have in the real-world in a coherent manner, rather than inducing them through verbal instructions.

In model 4B it was found that even with the introduction of a payoff function sometimes there are multiple strategies that achieve (near) optimal performance. In that sense, a payoff function does not always constrain the optimal performance space as much as desired. In Chapter 5 I will use mathematical models to investigate in more detail whether in a task environment like the one used here "ideal payoff curves" can be formed in which there are unique global maxima.

Contribution 4.2: A formal analysis of how task constraints, individual differences, and objective influence dual-task interleaving

The novel methodology allowed me to make a second contribution: to formally assess how task constraints, individual differences in skills, and objective (as formalized in a payoff function) shape human multitasking performance. Until now, other comprehensive theories of multitasking have focused only on some of these aspects – most notably on the interplay between the task environment and the cognitive architecture (e.g., Meyer & Kieras, 1997a, 1997b; Salvucci, 2005; Salvucci & Taatgen, 2011; Wickens, 2008), and the role of objectives (e.g., Brumby et al., 2009; Gopher, 1993; Horrey et al., 2006; Iqbal et al., 2010; Levy & Pashler, 2008; Navon & Gopher, 1979).

The role of individual differences in multitasking performance has typically been less well studied. There has been some focus in the process-oriented cognitive modeling community on explaining individual differences in performance. However, this focus has mostly been on bracketing performance for different types of individuals, such as fast and slow users (Card et al., 1983), or old and young participants (Meyer et al., 2001). More graded models of differences in skill have been applied, but are few in number. Most notable are the model of working memory by Lovett, Daily, and Reder (2000), in which performance on a new task was predicted based on an independent calibration of working memory skill, and the models of individual strategic differences in performance of the PRP task by Howes, Lewis, and Vera (2009).

More work on understanding individual differences in multitasking is needed, especially given that recent studies have confirmed that such differences exist, but have not identified the causes of these differences (Ophir et al., 2009 ; Watson & Strayer, 2010). Here, I demonstrated how characteristics of skills that are measurable outside the multitasking context (i.e., single-task typing performance) systematically change the utility of different strategies, and how this consequently influences and changes multitasking performance. That is, differences in multitasking were not attributed to differences in "general multitasking skill" (as was proposed by e.g., Gopher, 1993), but to differences in single-task performance (similar to proposals by e.g., Salvucci & Taatgen, 2011). Moreover, it was explored how these factors interact with the other factors that have previously received more attention in the literature: task environment, cognitive architecture, and objectives (e.g., see Figures 4.21-4.26).

A hallmark of the model is that it expressed performance on several measures. The fit between human performance and optimal performance as predicted by the model was not perfect, but still impressive. In particular, on most measures there were high R-squared values, and ANOVA analyses on the model data mostly replicated the same effects as observed in the human data. This result was especially impressive given the low 'complexity' of the model: the general structure of (the eventual selected) model 4B was identical to the structure of model 4A, and the only parameters that were changed were calibrated to an independent data-set (single-task performance).

Contribution 4.3: A re-appreciation of the flexibility of human performance

A third contribution of this work is that it helps to re-appreciate human flexibility in dual-task performance. This view has been around for longer (e.g., Moray et al., 1991; Navon & Gopher, 1979; Norman & Bobrow, 1975), yet computational cognitive architecture research has not yet fully understood the nature and causes of this flexibility until now (e.g., Meyer & Kieras, 1997a, 1997b; Salvucci & Taatgen, 2008, 2011). Instead, one of the dominant views is that people have "default" ways of interleaving given the characteristics of the task and of the cognitive architecture (esp. Salvucci & Taatgen, 2008, 2011).

In contrast, my work clearly demonstrates that there is flexibility within the performance space defined by the cognitive architecture and the task characteristics. This flexibility, and variability, exists both within individuals (e.g., as a result of different objectives) and between individuals (e.g., as a result of different skill levels). In this sense, the work in this Chapter is opening up a new area of research: to better understand variability in multitasking performance.

4.7.3. Reflection on model development and potential extensions

Reflection on model development and avoidance of over fitting

The model development started by modeling average human performance (model 4A). When individual differences in performance were encountered in experiment 4B, this was used as an argument to change the model structure to take these individual differences into account. In addition, new parameters were added to account for typing errors.

As with every modeling effort, the introduction of new parameters comes at the risk of over parameterization and over fitting (e.g., Pitt & Myung, 2002; Pitt, Myung, & Zhang, 2002; S. Roberts & Pashler, 2000). Three approaches were taken to avoid over fitting for model 4B. First, the number of free parameters of the models was kept to a minimum and only essential assumptions about performance were made. In practice, this meant that parameters were set to measurable units (e.g., interkeypress intervals), without making further assumptions about the underlying cognitive processes of these parameters (e.g., without assuming a series of actions, or production rules, that lead to a keypress). In this way, the assumptions were grounded in data. Moreover, although additional components were added to model 4B, it was very clear from the behavioral data that these new components (e.g., typing errors, individual differences in performance) were present in human performance. Therefore, not including these components would at least decrease the qualitative fit of the model.

Second, model complexity was taken into account when selecting the model with the best goodness of fit value. The eventual chosen model had a fit that was almost as good as the best fitting model, but was very simple. It followed the same structure as model 4A, with the addition of typing errors and a calibration of the typing model to single-task data. The fit is very impressive, given that the model could predict dual-task performance reasonably well, without any calibration to dual-task data.

Third, the model was assessed on multiple dependent variables (e.g., score, trial time, maximum deviation), and the larger strategy space was taken into account in this assessment. In this sense, the models were not fit to optimize a particular variable, or a particular strategy. In addition, the model was analyzed in various ways (i.e., R2, RMSE, ANOVA, slope of line y=x).

In this context it should be noted that, as is the case with any quantified theory, a good fit can never *prove* that the model is correct (Pitt & Myung, 2002; S. Roberts & Pashler, 2000). A good fit and correct predictions of performance for novel settings can only support the theory.

Note that my method for assessing the complexity of the model and for ranking the goodness of fit of the various models was very informal. In both cases, I assigned a rank value for various modeling components to each model and used the ranking of these values to handpick the model that provided the best fit, given the complexity. More formal methods such as the Bayesian Information Criterion (Schwartz, 1978) or Akaike Information Criterion (Akaike, 1973) could be applied to this end (see also, Pitt & Myung, 2002). These measures were not applied here, as the required definition of "number of free parameters" is not straightforward. In my models, the actual number of parameters was the same across most models, and the models differed in the *dataset* to which they were calibrated.

Moreover, in contrast to the way free parameters are used in most other modeling frameworks, I did not run an automated fitting procedure to find the best fitting parameter values. Rather, parameters were directly grounded in observed measures. Therefore, although the model had extra parameters, these were not completely "free" in the same way as in other model approaches, where values are based on the outcomes of fitting algorithms.

Reflection on possible model extensions

Given the relatively simple structure of the model, there is much room for extensions and further investigations. These extensions can broadly be categorized as: explaining more of the underlying psychological processes and moment-to-moment performance, accounting for more variability in performance, predicting a wider set of strategies, predicting learning, and conducting complementary analytic analyses.

First, more details of the underlying psychological processes and the moment-to-moment performance could be given for most components of the model. Such theories can provide an account of performance at different levels of abstraction (Newell, 1990; Salvucci & Taatgen, 2011, Chapter 1), see also discussion in Chapter 2 of this thesis. In particular, the model for typing errors is underdeveloped. For example, it does not offer an account for the exact position at which errors are made, for example due to the frequency of interruptions (e.g., Monk, 2004), as no such correlations were found in the data. Also, participants might strategically decide to speed-up or slow-down typing performance, which might shift their speed-accuracy trade-off (e.g., Guiard, Olafsdottir, & Perrault, 2011; Wobbrock, Cutrell, Harada, & MacKenzie, 2008). This could be added as a further dimension of strategic variation. See also Smith et al. (2008) for a more detailed Cognitively Bounded Rational analysis of human error.

The second, related, way in which the model could be improved is by taking even more of the variability of performance into account. Most model parameters only have a mean value. However, participants' performance might vary between trials, and samples from distributions might be better at predicting this variability. If the number of simulations would be large enough, this should still approximate the mean values used here.

Third, the strategy space might be broadened in two ways. First, the model was only used to explore simple strategies in which a consistent number of digits was typed during each visit. However, in practice people might have used strategies in which there was not a consistent pattern (e.g., type 8 digits, track, type 4 digits, track, type 3 digits, track, type 5 digits). Another way in which the strategy space might be broadened is by refining the typing component of the model. In the current model the assumption was built, that participants type digits in the same visit to the typing window during which they studied those digits. However, participants might have used more complex strategies, for example by studying digits in one visit to the typing window, and typing them in a subsequent visit. To account for such strategies, more detailed knowledge of people's moment-to-moment actions is required, as captured for example using eye-tracking (e.g., Hornof & Zhang, 2010). Such data was not available in this study[20].

The model did explore extreme strategies (no interleaving, and interleaving after every digit), and many strategies in between those extremes. It is therefore expected that performance of more "complex" strategies lies somewhere within the performance brackets of these extreme strategies (similar to bracketing approaches in e.g., Card et al., 1983; Kieras & Meyer, 2000).

A fourth way in which the model can be improved is by providing a formal theory of how people adapt over time to the payoff function: a theory of learning. My account has only described human performance and compared it to predicted performance by the model (see model results of model 4B). However, no formal theory or model of these learning patterns was given.

Previous research has proposed theories of learning for multitasking settings, see for example Erev and Gopher (1999) and

[20] In addition, if eye-tracking data was collected on top of the already rich log-files, the speed at events could be registered by the software would diminish.

Salvucci & Taatgen (2011). However, these theories are not yet at a level of sophistication such that they can directly be applied to the current context. In particular, it is unclear at what level of granularity feedback on performance is mentally processed, and how experience with one strategy is generalized to other strategies. Insights from hierarchical reinforcement learning might prove valuable here. This novel method allows for a model to learn both the utility of small consistent action units, while at the same time learning the utility of larger units (e.g., strategies) that are formed out of these smaller units (Botvinick, 2008).

Finally, a mathematical or analytical model might complement the approach taken here, by offering insights into the structural (or mathematical) ways in which constraints influence performance. As the current model uses Monte-Carlo sampling, the series of simulations might (hypothetically, due to sampling) not converge with the predictions that a purely mathematical model offer. A mathematical model will therefore be developed in Chapter 5.

4.7.4. Possible improvements of the experiment

The modeling analysis suggested that participants did not apply optimal strategies consistently on every trial. If we assume that the model is correct, this discrepancy might be due to several shortcomings in the experiment. Some of these were already discussed before. Below I discuss these shortcomings in more detail together with possible solutions.

First, participants only performed a relatively short series of trials, which might not be enough to adapt performance optimally to the condition at hand. Indeed, in Figure 4.23 for example it can be seen that participant 112 in the high noise, small radius condition with payoff function B adapted performance in the first condition over time, but did not yet reach "optimal performance", despite the trend line going there. The same pattern is present for participant 209 (see Figure 4.26).

An increase in the number of trials would allow for more time to adopt an optimal strategy. At the same time, more might be needed than an increase in exposure, as participants might not be willing to explore alternative strategies. Again, the analysis of learning supports this statement, as it demonstrated that some participants (e.g., participant 106 in Figure 4.22 and participant 202 in Figure 4.24) tended to stick with strategies that achieved good payoff values in the past, both within and between conditions (a strategy transfer effect, Poulton, 1982). This "exploitation" of good strategies seems reasonable, given that in the study all dual-task trials contributed to the remuneration of the participant. Although extensive "exploration" of alternative strategies might have led participants to find optimal strategies, it might come with occasional losses and the exploration might not leave enough trials to then exploit the newly discovered strategies. Participants might therefore be satisficing performance (Simon, 1956).

A workaround for this problem would be to split the experiment in two phases: an explicit exploration phase in which performance does not add to payment (but feedback is provided), and an explicit exploitation phase in which performance does contribute to payoff (similarly to the way Reinforcement Learning models often have an exploration and exploitation phase, e.g., Sutton & Barto, 1998). Participants might be more willing to explore if there are no risks involved.

To guarantee experience with all, or a substantial number of, strategies a no-choice/choice paradigm might be used (Brumby et al., 2011; Howes, Duggan, Kalidindi, Tseng, & Lewis, in prep.; MacLean, Barnard, & Wilson, 1985; Siegler & Lemaire, 1997; Walsh & Anderson, 2009). In such a paradigm, participants first are forced to apply specific strategies to solve a problem (no-choice phase), after which they are free to choose a strategy that they think will be best, given their experience (choice phase). This method has the advantage that all participants experience all conditions in an equal number of trials. However, as a down-side, this method might not scale to a situation in the real world, where people do experience different alternatives, but not necessarily all alternatives.

A strategy-transfer effect can also be avoided by running a complete between-subjects experiment. However, this will increase the number of participants required, and might limit the number of interactions that can be tested (e.g., between different payoff and different task conditions).

The experiment might also have suffered from floor and ceiling effects on performance. Most importantly, the string of digits that needed to be typed had a fixed number of digits (20 per trial). Once a participant paid one visit to the digit window, the set of available strategies for the next visit was reduced (similar to e.g., Gray et al., 2006). This effect can be trimmed down by making the string of digits longer (possibly, infinite). It is to be expected that in a situation like this participants adopt more stable strategies, in which a roughly similar number of digits is typed on every visit to the typing window.

Another component that might have reduced participants' ability to optimize performance is that the task environment is noisy, and therefore difficult to predict, especially in the high noise condition. Increased noise in the movement of the cursor makes it harder to predict the movement of the cursor. Moreover, as the performance on the tracking task is a component of the payoff function, having more noise in the movement of the cursor can also lead to more variability in the experienced payoff scores per trial (i.e., on some trials a strong penalty might be applied, on other trials a similar strategy might not lead to a penalty at all). Taken together, these effects might make it harder for the participant to identify optimal performance.

These effects of noise on strategy selection have been noted before in dual-task settings (Kopecky, 2008). Even richer accounts of the effect of noise on performance have been given in for example the work by Maloney and colleagues, and others on the effect of motor noise on rapid movements (e.g., Maloney, Trommershauser, & Landy, 2007; Trommershauser et al., 2003a, 2003b, 2008) and work by Bogacz and colleagues on the effect of noise in simple speeded two-choice decision tasks (e.g., Bogacz, 2007; Bogacz et al., 2006; Bogacz, Hu et al., 2010; Bogacz, Wagenmakers et al., 2010). See also a discussion in Tseng et al. (submitted).

One way to work around this problem is by having more experimental control over the deviation of the cursor. For example by updating the position of the cursor less frequently, but with a higher mean value and lower noise value. This was used in a Bachelor's project by one of my students (discussed below). In addition, the effect of noise on ability to optimize could be investigated as a separate research question.

In the current experiment, the payoff functions all followed the same general structure (see Equations 4.1 - 4.3). It is to be expected that optima change when the payoff functions have a completely different form - for example when the tracking penalty is based on absolute distance from center of the target instead of time outside of a target area. A related issue is that the differences in payoff between strategies might not have been large enough to encourage participants to do their best. For example, although the payment rate in the study by Hornof et al. (2010) was very similar to the rate in my experiment (with them: 10 dollars per hour), they had a participant perform their task over multiple days and provided bonuses of around 11 dollars per day. Participants might have felt more committed to perform successfully given these larger amounts.

A final, very different criticism might be that people are not at all able to adapt to a payoff function, as they might not be able to represent scores internally (e.g., Chater & Vlaev, 2011; Vlaev et al., 2011). Although this theory can be posited as being on the opposite end of the spectrum to the cognitively bounded rational analysis hypothesis (Howes et al., 2009), the two theories are probably not that far apart. For example, both theories argue that context effects are very important, see also (Kurniawan et al., 2010; Vlaev, Chater, & Stewart, 2007). Moreover, my experiments did not require participants to *assess* an internal representation of the value of a strategy. Rather, the payoff function made the value of strategies explicit (cf. Howes et al., 2009), and in that way alleviated the problems that would arise if one assumes that people can not assess such values by themselves (Chater & Vlaev, 2011; Vlaev et al., 2011).

4.7.5. Studies by students that have looked into some of the above issues

To address some of the limitations on my studies, two students that I co-supervised have run studies that avoided some of these issues, using slightly modified versions of the task environment introduced. First, George Farmer ran a study in which the following changes were made: payoff function was manipulated within subjects, payoff scores had a wider range of values, payoff functions had a slightly different nature, trials lasted longer, and the string of digits was infinite to avoid floor and ceiling effects (Farmer, 2010; Farmer et al., 2011).

Similar to the findings here, participants in Famer's study adapted their performance to the payoff function at hand (Farmer, 2010; Farmer et al., 2011), although there were also some hints at strategy transfer effects. Similar to the findings here, it was found that people had a harder time in achieving optimum performance when the noise in the cursor movement was high (noise was manipulated in a similar way to the manipulation here).

To further address the issues of variability in feedback, transfer of strategies, and increased experience, Shuo Yu ran a study for her BSc dissertation (Yu, 2011). In Shuo's experiment, participants again had to type in an infinite string of digits[21]. The movement of the cursor was

[21] The string of digits was slightly structured: the participant could only see 24 digits at a time, structured in four rows of six digits each

tightly controlled. It moved for a fixed distance (with a very small noise component added on top) for every digit that was typed, such that the moment at which it crossed the target boundary became more predictable. Shuo manipulated feedback (given after every visit to the typing window) between participants. One group received feedback on the maximum cursor deviation. The other group received feedback from a payoff function. Due to the better predictability of the cursor movement, the payoff values were also less variable, and thereby easier to predict. As a result of these manipulations, participants in the payoff condition performed very consistent. A subsequent modeling analysis that I conducted (not reported) also demonstrated that the majority of participants in the payoff condition performed optimal: they achieved the maximum score, given their personal characteristics.

4.8. Conclusion

In conclusion, the work in this Chapter provided support for the idea that people adapt their multitasking performance to three factors: task constraints (noise, radius), cognitive constraints and individual differences in performance (e.g., typing speed), and payoff function (a formal way of capturing objective). The modeling analysis suggests that people adapt their performance in such a way as to try and maximize the payoff value. This is not to say that performance was optimal on every trial. Several explanations have been given for this. Some are related to the learning process (e.g., strategy transfer and exploitation of successful strategies), others might have to do with the difficulty of the task (e.g., the noise in the feedback). In the next Chapter a mathematical model will be developed to give a more formal explanation of how constraints exactly influence performance, and on the limitations of using payoff as a way to influence human performance.

4.9. Appendix to Chapter 4

Table 4.A.1: Parameter values for the typing model component in model 4B (for the eventually selected model) for all participants in payoff condition B

Parameter	**Estimated using data from**	**Payoff condition B, Participant number**											
Participant Number		**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**	**111**	**112**
mean interkeypress interval (msec)	single-task	309	198	394	470	317	384	286	264	184	184	394	382
mean nr errors per trial	single-task	0.3	1	0.45	0.65	0.8	0.2	1	0.5	0.2	2.2	0.6	0.25
mean post-error slowing time (msec)	single-task	67	203	67	105	-9	-15	6	41	57	210	-33	91

Table 4.A.2: Parameter values for the typing model component in model 4B (for the eventually selected model) for all participants in payoff condition C

Parameter	Estimated using data from	Payoff condition C, Participant number											
		201	**202**	**203**	**204**	**205**	**206**	**207**	**208**	**209**	**210**	**211**	**212**
mean interkeypress interval (msec)	single-task	388	255	276	224	405	443	403	451	211	259	226	290
mean nr errors per trial	single-task	0.4	0.35	0.4	0.85	2.75	0.4	0.15	0.5	1.6	0.35	0.35	0.9
mean post-error slowing time (msec)	single-task	617	448	-57	93	8	248	376	-15	87	-45	91	173

Table 4.A.3: Summary of statistical effects in experiment 4B
(when Interkeypress interval is not part of the ANOVA)

		Dependent variable					
		Total trial time	**Maximum cursor deviation**	**Total time cursor outside target**	**Maximum nr of digits per visit**	**Mean visit time, typing window**	**Mean visit time, tracking window**
Independent variable	**Payoff function (P)**		**	.	***	***	
	Noise (N)	***	***	***	***	***	***
	Radius (R)	***	***	***	***	***	***
	Interkeypr. interval (I)		**	.	.	*	
	P x N						
	P x R	*	*	**			
	N x R						

.:   .05 < p <= .10;
*:   .01 < p < .05;
**: .001 < p < .01;
***: p <= .001

Chapter 5. A Reflection on Ideal Payoff Manipulations in a Tracking-while-Typing Scenario through Mathematical Modeling

Abstract

In this Chapter I investigate whether and under what conditions "ideal payoff manipulations" can be performed in dual-task settings that are similar to the tracking-while-typing setting from Chapter 4. That is, when can payoff functions be applied in such a way that changes in the payoff function coincide with changes in the predicted optimum strategy? A mathematical model of a tracking task is developed following principles from Pascal's triangle. The model is then extended to two simple dual-task scenarios. Results show that payoff functions scale effects that are present in the performance data of the individual tasks. As such, they can be used to emphasize the effects that underlying constraints (by the task, cognitive architecture, and individual skills) have on performance. This can be useful to distinguish predictions of competing theories, and to explore the effects of different constraints on performance. The contribution of this Chapter to the literature is that it provides a mathematical demonstration of how constraints and payoff function systematically influence the shape of the payoff curve and the uniqueness of optimal strategies within a specific task environment of interest. Limitations and implications are discussed.

5.1. Introduction

In Chapters 3 and 4 of this thesis, I demonstrated how the predicted optimal strategies for interleaving two tasks shift with changes in the task constraints, cognitive constraints (e.g., switch costs), individual differences in skill level, and payoff function. People were reasonably good at adapting their performance to these factors. However, what are the limitations to such an approach? In this Chapter, I will describe a mathematical model of the tracking-while-typing task of Chapter 4. This model is used to explore whether payoff functions can, in principle, be used to make any arbitrary strategy optimal. If this is the case, then the method is potentially very powerful, as it can then be used to see whether people's performance can be moved away from any 'default' ways (Salvucci & Taatgen, 2011) of interleaving tasks that they might have. By moving performance away from its 'default' state, a wider range of human performance can be explored.

To investigate this, I will use mathematical models to see how performance varies with a change in constraints, and how payoff functions can then be applied to change the utility of different strategies. No comparison will be made with human data, as the empirical question of whether people perform optimally (e.g., as explored in Chapter 4) is a different one from the core question of this Chapter: can payoff functions be used, in principle, to make any arbitrary interleaving strategy optimal in a dual-task setting? I will call this an ideal payoff manipulation.

The remainder of this Chapter is structured as follows. First, I will define what an ideal payoff manipulation is. Then, a mathematical model of a simple tracking task will be developed using Pascal's triangle. This is followed by development of two simple dual-task models that incorporate this tracking model. These models are used to explore whether ideal payoff manipulations can, in principle, be accomplished in a dual-task scenario. The contribution to the literature, implications, and limitations of the work are discussed in the general discussion.

5.2. Ideal payoff curves and ideal payoff manipulations

In this section I will introduce definitions of a payoff function, a payoff curve, an ideal payoff curve, and an ideal payoff manipulation. To start, a payoff function is a function that translates *performance* on a task, or in the case of a multitask environment performance on multiple tasks, into a *single, general, explicit, and objective currency* in a *consistent* manner. In Chapter 4 I used financial incentives as currency, but in other domains this could be another quantity, such as food, time, or accuracy. As a rule-of-thumb, the output of the function should be meaningful to the participant. For example, the payoff value should increase if and only if the performance value of the task improves.

Participants can use the output of the payoff function to identify how well they are performing, by comparing payoff values across trials and strategies. Similarly, a model can be used to compare the payoff values of different strategies (as done in Chapter 4). In this way, a *payoff curve* can be generated that shows how payoff fluctuates as a function of the applied strategy. An "ideal" payoff curve has four properties:

1. It has one global maximum.

2. It has no local maxima other than the global maximum.

3. The set of strategies that have a payoff value close to the maximum is small[22], and these strategies are very similar to the strategy that achieved the global maximum value.

4. The distribution of possible payoff values is consistent and narrow, such that the mean payoff value of a strategy is representative of the distribution of values.

Figure 5.1 shows an example of an ideal payoff curve. The black line is the curve, which depicts mean payoff score per strategy. The grey dots depict the distribution of values. Note that this curve is a polynomial curve that approaches the shape of a parabola. However, a parabola shape is not a prerequisite for an ideal payoff curve.

What happens if one or more of the criteria are not met? First, if there are multiple strategies that achieve a global maximum value, then the participant might adopt any of these strategies to confirm to a typical instruction to maximize payoff. Which of the "optimum" strategies is adopted might differ between and within participants, and can make it hard to identify consistent performance in an experiment. This is especially the case when the strategies are not similar in nature. For example in a dual-task context, if they require very different interleaving patterns, such as focusing completely on "task A" or focusing completely on "task B".

[22] What constitutes a "small" set, and what values are considered "close" to the optimum depends on the distribution of payoff values across the strategy space, and on the maximum and minimum payoff values. This will differ from one context to the next.

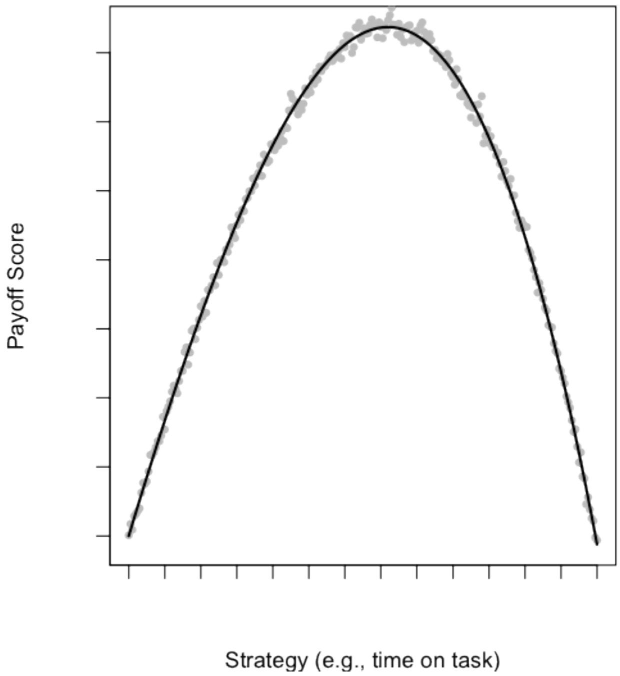

Figure 5.1: Example of an ideal payoff curve.
One global (and local) maximum, with a clear peak, and little variance.

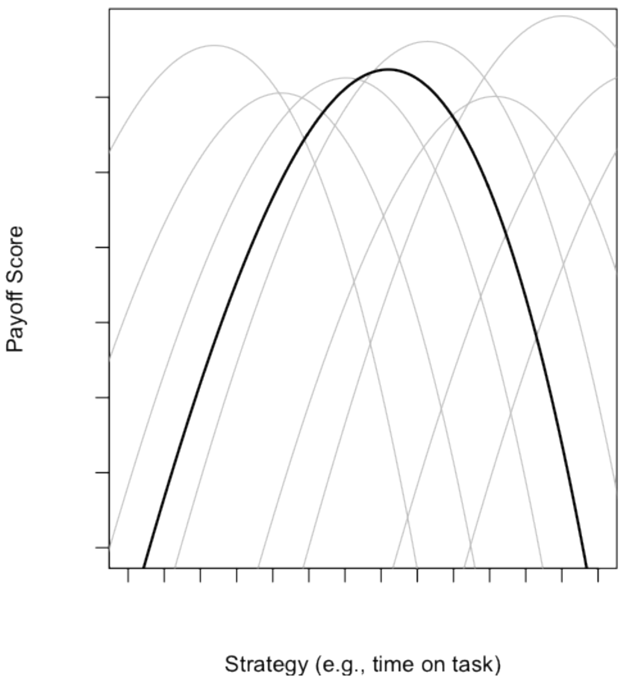

Figure 5.2: Example of an ideal payoff manipulation.
With a change in payoff function, different strategies become the optimal strategy (different grey curves).

Second, if there are one or more local maxima other than the global maximum, participants might get stuck at such local maxima. Again, this will not help in identifying whether people can, in principle, achieve optimum performance.

Third, if there are many similar, but slightly different, strategies that achieve very similar payoff values as the optimum value, then the same caveat holds as when there are multiple optima: There might not be consistent performance within and between participants. A noteworthy case is when the shape of the payoff curve has diminishing returns: if the rate of increase in payoff between two suboptimal strategies that precede the 'optimal' strategy is higher than the rate of increase between the second suboptimum and the optimal strategy. Here, people might settle on the strategy where the diminishing returns start, as predicted by foraging theories of cognition (Payne et al., 2007; Pirolli, 2007 ; Pirolli & Card, 1999).

Fourth, if the distribution of possible values for a given strategy is not consistent and narrow, then any sample set of values that the participant experiences might not be representative of the mean payoff value of those strategies. Therefore, it might again be hard to distinguish the optimum strategy from others, and to adapt performance accordingly.

If payoff is successful in manipulating performance, then in an ideal setting this can be used to make every strategy "optimal" for at least one payoff function. Schematically, this is represented in Figure 5.2 for a situation where only the mean values per strategy are plotted. The black curve is the same payoff curve as shown in Figure 5.1. Next to this curve, multiple alternative curves, generated by hypothetical alternative payoff functions, are plotted in grey. The situation in Figure

5.2 is considered an *ideal payoff manipulation*, because of two properties:

1. Each curve is consistent with the definition of an ideal payoff curve (assuming the variance in the data is also following the ideal pattern).

2. Across curves, each strategy is the optimum of at least one ideal payoff curve. That is, through a manipulation of payoff function, any arbitrary strategy can be made optimal.

Note that with a manipulation of payoff function, the shape of the curve might change in various ways. First, by definition, the *location* of the optimum strategy can change: a horizontal translation (as applied to generate Figure 5.2). Second, the (maximum, minimum, and distribution of) payoff *values* might change, for example due to a vertical translation (as applied to generate Figure 5.2). Finally, the *shape* of the curves might change in other ways, for example it might stretch or shrink due to multiplication factors (see also Chapter 6 for a further discussion on these types of manipulations).

With this definition of an ideal payoff manipulation, an investigation can now be made whether such manipulations are possible in a dual-task setting that is similar to the tracking-while-typing task environment of Chapter 4. However, I will start with the description of a single-task model. The single-task model is a simple mathematical model of the movement of a cursor that moves around freely until the model (or participant) decides to stop the movement. The objective is to let the cursor move around freely until the point where it crosses a fixed boundary (similar to the tracking task in

Chapter 4). Pascal's triangle is used in its development. The model is used to explore whether an ideal payoff manipulation is possible.

5.3. Model 5A: Modeling first-order movement of a cursor using Pascal's triangle

5.3.1. Model development

The first model simulates a first-order one-dimensional moving cursor (a random walk). It is assumed that the cursor is in the middle of an equal scale, and moves in discrete steps, with one movement per tick (or time stamp). At any point in time, the cursor can either move left or right from its current position, with equal probabilities for moving each way.

The predicted movement of this cursor can be captured using Pascal's triangle, as is done in Table 5.1. At tick 0 (first row of Table 5.1), the cursor is at position 0. At tick 1 (second row of Table 5.1), the cursor has moved either left or right, to a distance of -1 or +1. The probability of each option is 1/2. On tick 2 (third row), the cursor could have moved left or right again from its position at tick 1. If it was at

Table 5.1: Pascal's triangle
The triangle be used to see how likely (cell) a certain position (horizontal axis) is, given the number of updates or ticks (vertical).

tick										Distance								
(time)	sum	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
0	1									1								
1	2								1	-	1							
2	4							1	-	2	-	1						
3	8						1	-	3	-	3	-	1					
4	16					1	-	4	-	6	-	4	-	1				
5	32				1	-	5	-	10	-	10	-	5	-	1			
6	64			1	-	6	-	15	-	20	-	15	-	6	-	1		
7	128		1	-	7	-	21	-	35	-	35	-	21	-	7	-	1	
8	256	1	-	8	-	28	-	56	-	70	-	56	-	28	-	8	-	1

position -1, it could have gone to -2 or 0. From position +1 it could have moved to position 0 or + 2. That is, on this tick there is a 1/4 chance to be on position -2, a 2/4 chance of being on position 0, and a 1/4 chance of being on position +2. In a similar way, the likelihood of being at any specific position at any tick can be calculated. This probability (e.g., being at position -4 on tick 6) can also be read from the table, by taking the value of a particular cell that corresponds to the required distance (here: the value for position -4 on tick 6), and dividing it by the value underneath the "sum" column (resulting here in: 6/64).

For the simulations in this Chapter, Pascal's triangle was generated once with a depth of 200 ticks, and then used to give estimates of positions of the cursor. These estimates were turned into two performance metrics: (1) the expected distance of the cursor, given the number of ticks that has passed, and (2) the probability that the cursor crossed any arbitrary boundary, given the number of ticks that has passed.

Once these performance metrics were identified, the model was used to simulate a hypothetical task, which is also part of the later dual-task models (models 5B and 5C). In this task, the cursor moves around freely, 'out of sight' of the model (or simulated participant). A known target boundary is set, which the cursor should not cross. The task of the model is to let the cursor move around as long as possible, without crossing the target. That is, it should decide when to stop the cursor movement.

The model gains points for every tick during which the cursor was within (or on top of) the boundary. The model loses points for every tick during which the cursor is outside of the target area. With the model for this task, gains and losses can be systematically varied to explore whether an ideal payoff manipulation can be accomplished.

5.3.2. Model results

Performance Metrics

Two performance metrics were investigated: estimated distance at each tick, and the probability that the cursor crossed a boundary, given the number of ticks that passed. In both cases, the absolute distance was considered, because taking the direction of the position (i.e., positive or negative) into account would result in a mean of 0 due to symmetries in the table.

Figure 5.3 plots the expected absolute distance as a function of the number of ticks passed in a log-log plot, with lines depicting the expected distance mean (black line), median (grey), first quantile (green), and third quantile values (red). The mean distance approximates a linear function in the log-log plot, which indicates that the underlying function of expected movement is (approximately) a power function. The data points follow a zig-zag pattern. This is because the positions of the cursor goes back and forth between odd and even distances (with even distances including position 0, which lowers the mean). The zig-zag pattern has the strongest relative weight during the first series of time ticks.

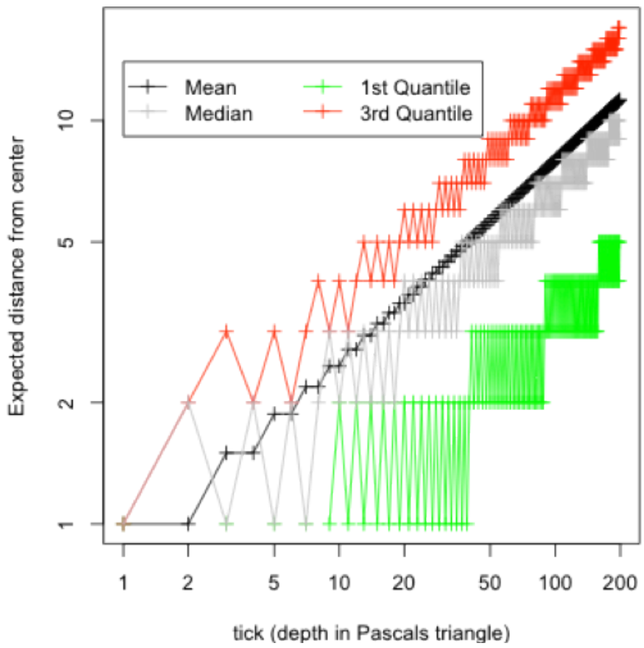

Figure 5.3: Expected distance from center given the number of ticks.
Plotted on a log-log scale. The figure plots the mean and median distance together with the first and third quantile.

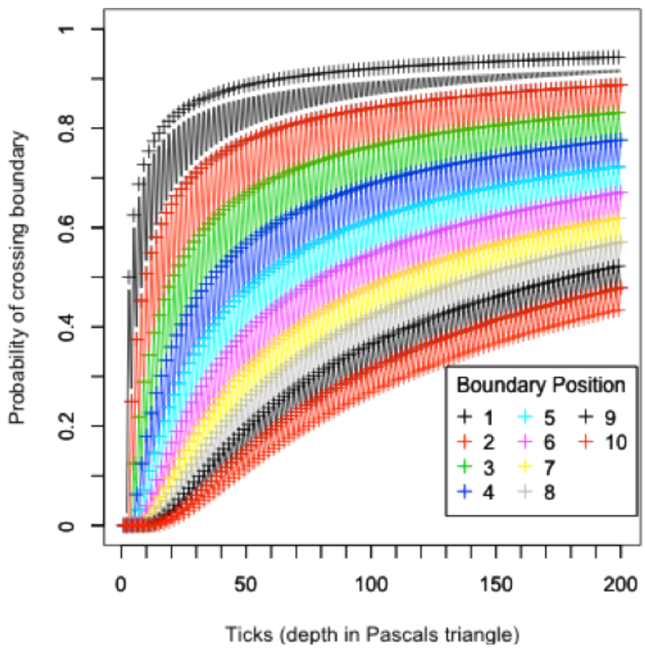

Figure 5.4: Probability that the cursor crosses a boundary (see legend) given the number of elapsed ticks

The triangle can also be used to get an exact estimate of the probability that the cursor crossed a particular distance from the center after a specific number of ticks. This can be done by looking up what proportion of mass is in cells that are beyond the intended distance. As an example, Figure 5.4 plots how the probability (vertical axis) that the cursor crossed a specific boundary changes as a function of the number of ticks (horizontal axis) for boundaries between 1 and 10 (colored lines, see legend). When generating this Figure, it was assumed that being on top of the boundary counted as not having crossed it.

Three effects are visible. First, the probability of crossing the boundary increases with time. Second, the probability increases more quickly when the boundary is close to center. Third, the probability follows a zig-zag pattern, similar to the zig-zag in the expected distance data (see Figure 5.3).

Are ideal payoff manipulations possible?

To introduce payoff, the basic model was applied to a task. In this task, the objective was to let the cursor move around freely as long as possible without having the cursor move outside of a specific target area, or boundary. In the task, the model gained points on each sample where the cursor was inside the target, and lost points on each sample where the cursor was outside of the target.

The strategy that the model applied determined how long the cursor should drift, and therefore how many samples were used to calculate overall payoff. The longer the cursor was left drifting, the more samples contributed to payoff. As long as the cursor stayed inside the target, the gains therefore accumulated. However, the longer the cursor drifted, the more likely it became that the cursor moved outside of the target area. This might result in a strong accumulation of losses. How should the model then balance these gains and losses?

To assess this, a payoff function was introduced. In contrast to a Monte Carlo simulation, Pascal's triangle allowed a calculation of the *exact* probability at each moment in time that the cursor was inside or outside of the target (i.e., see Figure 5.4). The expected payoff value on any arbitrary tick *i* (or sample) for this tracking task (labeled *tr* in the equations below) can then be calculated using the probabilities and the known loss and gain components, as done in equation 5.1:

$$Payoff_{i,tr} = \left(1 - P_i\left(outside\right)\right) * Gain_{tr} + P_i(outside) * Loss_{tr} \quad \textbf{(Equation 5.1)}$$

As can be seen, the expected payoff on any given sample *i* is a linear function of the probability that the cursor is outside or inside the target area.

Recall that in the task environment the model had to let the cursor drift for as long as it wanted. On every sample it gained a payoff value, using Equation 5.1. Over the series of samples *i* up until the point where the cursor is stopped (*N*), the total expected payoff value is then an accumulation of the values experienced on every individual sample *i*, and can be expressed as follows:

$$CumulativePayoff_{tr,1\to N} = \sum_{i=1}^{N}\left[\left(1 - P_{i,tr}\left(outside\right)\right) * Gain_{tr} + P_{i,tr}\left(outside\right) * Loss_{tr}\right]$$

(Equation 5.2)

Note that the cumulative payoff increased with the number of samples, but only if the gain component is relatively high in comparison to the loss component and if the probability of being inside the target remains relatively high.

Did this function allow for an ideal payoff manipulation? That is, can changes in the weights of gains and losses in Equation 5.2 change the payoff function such that across payoff functions different strategies are optimal? This was explored below. To start, one probability curve from Figure 5.4 was chosen. A boundary of three was chosen, as this curve covered most of the probability space [0-1] within 200 ticks. With this curve, payoff was systematically varied as follows. Gains were varied between 0 and 150, in increments of 5. Losses were varied between 0 and -150 in increments of 5. This gave a total of 961 (31 x 31) different payoff functions.

As an example, Figure 5.5 shows 4 of the resulting 961 payoff curves and provides their associated equations (see legend). Note how each function only has one global maximum (highlighted with an open circle), and how the location of the maxima differs between curves. Also note how the steepness of the points to the left side of the maximum increases with an increases in the gain component. For example, compare the gain components for the light grey line (gain of 90) and the second lightest grey curve (gain of 40). The line with component 90 is more than twice as steep. Similarly, the steepness of the curve on the right side of the global maximum increases with an increase of the loss component. For example, compare the steep black line (loss of -50) with the second lightest grey line (loss of -15).

Certain global maxima in the payoff curve could be achieved by multiple payoff functions. For example, Figure 5.6 plots four figures, that have the best strategy at four different locations, namely at 3, 9, 49, or 99 ticks. Within each plot, different lines show the payoff curves for

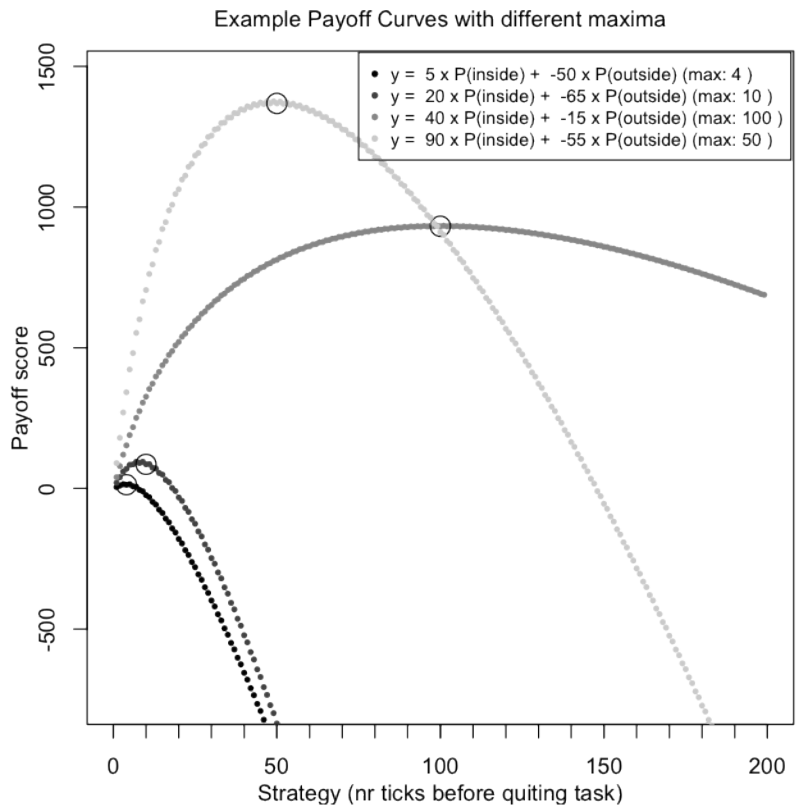

Figure 5.5: Four example payoff curves
with different global maxima and their associated equations.

the different manipulations that have the peak at this location. Not all curves have an ideal shape. In particular, some have a broad plateau where strategies that are close to the optimum strategy have scores that are similar to the maximum score. Similar to the findings for the curves in Figure 5.5, the curves that had the most clearly defined peaks had relatively large magnitudes for the gain and loss components.

Taken together, this suggests that ideal payoff curves can be generated in this task environment. Can the task environment also be used for an ideal payoff *manipulation*? That is, can different strategies be made the optimal strategy solely through a manipulation of the payoff function? To investigate this, I inspected where global maxima occurred across the different payoff functions.

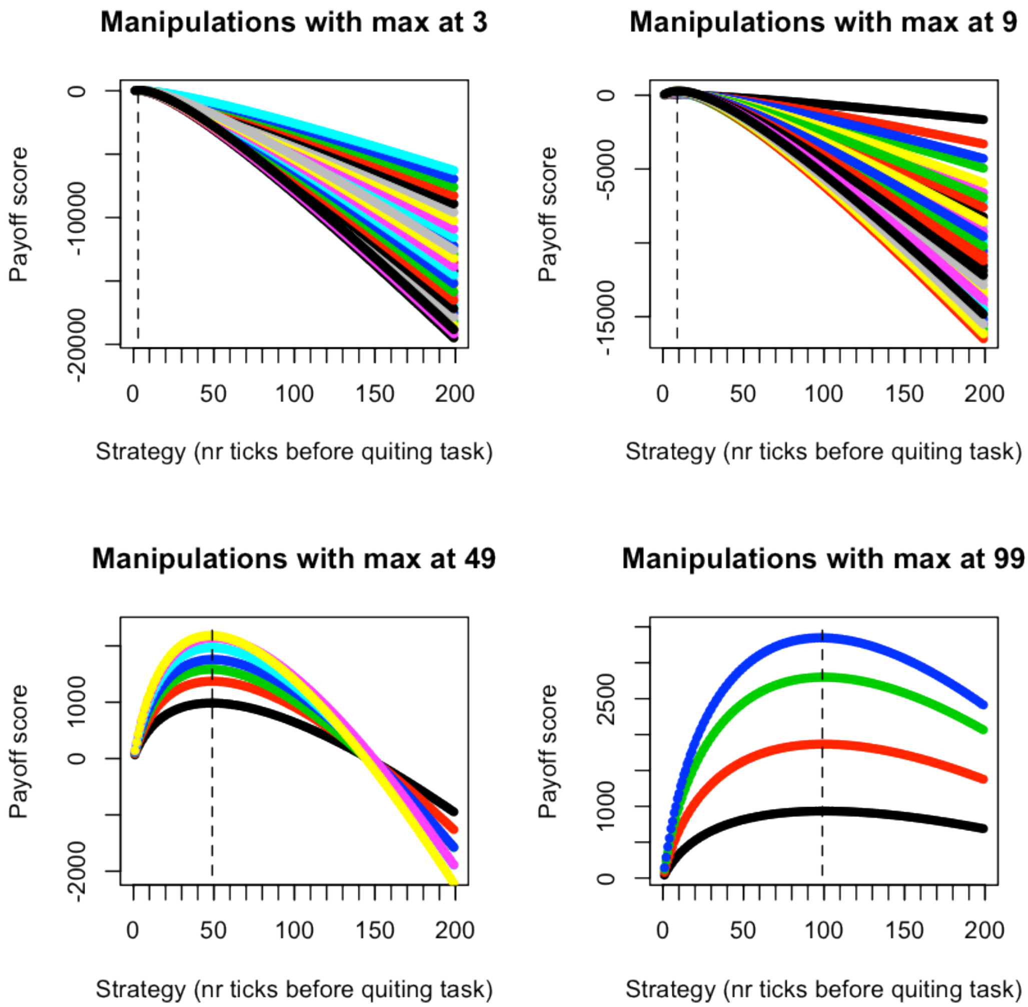

Figure 5.6: Payoff curves (different colors) with optima at different locations namely at 4 (top-left), 10 (top-right), 50 (bottom-left), & 100 (bottom-right).

The majority of payoff curves had only one global maximum. More than one global maximum occurred when the gain component was 0. In this case, the maxima were for strategies 1 till 3 (where the cursor is still inside). When the loss component was also 0, all points had an equal score of 0. Apart from these structural points, there were two maxima (located at strategies 3 and 5) when the gain component was 15 and the loss component was -145.

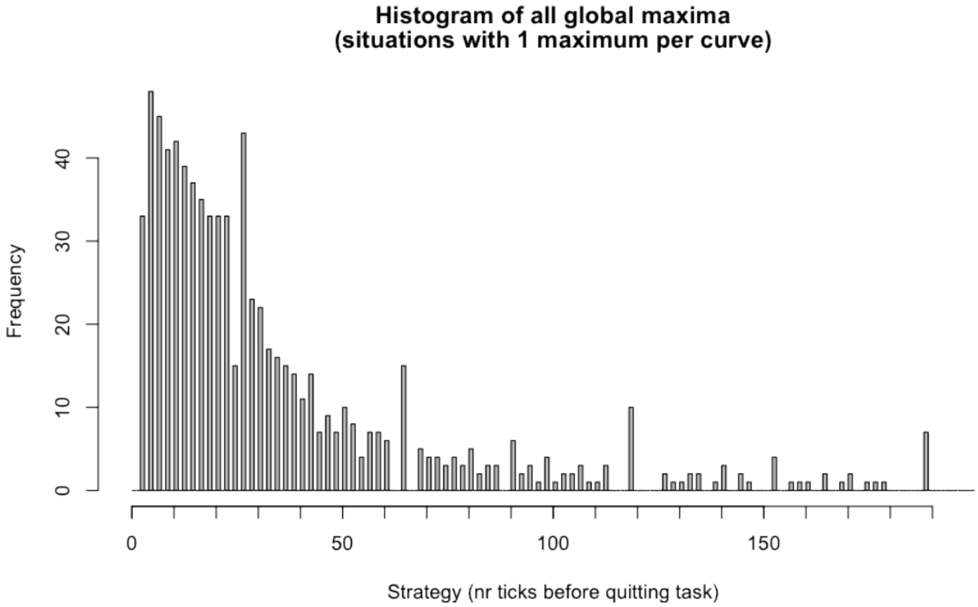

Figure 5.7: Histogram of the location of global maxima in the payoff curve when payoff function is manipulated. The model had the boundary at 3.

In all other cases, there was only one global maximum per curve. Figure 5.7 plots a histogram of the location of the global maxima of these functions. Strategies in which the maximum was at positions 199 or 200 were omitted from the Figure, as these strategies could be an artifact of the fact that no strategies larger than 200 ticks were investigated.

The Figure suggests that an ideal payoff manipulation is possible in this task environment, as different payoff functions have different optimum strategies, and this set of optimum strategies covers most of the strategy space. The first possible global optimum is at strategy 3. Note that 3 is also the position of the target boundary, and the last tick at which the cursor is guaranteed to be inside the target area under all circumstances. After this, only odd strategies are ever optimal. This is due to the zig-zag pattern of the probability function (see Figure 5.4), and having the position of the boundary at an odd distance. The probability of being inside is (for 3 or more ticks) always higher on odd ticks than on the two surrounding even ticks.[23]

Note that as the number of ticks increases, there are more possible positions (or states) at which the cursor can be. This is because the number of non-empty cells in Pascal's triangle increases with an increase in time (see Table 5.1). Because of this wider set of possible positions of the cursor, the exact location of the cursor on any given trial or simulation will get harder to predict with an increase in time. This will inherently also lead to variation, or noise, in the experienced payoff signal for this value. Therefore, when a curve is generated that predicts that the optimum strategy involves long waiting (i.e., a strategy that is on the right side of Figures 5.6 and 5.7), this curve might not conform to the definition of an ideal payoff curve (specifically, to point 4 of the definition). This is because the noise in the payoff signal might make it hard to identify the optimum.

5.3.3. Discussion of results

The results of model 5A suggest that an ideal payoff manipulation is possible in this simple task environment. Through a manipulation of the payoff function, different strategies can be made optimal (see Figure 5.7). Moreover, each of the resulting curves is, or approximates, the shape of an ideal payoff curve (i.e., see Figures 5.5 and 5.6).

There are four precautions, however. First, optima can only occur for strategies that wait at least the minimum number of ticks that is required to reach the boundary (in the case of the example in this

[23] For situations with an even boundary, the optima are at even peaks.

model, 3 ticks for a boundary at 3). Second, strategies that require long waiting will have more variation (or noise) in their payoff signal, and might therefore not comply with all characteristics of an ideal payoff curve. Third, if the boundary is at an odd position, optima can only occur for strategies that stop at an odd number of ticks, and for an even boundary, the optima are only for strategies that stop at an even number of ticks. Fourth, the global maxima will be most pronounced in situations where relatively large magnitudes for gains and losses are used (because in these cases the effects of gains and losses are scaled more steeply, see Figure 5.5 for examples).

These results are mostly straightforward, and very useful. This is because they provide exact predictions of probabilities and mean values, due to the mathematical grounding of the model. This contrasts with the model results in previous Chapters. Those models used Monte Carlo sampling to approximate probabilities. Although the results of Monte Carlo sampling will approach the true mean when a large enough sample is used, there is no guarantee that this will be reached.

In the above simulations, "tick" was not mapped to a concrete time unit (e.g., seconds or milliseconds). This can be done by defining how long each tick lasts. For example, in the model of Chapter 4 this was set to 25 milliseconds. In a similar fashion, the "distance" parameter in the model can be rescaled to correspond to a specific number of pixels. These mappings are not made here, as the resulting models are only linear translations of the simpler model. However, they are needed to see what part of the strategy space is *feasible* for participants. For example, if one tick corresponds to 25 milliseconds, then human participants will not be able to apply strategies that require quick responses, as the time for a basic perception-action cycle is larger than one tick.

5.4. Model 5B: Introducing a second task with continuous payoff

Can ideal payoff manipulations also be performed in a dual-task scenario? To explore this, model 5A will be extended with a secondary task.

5.4.1. Model development

The dual-task environment will only allow the model to attend to one task at a time, similar to models in Chapter 4. The model has to decide when to interleave one task for the other, and will incur a time cost when switching its attention.

One of the tasks is the tracking task as modeled in model 5A. In this task, the cursor moves around randomly whenever it is not attended to. Once the task is attended to, the model can actively correct the position of the cursor. The time needed for these corrections depends on three factors. First, on the distance at which the cursor is found. For simplicity, I assume that the cursor is always at the mean expected distance, given the number of ticks that have passed since the last active correction (see Figure 5.3). Second, the time to correct depends on whether the random movement of the cursor continues while the model is correcting. For simplicity, I will assume that this is not the case. Third, the time to correct depends on how many distance units are traversed per tick when the model is correcting the movement; I call this the "correction distance per tick". The higher the correction distance per tick, the faster the cursor can be brought back to center. If correction distance per tick is not exactly equal to 1, then the model cannot be moved exactly to center from all distances. For example, if correction distance per tick equals 2, and the initial distance equals 3, then the model can either move the cursor to +1 or -1. For simplicity, I built in the assumption that the cursor always ends at center position (i.e., 0) after correction. The time needed for this correction is set as the ceiling value of the start distance divided by the correction distance per tick. For example, to travel from a distance of 17 with a correction distance of 4 units per tick takes 5 ticks (instead of the exact value of 4.25 ticks). It is assumed that the cursor is moved an equal distance on every tick.

The secondary task that is used in this model is the simplest task possible, as it does not require any active action. As soon as the task window of this task is open (and therefore, the tracking task is closed), the model gains points. When the window is closed, the model loses points. For ease of reference, this task is therefore called the "passive attention task".

Within a given trial time (or total number of ticks) the model iterates through four distinct steps. The first step is attendance of the passive attention task. It is assumed that the model always starts with this task. This task is attended for a specific number of ticks, varied between strategies. In the second step, the model switches from the passive attention task to the tracking task (e.g., similar to the button press on the joystick in Chapter 4). In the model, this is set with a single parameter. In the third step, the model actively corrects the position of the cursor (as described above). In an experimental set-up, a resumption cost might be incurred here, where the participant first needs to identify the current position of the cursor before deciding how to correct it. For simplicity, the model does not incur such resumption costs. In the fourth, and final step, the model switches attention back to the passive attention task. After this step, the model continues with step 1: attending the passive attention task. The model iterates through these steps until the trial finishes. Note that the window of a task is open when that task is attended to, but is also open while the model switches attention to the other task (except for during the last tick of this switch).

In the simulation below, the following model parameters were set. The total ticks per trial was 50, the boundary of the tracking task was set at 3, the switch costs (both ways) were either 2 or 10, and the correction distance per tick was set at either 1 or 4. A wider set of parameters was explored initially, but did not provide additional insights about the possibilities to perform an ideal payoff manipulation in this task environment. In addition, exploring a wide set of parameters was not of interest, as the aim was not to develop an accurate model of human performance.

To calculate the total payoff per trial, the payoff on the tracking task was calculated as before (see Equation 5.2). This value was summed with the payoff of the passive attention (*p.a.*) task. For the passive attention task, the payoff depended on the total number of ticks during which the window of this task was open, T_{open}, in relationship to the total ticks in a trial, T_{trial}:

$$Payoff_{p.a.} = T_{open} * Gains_{p.a.} + \left(T_{trial} - T_{open}\right) * Losses_{p.a.} \qquad \textbf{(Equation 5.3)}$$

5.4.2. Model results

Before introducing different payoff curves, the model was used to explore how changes in the constraints of the model (switch cost, and correction distance per tick) influenced two performance metrics: expected total ticks during which the cursor is inside the target area, and expected total ticks during which the passive attention window is open. Figure 5.8 shows how these two values (vertical axis) vary as a function of strategy (horizontal axis: ticks spent on passive attention task per visit). Between the plots on different rows of the Figure, the values of the correction distance per tick parameter are varied between 1 (top row) and 4 units per tick (bottom row). Between the columns of the Figure, the switch cost is varied between 2 (left column) and 10 ticks (right column).

The expected total ticks that the cursor is inside its target (black dots in Figure 5.8) and the total ticks during which the passive attention task is open (grey dots), in general vary with changes in strategy (horizontal axis), correction distance per tick (rows), and switch costs (columns). The exact variation depends on the parameters at hand and is not too interesting, as the constraints (and therefore results) are not grounded in human data, and therefore do not make predictions about human performance.

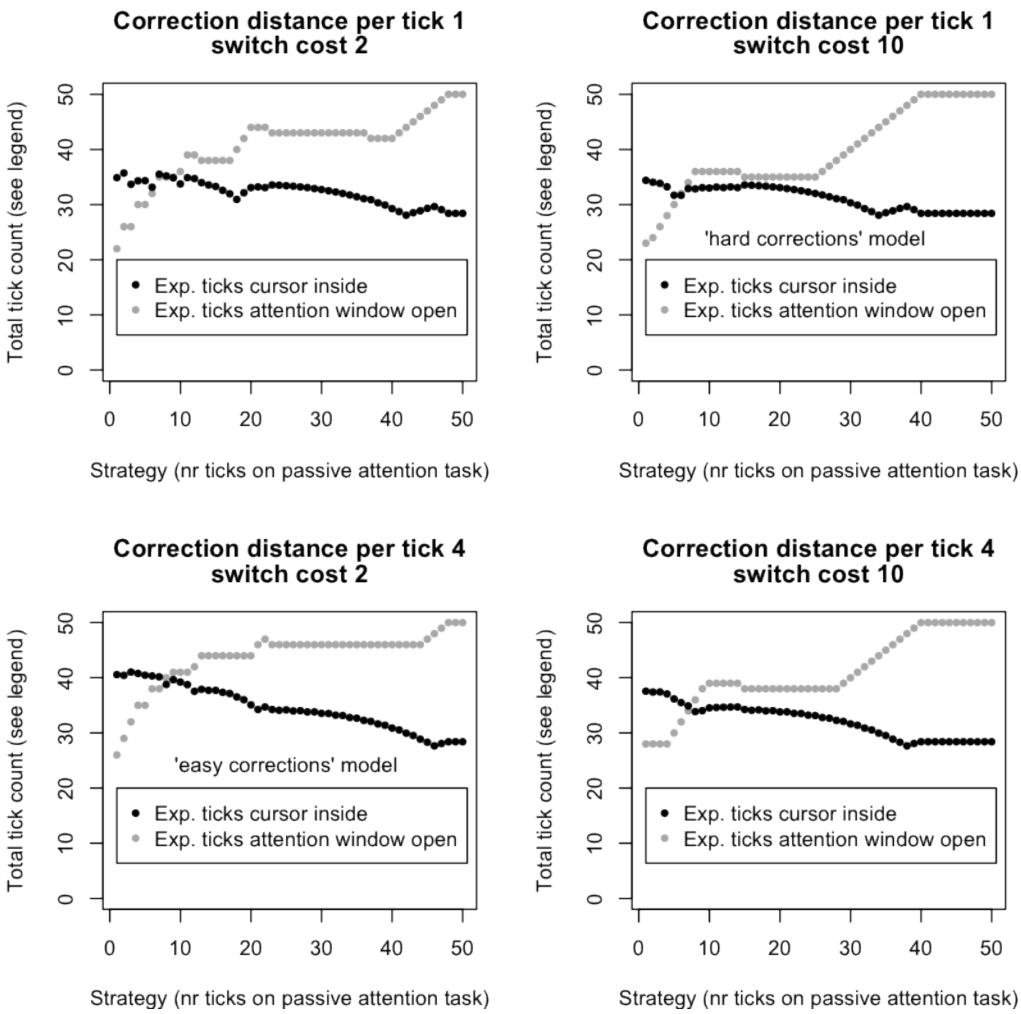

Figure 5.8: Expected total ticks
(vertical axis) that the cursor is inside the target (black), and that the passive attention window is open (grey), given the strategy (horizontal). For different correction distances per tick (rows) and switch costs (columns).

The general pattern is more interesting, because of its transfer to other dual-task settings. In general, as more ticks are not spent on tracking (either due to a strategy, or due to a large switch cost) and as correction of the tracking task is made harder (as set by the correction distance per tick parameter), the total number of ticks that the cursor is expected to be inside its target area reduces. Where this is due to a strategy choice or a switch cost, the total ticks that the passive attention task is open increases[24]. The curves have local maxima and plateaus in them, as different strategies sometimes end up in exactly the same number of visits and (given the switch cost during which the passive attention window is open) in the same number of ticks spent on this task.

Independent of whether this data fits human performance, it can be used to explore whether, in principle, an ideal payoff manipulation is possible. To explore this, two of the four conditions in Figure 5.8 were selected for further analysis. These models were chosen as the shapes of the performance curves (see Figure 5.8) differed the most between these two conditions. One model had a switch cost of 2, and a correction distance per tick of 4 (bottom left), which for ease of reference will be called the "easy corrections" model. The other model had a switch cost of 10, and correction distance per tick of 1 (top right), which for ease of reference will be called the "hard corrections" model.

With these models, the gains and losses that were applied per tick were varied systematically. For the tracking task, gains were varied for all values between 0 and 100 with increments of 5, and losses were varied between -100 and 0 with increments of 25. Gains and losses on the passive attention task were varied in a similar way. This gave a total of 11,025 different payoff functions (21 x 5 x 21 x 5).

Similar to the analysis for model 5A (see Figure 5.7), the locations of the global maxima across the different payoff functions were explored, to see if an ideal payoff manipulation is possible. In both models, the simulations that had all gains and losses set to 0 did not

[24] Note that during a switch from the passive attention task to the tracking task, the passive attention task is open for all but 1 tick.

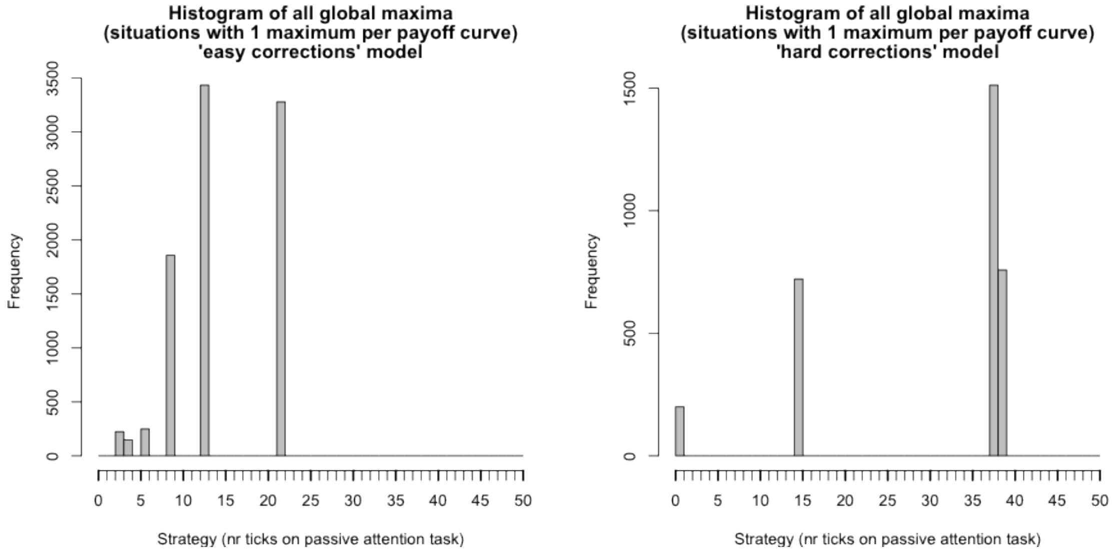

Figure 5.9: Histograms of the location of global maxima in the payoff curve when payoff was varied, for payoff curves with a single maximum. Left shows results for the 'easy corrections' model, right for the 'hard corrections' model.

have any maxima (as all strategies had a payoff value of 0). For the easy corrections model, eight simulations had two peaks (at positions 3 and 4), and 1,824 simulations had three maxima (always at positions 48, 49, and 50). However, the majority of simulations (9,192) had only one global maximum.

For the hard correction model, most simulations had 11 global maxima per payoff curve (7,833 simulations, all with the maxima at positions 40-50), and therefore did not conform to the characteristics of an ideal payoff curve. The remaining 3,191 simulations had only one global maximum.

A precondition for an ideal payoff manipulation is that there is only one global maximum per payoff curve. Therefore, the subset of payoff curves that fulfilled this criterion was further explored. Figure 5.9 shows a histogram of the locations of the global maxima for the easy corrections model (left) and the hard corrections model (right). Two aspects are immediately apparent in these histograms. First, the maxima are *not* distributed across the entire strategy space, suggesting that an ideal payoff manipulation is not possible. Second, the location of the maxima differs between the two task environments.

At closer inspection, the position of the maxima can be related to the results that were plotted in Figure 5.8. Global maxima in the payoff curve (i.e., peaks in the histograms in Figure 5.9) only occur for strategies that have a *local* maximum in the expected total ticks for the tracking task (black line, Figure 5.8) or the passive attention task (grey dots, Figure 5.8) at that point. In hindsight, this is perhaps not surprising, as the payoff manipulation is used to scale these graphs.

These results show that an ideal payoff manipulation was not possible. However, did the functions that have only one global maximum at least adhere to the properties of an ideal payoff curve? That is, was the global maximum clearly identifiable, and were there no other local maxima? Rather than exploring this for all possible curves, a subset of curves was selected. Figure 5.10 shows plots for the easy corrections model, when the optimum strategy was at position 3, 6, 13, or 22. Per plot, ten example curves that share the same optimum strategy (indicated with the dashed line) are plotted. It can be seen that the curves do not have the shape of an "ideal curve", because the plots have local maxima, and the areas surrounding the optimum strategy often have payoff scores that are very similar to the maximum score.

It is also clear, that some of the curves mostly emphasize performance on the tracking task, as evident by a similarity in shape with the performance plot in Figure 5.8. For example, the plots that have local maxima at positions 3 and 6 in Figure 5.10 (the Figures on

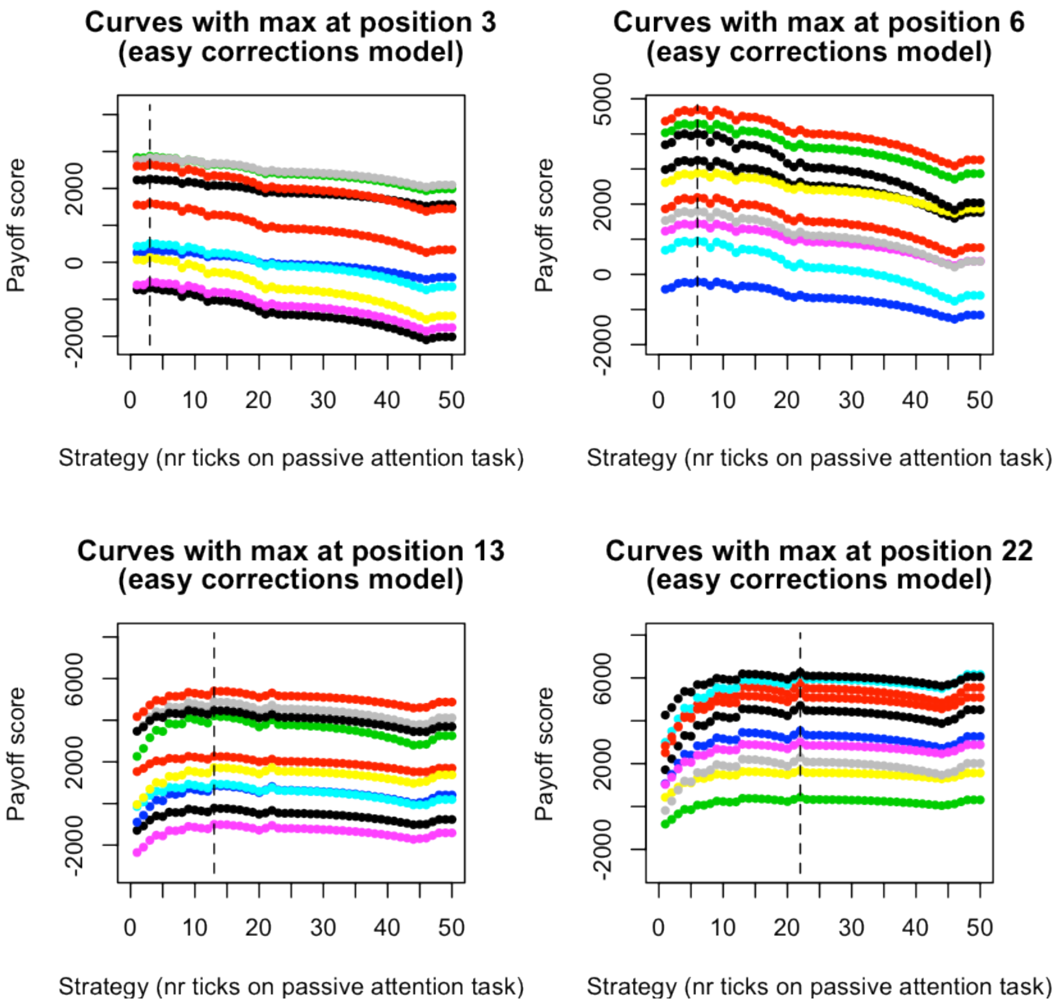

Figure 5.10: Example of payoff curves with a single global maximum.
Per plot, the maximum is at a fixed place. Maxima differ between plots.

the top left and top right of Figure 5.10) follow the shape of the performance curve of the tracking task (i.e., the black dots in the bottom left plot in Figure 5.8). Other plots, for example with maxima at positions 13 and 22 (bottom row of Figure 5.10) show more similarities with the performance curve of the passive attention task (i.e., the grey dots in the bottom left plot in Figure 5.8).

Did these results only emerge because of the chosen sample of gain and loss values, or because of the choice for linear scaling? To investigate this, I applied six other payoff manipulations to the data of the easy corrections model. In one manipulation, a random sample of 10 payoff values between 0 and 100 was applied to each of the values for gains and losses on each task (i.e., 10 x 10 x 10 x 10 = 10,000 unique combinations), to explore effects of a sample bias. In a second simulation, the effect of the magnitude of the gains and losses was explored by multiplying the initial chosen values with a factor of 100.

In four additional explorations, the original sample of gains and losses was used, but non-linear functions (logarithmic and exponential functions) were applied. In these functions, the gains and losses components were either multiplied with the tick count within the function (e.g., log[gainsWeight x ticks + 1] and exp[gainsWeight x ticks+1]), or used as a factor to multiply the result from the function (e.g., gainsWeight x log[ticks + 1]).

The results are not plotted here, because they did not provide significant additional insights. The maxima occurred at similar locations as before, although the frequencies with which they occurred were of course different (with some maxima not showing up in some simulations). When inspecting individual payoff curves, only the curves that used a logarithmic or exponential function had very characteristic shapes. In line with their mathematical characteristics, the payoff curves of the logarithmic functions were relatively flat, whereas curves of the exponential functions were relatively well pronounced, with (in general) few or no local maxima, and a clear global maximum (see also Chapter 6 for a more detailed exploration of how the underlying performance functions influences the shape of the payoff curve).

5.4.3. Discussion of results

From this data it can be concluded that, at least for this task environment, the payoff manipulations can only be used to scale the underlying performance space of the tasks. Global maxima on the payoff curve can only arise at positions where the performance curve has local maxima. The local maxima on the performance curve themselves arise because of the constraints that come from the task and cognition. For example, in the current model the switch cost and the correction distance per tick for the tracking task both provided severe limits on how long each visit to the tracking task would take, and therefore limited the amount of time that the passive attention task could be attended to. This created local maxima in the performance curve. In that sense, task constraints and cognitive constraints are hard constraints (Gray et al., 2006; Howes & Young, 1997) that severely narrow down the plausible strategy space, and rule out some strategies from ever becoming optimal under those constraints (see also theoretical framework in Chapter 2).

5.5. Model 5C: Typing and tracking

In the previous model the secondary task was kept very simple. Might the results change when this task involves more steps, such as in the typing task from Chapter 4? To investigate this, I explored how model performance changed when the secondary task was changed into a typing task. To incorporate this change, it was assumed that the model could only type 'digits' when the passive attention window was open.

Moreover, it was assumed that it took a couple of time units to type a digit (varied between simulations).

To translate performance into a payoff score, the model gained points for the tracking task as before. For the typing task, the model gained a fixed number of points (varied between simulations) for every digit it typed. Note that with this payoff function, the objective of the model was not to type a fixed set of digits as fast as possible (as in Chapter 4), but to type as many digits as it could, given the interleaving strategy (similar to Farmer, 2010; Farmer et al., 2011).

5.5.1. Model development

The basic structure of the model was kept similar to the "easy corrections" model of model 5B. The secondary task model was refined to include a typing component. While the secondary task window was open, the model could type 'digits'. It was assumed that within a simulation the typing time was similar across all digits. The model was used to explore what the effect of different typing times was on performance, by varying this between simulations with integer values between one and six ticks per typed digit (i.e., giving six different simulations).

The model was first used to generate a performance space that characterized how many digits the model could type during a trial, given the strategy and the required time for typing. The performance space for the tracking task was generated as before (see black dots in Figure 5.8). These two performance spaces were then translated into a payoff space as follows. For the tracking task, the model gained and lost points following the equations in model 5B. The gain component was varied between 0 and 50, in steps of 10, and the loss component was varied between 0 and -50 in steps of -10. For the typing task, the model gained points for every digit it typed. The gains per digit was varied between 0 and 100 in steps of 10. This gave a total of 396 different payoff functions.

5.5.2. Model results and discussion

The results did not produce any novel insights and will therefore be discussed only briefly. The performance space of the tracking task varied as before (see Figure 5.8). The performance space of the typing task varied as expected: in general, more digits could be typed per trial if (1) the time needed to type a digit was reduced, or (2) more digits were typed per visit to the typing window (i.e., as a function of strategy). Given that different strategies resulted in a similar number of total digits typed, the performance space for the typing task had local maxima and plateaus (similar to how these were present in the performance space of the passive attention task in model 5B, see Figure 5.8).

When generating payoff curves, the set of strategies that could be made optimal due to a variation in payoff function was again limited. Across payoff functions, only a limited set of strategies could ever become a local maximum. Similar to the results for model 5B, these strategies were strategies where the performance space of either the typing task or the tracking task had a local maximum. Moreover, when inspecting the underlying payoff curves, these curves did not adhere to the properties of an ideal payoff curve, because they had local maxima, and mostly did not have well defined peaks (similar to results in Figure 5.10).

5.6. General Discussion

5.6.1. Summary of this Chapter

In this Chapter a mathematical model was developed to explore whether an ideal payoff manipulation can, in principle, be applied to dual-task settings that contain a tracking and a typing task. An ideal payoff manipulation allows for different strategies to emerge as global optimum with a change in payoff function (for a detailed definition, see section 5.2). A model of random movement that used Pascal's triangle (model 5A) was used as a stepping-stone to explore performance in two dual-task scenarios (models 5B and 5C). Results suggest that ideal payoff manipulations are not possible if the underlying performance metrics of the tasks have local maxima at positions other than the global maximum. This was the case in the dual-task scenarios that were explored in this Chapter. In such cases, the payoff function can only be used to scale the performance space and to turn local maxima into global maxima.

The local maxima in the performance space, and the resulting potential global maxima on the payoff curves, are influenced by the specifications of task and cognitive constraints. If the specification of the constraints changes (for example, the size of the switch costs), then the location of the local maxima in the performance space also changes (e.g., see Figure 5.8). This subsequently changes the possible locations for global maxima on the payoff curves (e.g., see Figure 5.9).

Because task constraints and cognitive constraints have this strong effect on the performance curves and payoff curves, they can be seen as hard constraints (Gray et al., 2006; Howes & Young, 1997), that severely narrow down the space of plausible strategies, and rule out some strategies from ever becoming optimal. This is in line with the description of the theory of cognitively bounded rational analysis (Howes et al., 2009) and its representation in Venn diagrams in Chapter 2 (see Figure 2.2 in Chapter 2).

5.6.2. Implications of results

The simulations in this Chapter illustrated how the position of global optima within a payoff curve is influenced by the weights of the components of the payoff function (e.g., settings for gains and losses), but most importantly by the characteristics of the performance space and the underlying assumptions about constraints. Are payoff functions then of any use if they only emphasize effects that are already present in the data (i.e., in the performance space) without incorporation of a payoff function?

The answer is yes. In Chapter 4 some of the reasons for preferring the use of payoff functions were already discussed. They offer participants a clear objective, which avoids subjective interpretation of priorities. They can also express performance on two tasks consistently in a single, objective ‘currency’. Therefore, they do not require trade-offs to be made between performance metrics on tasks with different currencies. The analysis in this Chapter added the following advantage to this list: payoff functions can be used to pronounce the effects that constraints have on performance more strongly. That is, they can help to turn local maxima in the performance space into global maxima on the payoff curve. As an example, see the differences between the curves with different global maxima in Figure 5.10.

For the models in this Chapter, the exact value of the constraints was not too important, as model data was not compared with human data. However, in most other cases, specifying the constraints correctly is of the upmost importance. Yet, the 'correct' values of the constraints might not always be known upfront. This poses two challenges. First, it poses a challenge for experimental design. If the constraints are not well specified before running the experiment (e.g., if not all sources of individual differences are known), then it might be hard to craft a payoff function that will emphasize the effects of the (unknown) constraints. This will make it hard to apply a useful payoff function in the experiment.

Second, once the experiment is run, different combinations of constraints might make different predictions about 'optimal' behavior. Whether one can then use the models to conclude whether people are optimal depends on the chosen constraints at hand. When I used my models for these purposes (see Chapter 3 and 4), I tried to minimize the risk of inferring incorrect conclusions, by grounding the values of constraints in measured (single-task) data.

An alternative to that approach is to use the model in an inverse way to what I have been doing. That is, rather than specifying constraints and then investigating whether people perform optimally (as I did in Chapter 4), one might assume optimality (i.e., assume that people conform to instructions), and then explore how predictions of optimality change with a change in constraints. The model of which the optimum strategies provide the best fit to human performance might then be chosen as giving the best description of the constraints. This approach is very similar to that taken in “optimal experimental design” research (e.g., for recent work see Nelson, McKenzie, Cottrell, & Sejnowski, 2010). In that work, human data is used to rule out parameter combinations that make inconsistent predictions.

Similar in spirit, but slightly different, payoff functions can be used to explore performance of two or more competing theories that differ in their assumptions about underlying constraints. In these cases, the payoff function can be used to emphasize performance differences between model predictions that otherwise might be hard to detect. This is the approach taken by Howes, Lewis, and Vera (2009), who use payoff manipulations to investigate which theories better explain performance in a Psychological Refractory Period (PRP) setting: those that assume serial processing of information (e.g., Anderson, Taatgen, & Byrne, 2005; Byrne & Anderson, 2001), or those that assume parallel processing (e.g., Meyer & Kieras, 1997a, 1997b).

Such a method can be of interest to further expand on the theories and results in Chapters 3 and 4 of this thesis. For example, building on the results of Chapter 3, payoff functions can be used to increase or diminish the speed-accuracy trade-offs that interleaving at ‘natural breakpoints’ offers. We have started to explore such avenues in a project with a Bachelors student (Yu, 2011, see also short discussion in Chapter 4).

Building on the work in Chapter 4, payoff functions can also be used to highlight more strongly the individual differences in performance in multitasking settings, and to explore their underlying causes. Again, individual differences in skills put constraints on the performance space, and might create local maxima in this space. A payoff function can be used to better pronounce these maxima on the payoff curve. Moreover, if two competing theories exist about the source of individual differences, payoff functions might be used to better discriminate the impact that these differences in constraints predict for dual-task performance.

5.6.3. Limitations

A limitation of this work is that only a small set of constraints, payoff magnitudes (or weights), and payoff functions (mostly linear, but also logarithmic and exponential) were explored. For this Chapter, the small set of constraints, payoff magnitudes, and payoff functions was enough to demonstrate that ideal payoff manipulations cannot always be made in the dual-task scenarios of interest in this thesis: tracking-while-typing (Chapter 4) and dialing-while-driving (Chapter 3). In Chapter 6, I will study how the underlying function of performance (e.g., linear, exponential, or logarithmic) influences performance in general in dual-task scenarios.

Another limitation of the work in this Chapter was that only mean performance (e.g., mean drift of the cursor in model 5A) was used for the majority of analyses. However, as was already noted in the discussion of the results in Chapter 4, the distribution of values and the sample of values that participants experience might differ from the mean values. This can then influence performance (see also the definition of an ideal payoff curve in section 5.2).

The payoff functions in this Chapter translated data from a performance space into payoff values. The effect that strategies had on payoff values was therefore indirect. Strategies influenced performance, and performance then determined the payoff value. In theory, one can also craft functions that provide a direct mapping between strategies and payoff values. Moreover, these functions can be shaped such that there is a unique maximum. For example, a block function or a band filter can be used to give a positive value when the desired strategy (or, in the case of a band filter, strategies) is applied, and a negative (or zero) value otherwise.

Such functions were not used in this Chapter, as they might not map easily to a participant's subjective experience. For example, imagine a function for the experiment in Chapter 4 that directly maps the time spent away from the tracking task to a payoff value, such that spending two seconds away is the optimum. This function would give a positive value independent of whether the cursor crossed the target boundary and independent of the number of digits typed. However, participants might be aware of these two components and try to relate them to their experienced payoff. This might result in an inconsistent mapping between performance and payoff. For example, they might be achieving a high payoff value in situations where relatively few digits were typed or where the cursor crossed the target boundary. Situations like this might make it difficult to learn how to adapt performance.

5.6.4. Contribution to the literature: a mathematical analysis of how constraints and payoff function systematically influence performance and payoff curves

This Chapter contributed to the literature by providing a concrete demonstration of how constraints systematically influence performance and payoff curves. Results suggested that payoff functions can not be used to make any arbitrary strategy optimal in the chosen task environment. That is, no ideal payoff manipulation was possible. It was found that manipulations of payoff can change which parts of the performance curve are emphasized more strongly. These effects in the performance curve arise themselves from constraints that are imposed on performance by the task, the cognitive architecture, and individual differences in skill.

Based on these results, a path for further research that involves payoff functions was suggested. Payoff functions can be used to compare the performance predictions from different theories, that have different constraints. In such efforts, payoff functions can be used to pronounce the differences in performance predictions (that result from constraints) more strongly in the payoff space. By comparing these predictions with human performance, a decision can be made about what set of constraints best describes human performance (i.e., as done in Howes et al., 2009). This approach requires a clear, concrete specification of competing theories (or models) before applying the payoff function, and it assumes that human performance is cognitively bounded rational. Using predictions of different models to distinguish different theories is similar in spirit to research on "optimal experimental design" (e.g., Nelson et al., 2010).

5.7. Conclusion

The application of payoff functions in experimental and model settings is valuable. They can transform performance spaces in such a way that only one global maximum emerges. However, the position of the global maximum depends on the payoff function at hand, and depends even more strongly on the constraints that are imposed on performance by the task environment and cognition. This systematic relationship can be exploited to investigate what constraints underlie human performance.

Chapter 6. Exploring which General Types of Tasks allow for Ideal Payoff Manipulations

Abstract

In this Chapter I investigate through mathematical analysis what types of tasks have most chance of allowing for optimal payoff manipulations. I distinguish tasks based on what the resulting output of such a task looks like: a linear, linear step, exponential, power, or logarithmic function. Gain and loss functions are then applied to explore which combinations of tasks allow for ideal payoff manipulations. Results show that settings that combine linear tasks with logarithmic tasks can produce payoff curves that are most like an ideal payoff curve. However, ideal payoff manipulations can only be approximated. The model was developed at a higher level of abstraction than what was done in other Chapters and did not consider moment-to-moment interleaving and related concepts and constraints such as switch and resumption costs. Rather, it focused on the general characteristics of the task before such more complicated factors are introduced. The contribution to the literature, implications, and limitations are discussed.

6.1. Introduction

In Chapter 5 I found through mathematical modeling that ideal payoff manipulations are perhaps not possible in a typing and tracking task paradigm. However, such a task combination is only one example setting out of a large set of task combinations. Here, I use mathematical modeling to explore whether ideal payoff manipulations are, in principle, possible for five different task types that differ in how performance increases as a function of investing additional time: linear, linear step, exponential, power, and logarithmic tasks. Using the mathematical definitions of the performance functions of these tasks, I will then explore what predicted combined performance might look like for any set of two tasks. At first, performance will be explored without considering payoff. This will be followed by the introduction of a payoff function that scales the gains to be won on the task. Then, I will also introduce a loss function, which will decrease the gain values based on the time not spent on the task. Based on the results of these manipulations, I will then look into more detail into a specific task combination (linear with logarithmic) to see whether ideal payoff manipulations are possible.

Before considering the details of this work I shall recap the definitions of an ideal payoff curve and ideal payoff manipulations that were introduced in Chapter 5 (see there for more details). A payoff function translates performance on one or multiple tasks into a single objective score. A payoff curve captures how payoff varies as a function of strategy. In an ideal payoff curve, the curve has one strategy that can be considered optimal (because it achieves on average the greatest payoff value) and this strategy is well defined within the curve. An ideal payoff manipulation is possible when different strategies can be made optimal, solely through a manipulation of the payoff function, while preserving all properties of an ideal payoff curve.

In this Chapter, I will explore for what general types of tasks ideal payoff manipulations are possible. This exploration will be conducted at a different level of abstraction than was used before. In previous Chapters, I explored performance for a *specific* task setting with inherent constraints. In addition, I also explored how performance changed as a function of the moment-to-moment interleaving pattern (i.e., the strategy alternative). Here, I will not specify the detailed constraints of the task nor the moment-to-moment interleaving patterns. That is, I will assume that there are no switch costs and no resumption costs. Rather, for a defined time period of 100 discrete ticks, I will explore what performance looks like when different *proportions* of time are allocated to a task, without exploring exactly how the proportions are allocated. For example, a strategy of spending 50 ticks on task A and 50 ticks on task B can be instantiated in several ways (e.g., spending 1, 25, or 50 ticks on a task per visit). I will not explore these finer grained patterns here. Rather, it will be assumed that the performance during individual visits to a task will accumulate such that summed performance across visits is similar to a situation where a longer visit was spent on this task (i.e., in the example, a visit of 50 ticks).

This analysis can give a first insight into how performance fluctuates given (1) the general pattern of dividing attention proportionally across tasks (e.g., 80:20, 60:40, or 50:50), (2) the applied performance function (e.g., linear or exponential), and (3) the payoff function. Once the influence of these three general factors have identified under what conditions ideal payoff manipulations might be created, future work can investigate how the moment-to-moment interleaving pattern influences results for specific tasks. In line with the models described in the preceding Chapters, it is to be expected that these will introduce further refinements on the location of optima within the payoff curve.

The remainder of this Chapter is structured as follows. First, I will give formal definitions and examples of the five different task types. Then, I will illustrate with a model how the shape of the payoff curves changes as a result of (1) different task combinations, (2) different gain functions, and (3) different loss functions. Based on the results obtained through these manipulations, I will explore one specific combination (linear with logarithmic) in more detail.

6.2. Five general task types

The starting point of this work is that different classes of tasks have different ways in which performance can be described. I consider five types in this Chapter: linear, linear step, exponential, power, and logarithmic. Table 6.1 summarizes these tasks and gives their mathematical definitions. The simplest definition of such a function is given under the header “minimal form”. The top row of Figures 6.1-6.3 also provides examples of how expected performance (vertical axis) changes for the different task types, as a function of the amount of time spent on the task (horizontal axis). I will now introduce each type and provide examples of each type of task.

The first type of task that I shall consider is a linear task. In such a task, each extra time step invested on the task contributes an equal amount to performance. That is, the rate of gain per time step (from now on: rate of gain) remains constant over time. A simple example is filling a bathtub with water that comes out of a hose. For each time step, the hose will release an equal amount of water. Therefore, with each time step, performance (i.e., how full the bathtub is) will increase at an equal rate.

The second type of task that I shall consider is a linear step function. For step functions, an increase in performance takes multiple time units, a step. In a linear step function, the rate of gain is constant across steps. Step functions can, in principle, be applied to any function. A linear step function is of particular interest here, because it can be used to describe performance on a typing task, such as the one used in Chapter 4 and 5. Each digit takes some time to enter, but once it is typed, a digit brings a user one step closer to having entered the complete string of digits. In Chapter 4 in particular, I have assumed that every correctly typed digit takes an equal amount of time to type, a step.[25] Given that I only explore one step function, I will refer to the linear step function simply as the step function.

For the remaining three types of tasks, and their associated functions, the rate of gain is not constant. For the exponential and power function, the rate of gain increases with every time step. In the

[25] However, this pattern changes when considering typing errors (see Chapter 4) or memory retrievals (see Chapter 3).

exponential function, x is taken as the exponent of the mathematical constant *e* (the base of the natural logarithm). In the power function, a power is applied to x (see also definition in Table 6.1). In this thesis, I will use a power of 2. As a result, the exponential function increases more rapidly than the power function.[26]

Exponential and power functions can be used to describe many phenomena in nature and in economics, for example the spread of a cold virus across the globe, or the increase in the amount in one's savings account due to compound interest. In psychology, exponential functions have been used widely, but with some modifications to the basic forms, to model phenomena ranging from the likelihood of being able to recall items from memory given the amount of practice (for a discussion, see Anderson & Schooler, 1991) to the likelihood that previous experience with actions and their utility will be exploited at the disfavor of exploring new actions (the Boltzmann equation, used in reinforcement learning approaches in for example, Anderson, 2007; Sutton & Barto, 1998).

The last type of task that I explore is a logarithmic task. For these tasks, the rate of gain decreases over time. As a result, some investment of time in a task is beneficial, as the first steps will improve performance rapidly. However, later on, the rate of gain becomes very small, and theoretically it is often not beneficial to invest time after the point of diminishing returns, as discussed for example in Pirolli (2007) and Stephens and Krebs (1986).

---

[26] By introducing a negative constant (for b in the full form equation), an exponential function can decrease over time. Such functions are not explored here, as it is assumed that investing extra time in a task will never lead to worse performance.

There are many examples in biology where logarithmic functions can be used to describe the patterns in which animals gather food (Stephens & Krebs, 1986). In psychology, logarithmic functions have been used to describe information searching behavior (Pirolli, 2007; Pirolli & Card, 1999), for example to model the amount of time that should be invested to find a better hotel when searching for hotel accommodation online. Logarithmic functions have also been used to study problem solving (Payne & Duggan, 2011) and the time that should be invested between multiple information rich tasks such as scanning documents (Duggan & Payne, 2009) and solving puzzles (Payne et al., 2007). In all these studies, the core idea is that there are diminishing returns in terms of the improvement in quality of the outcome (i.e., finding a cheaper hotel, finding another solution to a puzzle) based on the increasing amount of time that is put into the task.

The tracking task of Chapter 4 can be approximated by a logarithmic function. If the cursor is far away from center, initial movements will bring it rapidly back to center (and increase performance). However, once the cursor has approached (or is at) center, performance can hardly be improved. In a similar way, the average drift of the cursor also follows a logarithmic pattern: in the beginning it will on average drift a lot, afterwards the average drift will diminish (see Chapter 5).

Given the basic set of tasks outlined above, I next provide a more concrete mathematical definition of each. Table 6.1 provides a mathematical description of each function. The minimal form provides the simplest form of the equation, which is also used to generate the base forms in Figures 6.1-6.3. For each equation, four coefficients can

Table 6.1: Mathematical form of each function.
Floor means that the division is rounded to the lowest integer value, "exp" means exponential, and "ln" mean natural logarithm.

Type	Minimal form	Scaled form	Full form
Linear	$y = x$	$y = a * x$	$y = a * (b*x+c) + d$
step	$y = \text{floor}(x/step)$	$y = a*\text{floor}(x/step)$	$y = a*\text{floor}((b*x+c) / step)+d$
Exponen.	$y = e^x$	$y = a*e^x$	$y = a*e^{b*x+c} + d$
Power	$y = x^2$	$y = a*x^2$	$y = a*(b*x+c)^2+d$
Logarith.	$y = \ln(x)$	$y = a*\ln(x)$	$y = a*\ln(b*x+c)+d$

be added, as done under the heading "full form" in Table 6.1. I explain each of these coefficients in turn.

First, the coefficient "d" provides the starting level of performance (i.e., the intersection with the y-axis when x equals zero). Higher values indicate having a "head start", whereas negative values denote a "delay" in performance. In mathematical terms, *d* creates a vertical translation. For example, in the earlier bath tub filling example (a linear function), a positive value of *d* can be used to describe the initial level of water in the bath tub.

Second, the coefficient "c" can also capture head start or delay effects. Mathematically, it creates a horizontal translation (i.e., it moves the intersection with the x-axis of the curve, where y equals zero). In psychological terms, it denotes situations where someone starts as if they are *c* time-units ahead (or behind) of normal performance. In the bath tub example, it would describe a situation where someone has filled the bath with the hose for *c* time ticks before starting the new count.

Finally, the coefficients "a" and "b" are scaling functions, which make it easier or harder to achieve a particular performance level within a particular time frame. Psychologically, coefficients b and a can therefore be seen to reflect the difficulty of the task.

Coefficient b is applied within the function. It leads to horizontal scaling, by making performance look as if already b times as many time units have passed. In the bathtub example, it would describe a situation where the width of the hose is b times as wide compared to normal.

Coefficient a is the only scaling function that is applied outside of the regular performance function. Manipulation of a leads to vertical scaling: it makes performance look a times as good as it normally would be rated. In the bathtub example this would correspond for example to a situation where a times as many hoses are used, or the bathtub is made a times as small.

Out of all the four coefficients, coefficient a is the only one that can be considered for a manipulation of payoff. This is because it has two properties: it is outside of the regular function and at the same time it can scale performance to put more weight on certain types of performances. All other coefficients have to do with characteristics of the task itself and of the participant performing them. Although these can also be manipulated in experiments (e.g., as done in Chapter 4 by manipulating the size of a radius), here I am not interested in those effects. Rather, the interest is on how the performance on a task changes as a sole result of manipulations of payoff. I will therefore now explore how manipulations of a changes the payoff curve for combinations of different task types.

6.3. Model 6A: Combining tasks with different performance characteristics

6.3.1. Model development

To explore how different combinations of general task types changes dual-task performance, I developed a mathematical model. As a starting point, I implemented the minimal form of each function (see Table 6.1) and explored how performance changed as a result of the total number of time ticks spent on the task. I then inspected performance for all combinations of tasks in a dual-task setting. For simplicity, ticks were varied between 1 and 100. For the step function, a step of 10 was used. For the power function, an exponent (or power) of 2 was used.

A final consideration revolves around scaling. Without applying some form of normalization procedure to the output of the minimal forms of each equation, performance across tasks would be difficult to compare because of the large differences in the order of magnitude in output values. For instance, the maximum score that could be achieved by each function (i.e., when x equals 100 in Table 6.1) lead to these values: 4.6 (logarithmic), 10 (step), 100 (linear), 10.000 (power), and $2.7 \cdot 10^{43}$ (exponential). To make the outputs of each function more directly comparable, I normalized the results of each function such that the largest value achieved by each was set to 100 points (i.e., by dividing each score by the maximum possible score for that function and then multiplying this with 100).

With these normalized functions in place, I was able to explore the effects of payoff. That is, I explored the effects of applying a factor *a* to the outcome of the basic function (see "scaled form" in Table 6.1). I first explored the effects of different gain factors (i.e., positive *a* values).

This was followed by the introduction of costs for not attending tasks. It was assumed that this cost function followed the same shape as the gains function, but that the scaling factor *a* had a negative value. Based on these explorations, I further investigated manipulations of *a* for a particular combination of tasks (linear with logarithmic) that seemed most promising of achieving ideal payoff manipulations.

6.3.2. Model results and discussion of results

Performance for different combinations of normalized minimal function

The black lines in Figures 6.1-6.3 show how performance changes based on the number of ticks spent on the primary task (horizontal axis) for different primary tasks (columns) and different secondary tasks (rows). Note that the output of each function was normalized such that the best possible output of each function had a value of 100. For performance of the black line in Figures 6.1-6.3, the scaling factor *a* was set to 1 (i.e., there was no additional scaling).

The strongest effects occurred when the primary task was either exponential or a power function (columns 2 and 3 in Figure 6.2). In these cases, the model predicted that focusing attention fully to only one task would achieve the highest payoff, independent of what the secondary task was. It also did not matter which task was focused on – just so long as one of the two got full attention – equal payoff values could be achieved. However, based on the slope of the curves, it is best to focus on the task that was not an exponential or a power function. This is because the slope of these alternative functions (i.e., on the left hand side of the plots in Figure 6.2) are relatively less steep, and therefore involved less risk: strategies that were close to the extreme strategy (i.e., that were close to the strategy that gave full attention to the secondary task) achieved scores that were relatively close to the optimum value. In comparison, strategies that dedicated most, but not all, attention to the power or exponential function (i.e., that were on the right side of the plots in Figure 6.2) had scores that were at least half of the optimal score. That is, these strategies were more "risky" and might in practice not be applied by human participants.

For situations where the primary and secondary task were made up only of linear and linear step functions (rows 2 and 3 of Figure 6.1), the model predicted that there should be no strategy preference. The payoff value of strategies remained relatively constant across strategies. In cases where a step function was involved, strategies that did not initiate new steps without completing them were preferred. For example, in this model the "step" equaled 10, so only strategies that were a multiple of 10 were optimal (i.e., strategies 0, 10, 20, … 100).

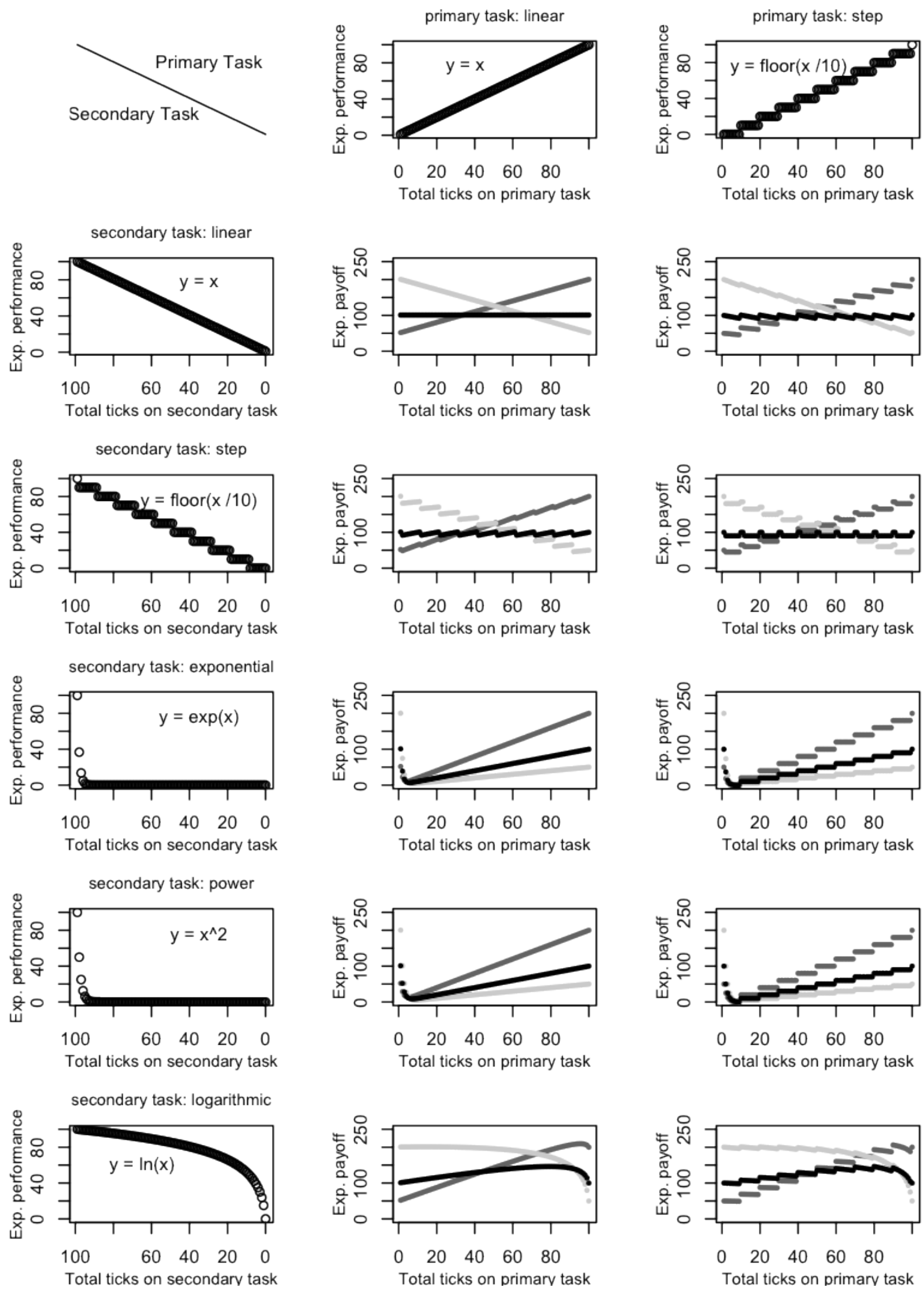

Figure 6.1: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for a linear function (2nd column) and a step function (3rd column).
Functions are normalized. Gain functions were then("a") applied. The black line shows a model with each task having equal weights. The dark grey line has a higher weight (2) for the primary task compared to the secondary task (0.5). The light grey line has the inverse (preferring the secondary task).

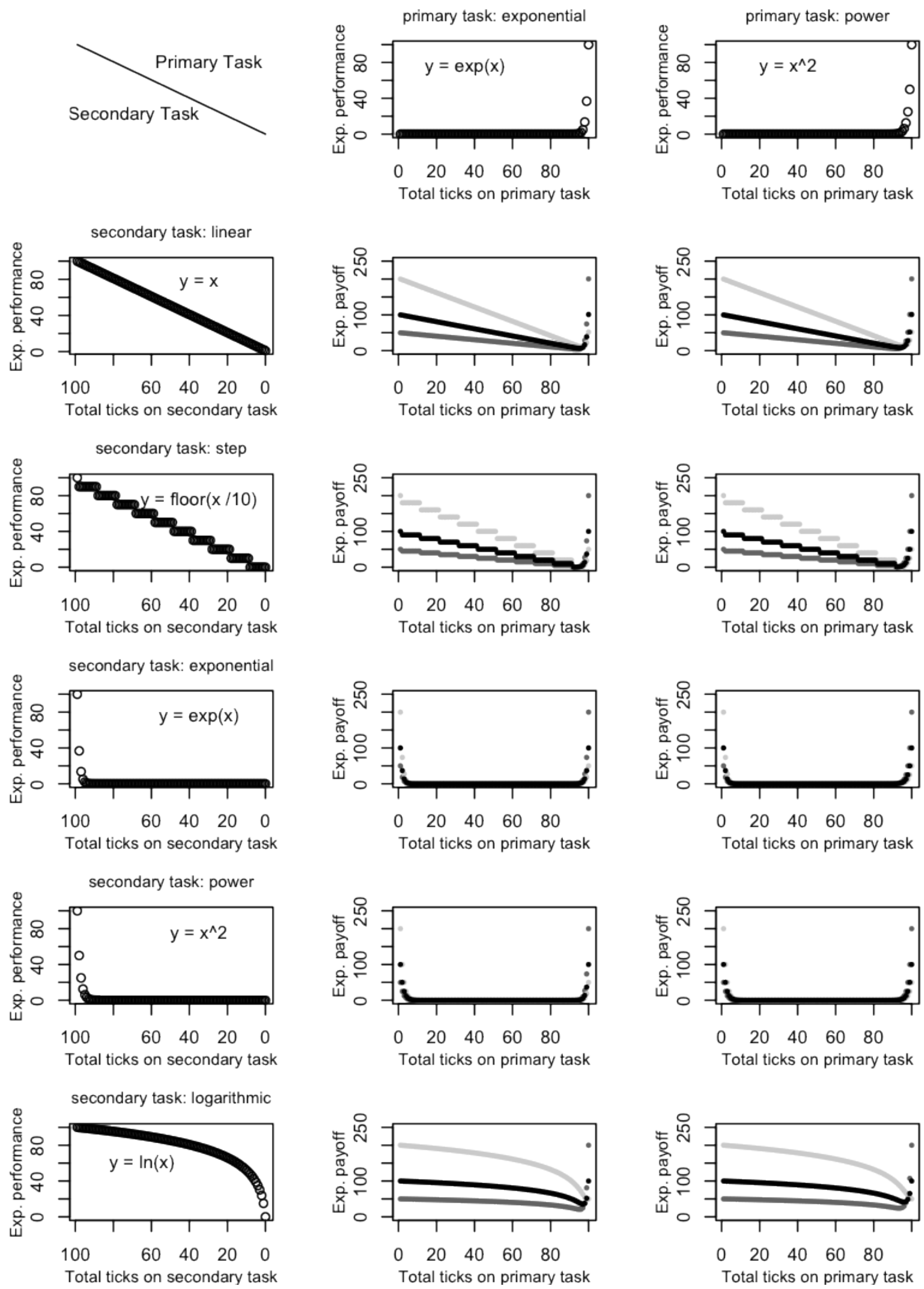

Figure 6.2: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for an exponential function (2nd column) and a power function (3rd row).
Functions are normalized. Gain functions were then("a") applied. The black line shows a model with each task having equal weights. The dark grey line has a higher weight (2) for the primary task compared to the secondary task (0.5). The light grey line has the inverse (preferring the secondary task).

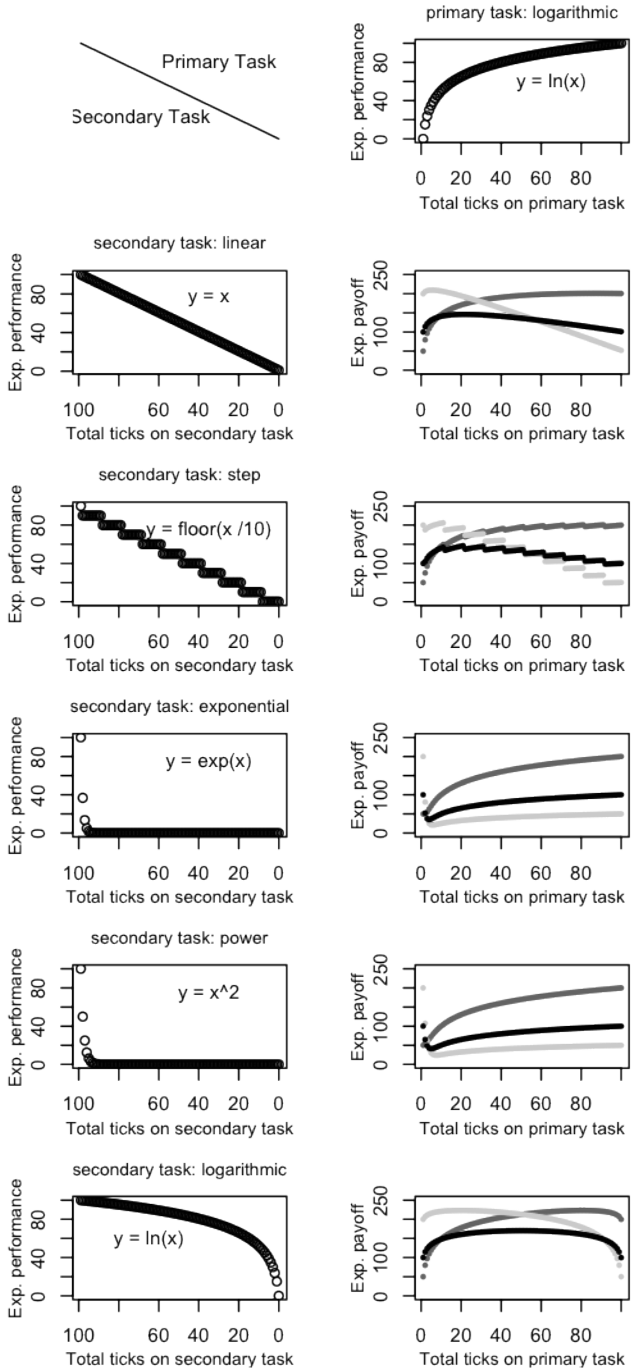

Figure 6.3: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for a logarithmic function (2nd column).

Functions are normalized. Gain functions were then("a") applied. The black line shows a model with each task having equal weights. The dark grey line has a higher weight (2) for the primary task compared to the secondary task (0.5). The light grey line has the inverse (preferring the secondary task).

When one of the tasks was logarithmic, and the other task was linear, step, or logarithmic (i.e., bottom row of Figure 6.1, and rows 2, 3, and 6 of Figure 6.3), the model did predict that there would be unique global maxima. When a (primary or secondary) task was linear or a step function, most attention should go to that task. This is visible in the bottom row of Figure 6.1 (where the logarithmic task was secondary), where the peaks of the black line are on the right side of the plot. In the second and third row of Figure 6.3, where the logarithmic task is the primary task, the peaks are on the left side of the plot. However, in all these cases at least some attention should be given to the logarithmic task, as visible by the fact that the global maximum was near, but not at, the extreme ends of the plot. That is, the model predicts that optimal performance requires a trade-off in the amount of time invested on each task.

This can be explained due to the nature of the logarithmic function: relatively strong gains can be achieved during the first few ticks that are spent on this task. For the same reason, when two equal logarithmic tasks are combined, the optimum strategy is to spend half of the total time on each task. Note however, that many strategies that surround this optimum also achieve similar scores, due to the nature of the diminishing returns curve underlying this function. As a result, this curve does not conform to the properties of an ideal payoff curve (see Chapter 5).

The main conclusion to take from this analysis is that in none of the combined cases an ideal payoff curve was present. In all cases there were either multiple global maxima (e.g., for all combinations of tasks involving an exponential or power function), or there were many strategies that had values close to the optimal value, leading to a less clearly defined peak (e.g., for the log-log combination and the log-linear and log-step combination).

The effects of introducing gains

Can introducing payoff functions that make one task more valuable than the other change these trends? To approximate this, I scaled the primary and secondary tasks using gain functions. That is, I used the equations for scaled functions from Table 6.1 (with values normalized between 0 and 100, before applying the scaling function *a*). I assumed that in all cases, spending more time on a task was beneficial, and therefore only used positive values of the scaling constant *a*. Rather than exploring a wide set of values, I only focused on situations in which there was a large contrast between the two tasks, assuming that other performances would lie in between these extremes. This was done by using values of 0.5 and 2 for *a*. In Figures 6.1-6.3 the extreme cases are illustrated. The dark grey line is a situation where the primary task (columns) was scaled with a value of 2 (i.e., performance was valued twice the normal value) and the secondary task (rows) was scaled with a value of 0.5 (i.e., performance was valued half the normal value). The light grey line reflects the inverse situation: when the primary task (columns) was scaled with a value of 0.5, and the secondary task (rows) was scaled with a value of 2.

In contrast to the situation without payoff scaling (the black lines), this time all combinations of tasks lead to predictions of a unique global maximum. When the primary task and secondary task consisted solely of linear or step functions (rows 2 and 3 of Figure 6.1), full attention should be given to the task that had the strongest weight: the primary task (dark grey lines) or the secondary task (light grey lines).

The Figure suggests that no intermediate strategies (i.e., dividing attention 60:40 between tasks) should be applied, as these lines always had a linear slope.

When the primary task followed a linear or step function and the secondary task was an exponential or power function (rows 4 and 5 in Figure 6.1), attention should again only be given to one of the two tasks, depending on the weight of the tasks. However, as before, spending more time on the linear or step function would involve less risk, as these tasks had a less steep grow curve. It might therefore be found that participants in an experiment choose these certain gains over the risky gains of the exponential function.

When the primary task was linear or step, and the secondary task was logarithmic, an interesting pattern emerged. When the task with the strongest weight was the secondary (logarithmic) task (light grey line), there was again only one optimal strategy: to spend all ticks on the primary task. However, when the strongest weight was put on the primary task (dark grey line), the shape approached the shape of an ideal curve: there was one global maximum, which was not completely at the extremes (i.e., not at “100 ticks” or “0 ticks”).

When the primary task involved either an exponential or a power function, the pattern was mostly as before. The optimal strategies were at the extreme ends of the strategy curves. However, in contrast to the situation without a scaling function, this time only one of the extreme strategies (i.e., far left, or far right) was optimal.

When the primary task was a logarithmic function, interesting patterns emerged when the secondary task was linear, step, or logarithmic. When the secondary task followed a linear or step function (rows 2 and 3 of Figure 6.3), the pattern was similar as described when these tasks were combined with a logarithmic task as secondary task (row 6 of Figure 6.1): the most well defined peaks occurred when strong weight was put on the linear or step function in comparison to the weight of the logarithmic task (light grey lines in rows 2 and 3 in Figure 6.3). However, a disadvantage of these peaks was that they were very close to the extremes. If this payoff setting would be used in an experiment, then a participant might decide to spend no time at all on the less important logarithmic task.

When both tasks were logarithmic, the location of the global maximum shifted from position 50 to a point closer to the extremes: the task with the strongest weight was preferred. However, note how these curves were again not shaped like an ideal payoff curve, because there were too many strategies that achieved values that were close to the global maximum value (which violates the definition of an ideal payoff curve, see Chapter 5).

Introduction of a loss function for not paying attention

In many situations there might not only be gains to be had when attending a task, but also costs associated with not attending a task. For example, in the studies discussed in the previous Chapters, there were costs associated with not attending to the driving task (Chapter 3) and the tracking task (Chapter 4). The longer the participant was away from these tasks, the more time the object drifted (i.e., the car in the lane or the cursor on the screen). This drifting lead to direct costs (i.e., being further from the center) as well as indirect costs (i.e., it would take longer to return the object to center).

In a third set of simulations, I inspected how the shape of the curve changed when penalties were introduced. It was assumed that the losses of a task followed the same shape as the gain function and that their base value (before scaling) was determined by the number of ticks not spent on the task. For example, if 40 ticks were spent on a logarithmic task, the (not normalized) gains would be log(40) and the (not normalized) losses would be log (100-40) = log(60).

As before, the functions were first normalized to produce scores between 0 and 100. In a second step, payoff functions for losses were applied. Two contrasts were applied, and are plotted in Figures 6.4 - 6.6 together with the baseline condition without losses (black line). In one condition the primary task (columns) had a loss function of -0.5 and the secondary task (rows) had a loss function of -2 (dark grey line). In a second condition the primary task had a loss function of -2 and the secondary task had a loss function of -0.5 (light grey line).

Similar to the results for the model with gain functions, in most cases these manipulations made sure that there was only one global maximum (when the primary task was an exponential or a power function, in some cases there were two global maxima). Typically, the optimal strategies were at the extreme ends of the strategy space. This indicates that in those cases, time should be spent exclusively on one task or the other, without consideration to the other task.

The most interesting results emerged when one task was logarithmic and the other task was linear or step and when the strongest penalty was applied for not attending the linear or step task. These are the dark grey lines in the bottom row of Figure 6.4 and the light grey lines in the second and third row of Figure 6.6. Here there was one global maximum, that did not lie at the extreme end of the strategy space. Note however, that this time these functions had a (small) local maximum and in that sense violated a property of the ideal payoff curve.

Further explorations of the combination of a linear and logarithmic task

The previous results showed that ideal payoff manipulations are most likely to be possible in situations where a linear task is combined with a logarithmic task, and in which the linear task has the strongest weight in the payoff function. To further explore whether such ideal payoff manipulations are possible, I explored a wider set of gain and loss scaling values for combinations of these two tasks.

In this simulation, the gain component of the linear task and (normalized) logarithmic task was varied between 0.5 and 5 in steps of 0.5 (i.e., 10 variations each). The loss components for both tasks were varied between 0 and -5, in steps of 0.5 (i.e., 11 variations each). The previous modeling results indicated that ideal payoff curves only arise when the linear task had a stronger loss or gain weight than the logarithmic task. Therefore, only situations were explored in which the gain component of the linear task was larger than the gain component of the logarithmic task, or in which the loss component of the linear task was larger than the loss component of the logarithmic task. This was the case for 8,470 simulations out of the full set of 12,100 simulations (10 x 10 x 11 x 11).

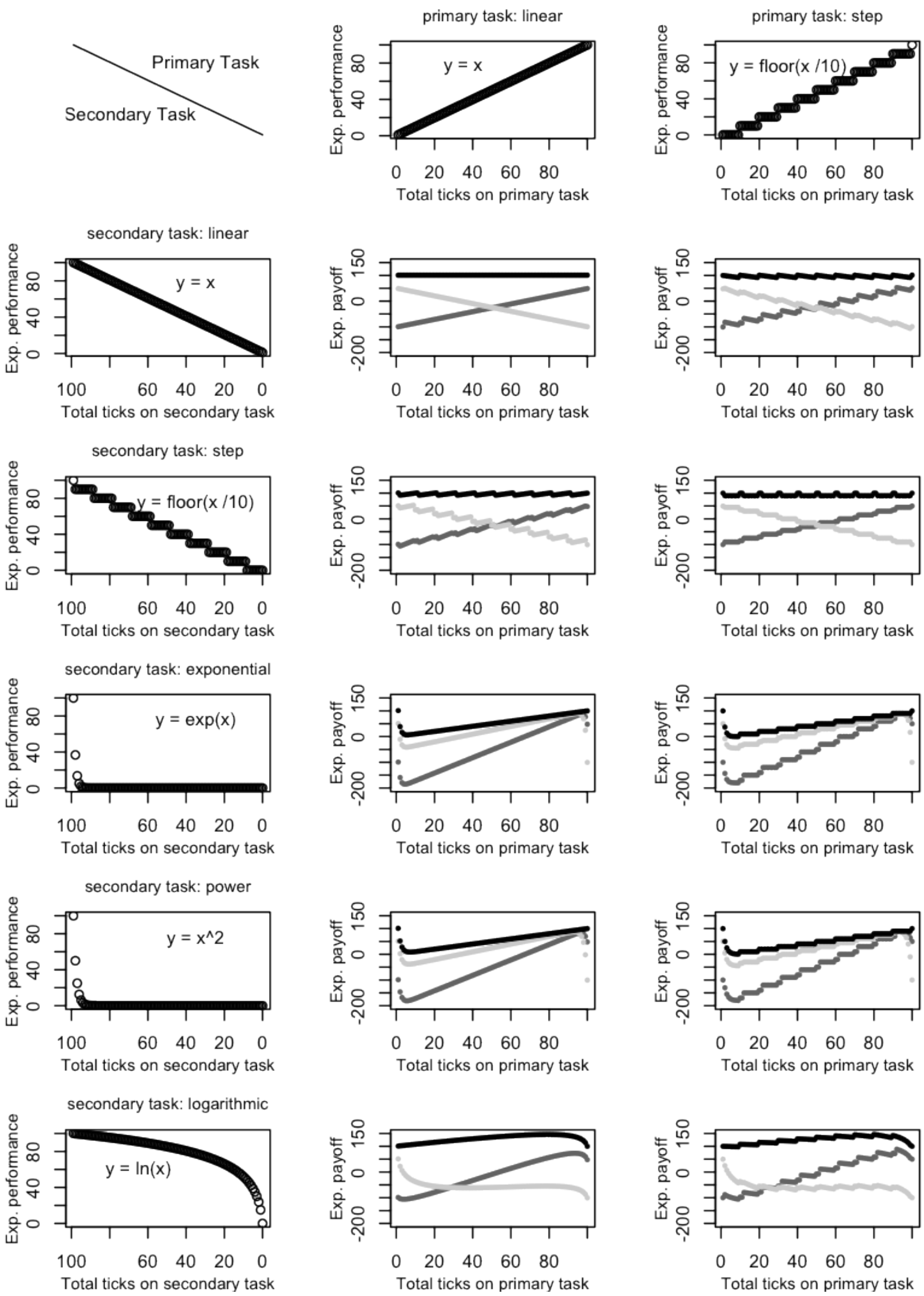

Figure 6.4: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for a linear function (2nd column) and a step function (third column).
Functions are normalized. Loss functions were then applied. The black line incurs no losses. For the dark grey line the losses are higher on the primary task (2) than on the secondary task (0.5). For the light grey line this is reversed.

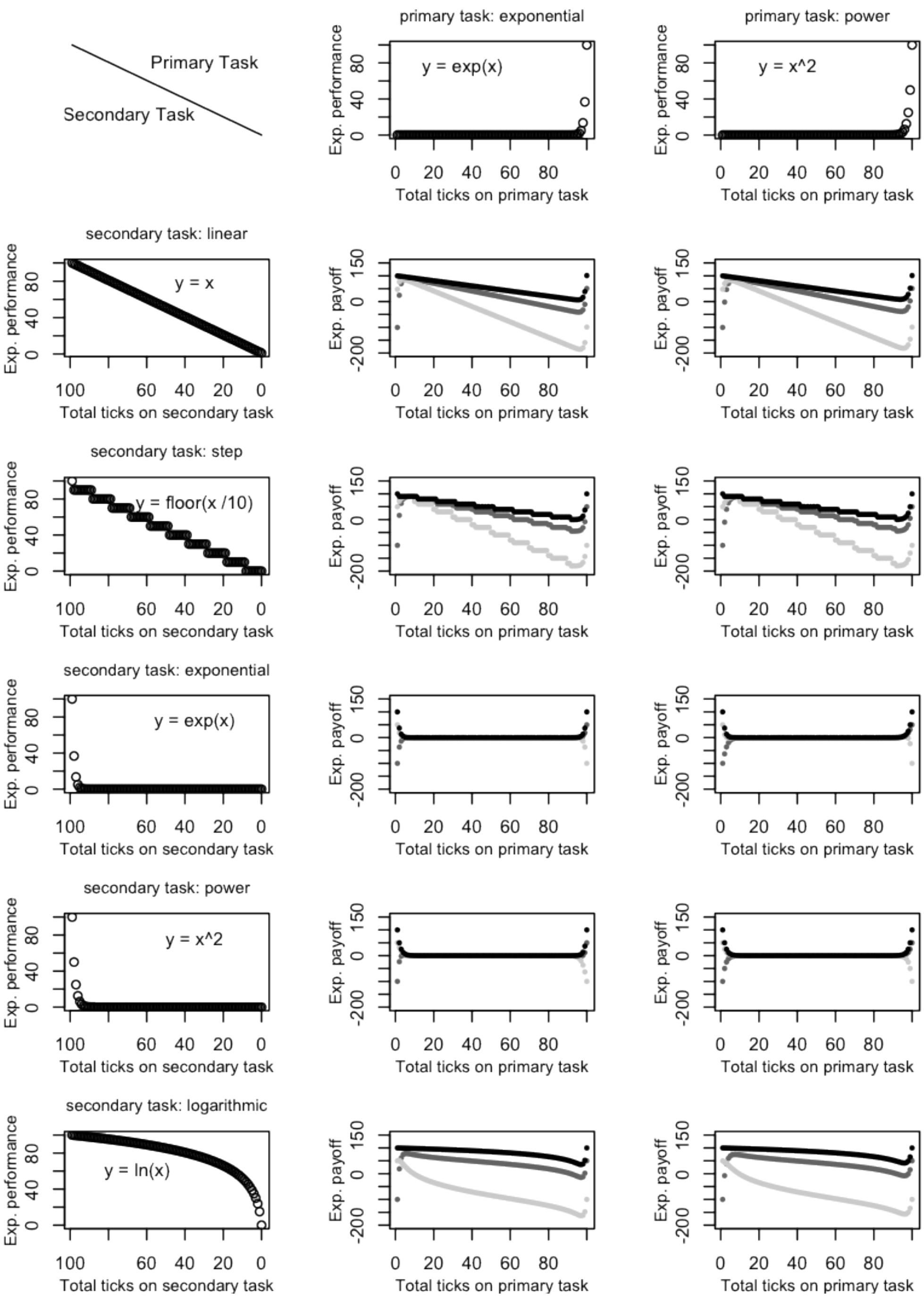

Figure 6.5: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for an exponential function (2nd column) and a power function (third column).
Functions are normalized. Loss functions were then applied. The black line incurs no losses. For the dark grey line the losses are higher on the primary task (2) than on the secondary task (0.5). For the light grey line this is reversed.

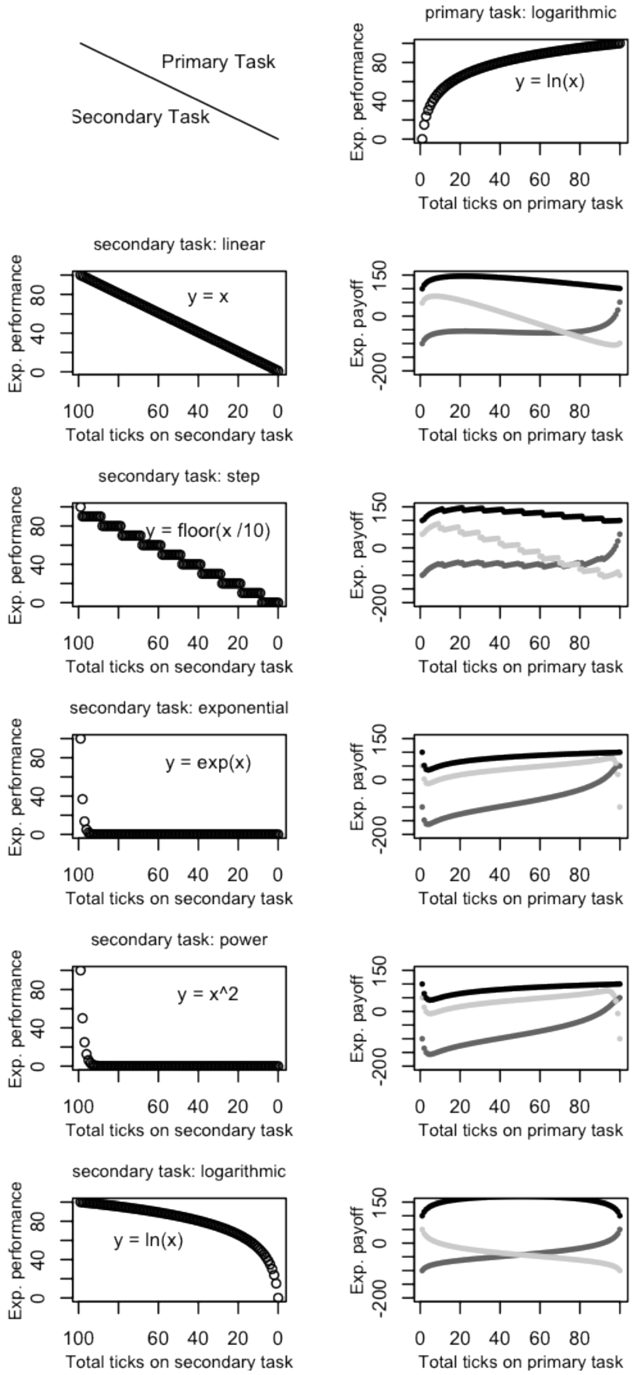

Figure 6.6: Examples of how performance increases when a primary task (columns) is combined with different secondary tasks (rows) for a logarithmic function (2nd column).
Functions are normalized. Loss functions were then applied. The black line incurs no losses. For the dark grey line the losses are higher on the primary task (2) than on the secondary task (0.5). For the light grey line this is reversed.

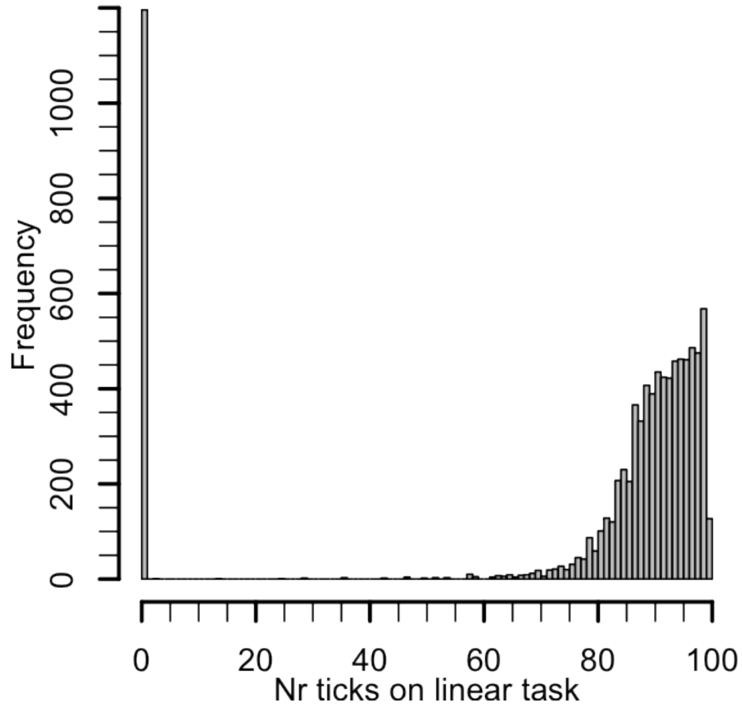

Figure 6.7: Histogram of the locations of global maxima when the payoff functions are varied for a combination of a linear and a logarithmic task (where the linear task has the stronger weights).

In each of these 8,470 simulations, the payoff curve had only one global maximum, of which the location differed between curves. Figure 6.7 plots a histogram of the location of the global maxima across payoff functions. The horizontal axis shows the number of ticks spent on the linear task. As a result of my pruning algorithm, that favored strong weights for the linear task, the majority of optima (for 7,260 payoff functions) occurred for strategies that spent more time on the linear task than on the logarithmic task (i.e., that are on the right half of the horizontal axis of Figure 6.7). Moreover, for all strategies that spent between 62 and 100 ticks on the linear tasks, there was at least one payoff function that made this strategy optimal. Between 51 and 62 this was more sporadic, which might be due to the limited set of simulations that I explored.[27]

The above finding satisfies one condition of the characteristics of an ideal payoff function: different strategies can be made optimal due to a change of the payoff function. However, one important property is not yet tested: whether the resulting payoff curves follow the characteristics of an ideal curve (see also definition in Chapter 5).

To explore this, Figure 6.8 plots example curves that have maxima at different locations, at (from left to right, and top to bottom) strategies 62, 67, 72, 77, 82, 87, 92, and 97. When possible, all curves that had the maxima at these locations were plotted. When more than 10 curves had a global maximum at this position, only 10 curves (randomly selected) were plotted.

A clear trend is visible in these plots: the further the maximum is to the right of the plot, the better defined the location of the global maximum becomes. For the plots that had a maximum at 62, 67, 72, 77, or 82 (the first five plots), the maximum is not clearly defined.[28]

---

[27] For 1.210 out of the 8.470 simulations the optimal strategy of the model was to spend more time on the logarithmic task (i.e., the number of ticks on the linear task was smaller than 50). Of these, the large majority (1.196 simulations) spent no ticks on the linear task. These strategies were not explored in further detail, as preceding analyses showed that these curves tended not to be in line with the characteristic of an ideal payoff curve.

[28] The absolute difference between strategies can be increased by increasing the value for the gain and loss components of the payoff curve. However, within such a curve, the relative difference between strategies compared to the possible maximum and minimum values will remain the same, as this linear scaling function will be applied to all values.

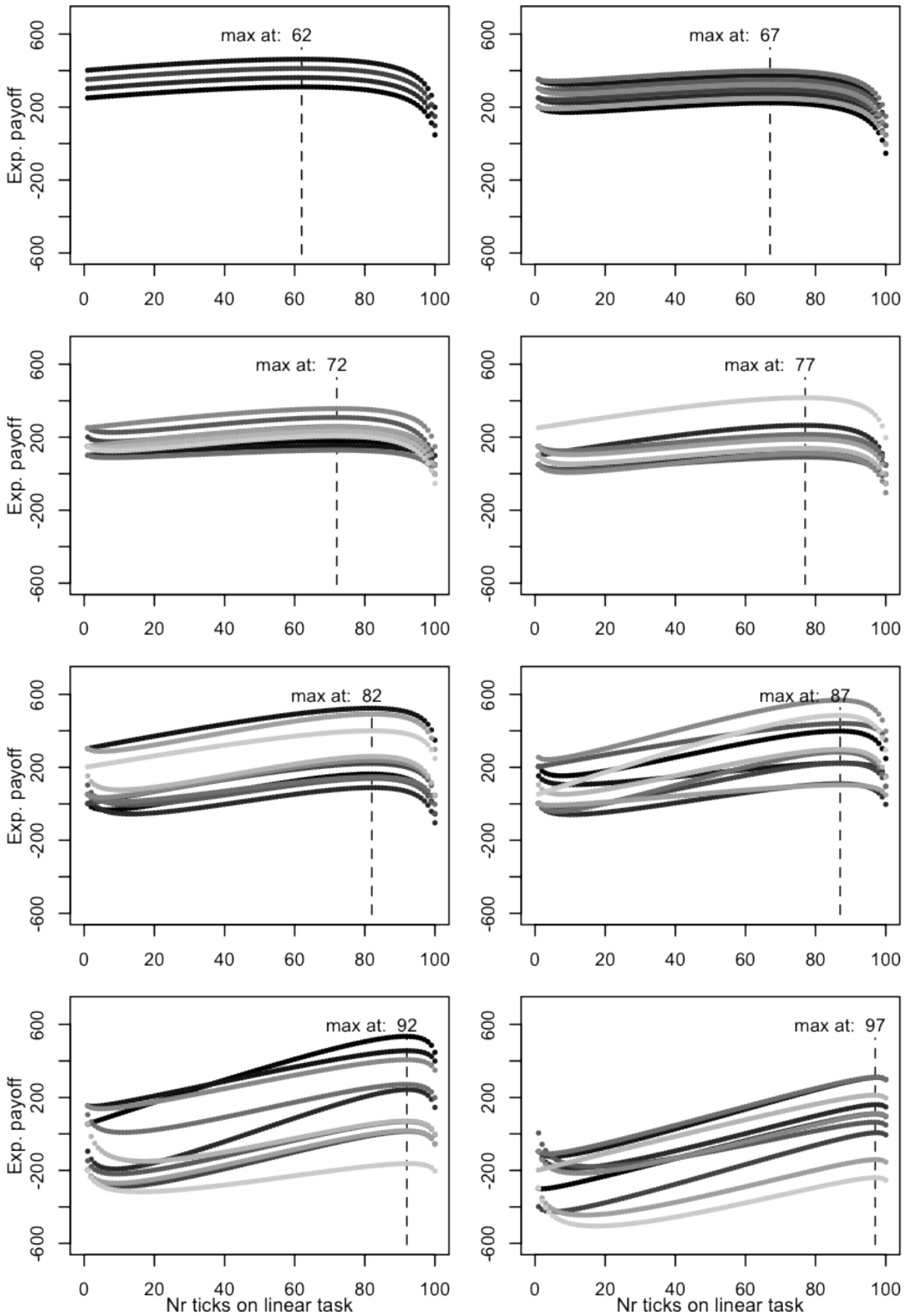

Figure 6.8: Example payoff curves sorted by the location of the global maximum. Note how the curves that have the maximum further to the right (towards the bottom of the Figure) have better defined peaks.

The above findings suggest that an approximation of ideal payoff manipulations is possible in this task environment that combines a linear with a logarithmic task. Payoff functions can be used to promote strategies that put strong weight on spending a relatively large amount of time on the linear task. One concern is how useful such a payoff manipulation would be. Given that the set of strategies that can be optimal lies close together (i.e., strategies 87-100), these can all be characterized at a higher level as “strategies that spend lots of time, but not all, on the linear task”. In an experiment, human participants might not pick up on more subtle aspects, such as whether they should spend 87 or 89 ticks on the linear task. Moreover, this might be difficult to distinguish due to variability in human performance or inaccuracy in measuring equipment.

6.4. General discussion

6.4.1. Summary of this Chapter

In this Chapter I have explored whether ideal payoff manipulations are possible for combinations of five generic task types that had different underlying performance functions: linear, linear step, exponential, power, or logarithmic functions. To study these effects, the analysis was done at a higher level of abstraction than in previous Chapters. In contrast to the work in the previous Chapters, in this Chapter detailed assumptions about constraints such as switch costs, and the role of moment-to-moment interleaving patterns were not considered. Rather, I explored how, in general, combinations of different tasks changed the predicted performance as a function of the total proportion of time spent on the primary task. I then introduced components that changed the payoff of tasks: the gains associated with spending more time on a task, and the losses associated with not attending the task.

The results of this model analysis showed how the shape of a performance and payoff curve is dependent on (1) the types of tasks being combined, (2) the gain component of the payoff function, and (3) the losses associated with not spending time on a task. The model predicted that trade-offs need to be made, and that at least some attention needs to be paid to each task (i.e., no task should be completely ignored), when at least one of the tasks had a logarithmic function. The most well defined peaks (a characteristic of an ideal payoff curve) occurred when two properties were met: (1) one task was logarithmic and the other task was linear, and (2) the strongest weight was put on the linear task through a manipulation of gains or losses.

A further exploration of this specific combination of tasks with different payoff functions found that although payoff manipulations can make different global strategies optimal, only strategies that strongly preferred the linear task (i.e., that spent over 82 percent of the total time on this task) had relatively well defined peaks. This is a problem, because it is not clear whether using payoff manipulations to promote different strategies within this small set will lead to insightful results. The set is small, and these strategies can all be characterized at a higher level as "strategies that spend lots of time, but not all, on the linear task". In an experiment, human participants might not pick up on more subtle aspects, such as whether they should spend 87 or 89 ticks on the linear task. Moreover, this might be difficult to distinguish due to variability in human performance or inaccuracy in measuring equipment.

6.4.2. Implications and further discussion

Uses of (approximately) linear and logarithmic functions in the literature

This modeling study suggested that not all arbitrary combinations of tasks can be used to give constrained predictions of the optimal strategies with which people interleave tasks in dual-task settings. If one wants to study these aspects, careful consideration should be given to the nature of the tasks in a dual-task paradigm and the mathematical function that approximates performance.

Incorporation of a task that follows a logarithmic function seems necessary. In particular, a combination of a linear and a logarithmic task seems to provide the most well defined peaks that require a balance of attention between tasks. Is there evidence of this in the literature?

Three particular interesting instances of successful payoff manipulations were identified. First, the curves from our own analysis (i.e., at the bottom of Figure 6.8) show close resemblance to the payoff curves that were generated by Howes, Lewis, and Vera (2009) in their studies of the PRP effect (i.e., see Figure 8 on page 738 and Figure 9 on page 740 of that paper). A translation of the two stimulus-response tasks in the PRP paradigm into a linear and a logarithmic component is not straightforward, and perhaps not even realistic. Nonetheless, the classical PRP prediction is that the response to task 1 is independent of the timing of the stimulus for task 2 (i.e., the reaction follows a linear function with a slope of 0), whereas the response to the second stimulus does depend on the timing of the stimulus (SOA) and seems to approximate the shape of an inverted logarithmic curve (e.g., see Figure 1 on page 721 of Howes et al., 2009). Future work should point out what is special about this combination of tasks in that it can lead to constrained predictions of optimality.

A second successful use of payoff manipulations was found in the work by Maloney and colleagues (Juni et al., 2011 ; Warren et al., 2012). Although slightly different in nature, both studies used a combination of a logarithmic and a linear function to generate predictions of optimal performance. For example, in the study by Juni et al. (2011), participants had to identify a hidden target on a touch-screen. The location of the target was randomly determined. To inform their decision about the location of the target, participants were allowed to sample different locations first: they chose a location, and the experimental software then revealed whether this location was part of the target area or not. Participants could determine themselves how many samples they wanted to collect before making a decision.

Performance on this sampling task approximated a logarithmic function. The more participants sampled, the more confident they could get about the position of the target. In particular, early samples quickly increased confidence (i.e., particularly if different areas of the touch-screen were explored). In addition, after sampling the entire touch-screen, participants would be 100% confident. The logarithmic function can be used to describe this rapid increase in confidence with later diminishing returns.

Maloney and colleagues also applied a linear loss function, by applying a fixed cost for every sample that was taken. So, the more locations a participant sampled before they made their decision, the more money they lost, in a linear fashion. The combination of the logarithmic probability curve and the linear loss function created a payoff curve: see Figure 1 in both papers (Juni et al., 2011; Warren et al., 2012).

The shape of the resulting curve satisfied all properties of an ideal payoff curve. In particular, in Warren et al. (2012) the peak of the payoff curve was well defined, as there the weight of the linear loss function was relatively strong (-6 points for every sample) compared to the gains range on the logarithmic function (between 0 and 1). This is in line with the prediction that followed from the mathematical models reported in this Chapter (i.e., see Figure 6.8).

A third example in which payoff functions were generated successfully, is a study by Jarvstad, Rushton, Hahn, and Warren (in press). In the experiments discussed in that paper, participants had to make a series of perceptual and cognitive decisions within a fixed time frame. The faster they responded to the stimulus of a single decision, the more decisions they could make within the fixed time frame. At the same time, hasty decisions could lead to incorrect answers and end up in costs. The participants therefore had to make a trade-off between the speed and accuracy of responses.

To generate payoff functions (what they call "efficiency functions") for performance within a fixed time frame, Jarvstad et al. went through two steps (see also Figure 1 in their paper). First, the accuracy of performance on a single decision, as a function of the time taken to make the decision, was approximated using a "Weibull function". Although this function is not logarithmic, it approximates many of the important properties of a logarithmic function: early investment of time in a decision increases accuracy, up till a point of diminishing returns, where a person is confident about the decision and extra time will not change this (i.e., where the maximum gains have been achieved and the rate of gain becomes 0).

In the second step, the performance on a single stimulus was accumulated for performance across a series of stimuli within the provided time-frame (recall that multiple decisions had to be made within a fixed time frame). In the studies by Jarvstad et al. (in press) participants gained a fixed amount of points for every correct answer and lost a fixed amount of points for every incorrect answer. This can be considered to approximate a linear function, as every answer contributed an equal gains to the payoff function. The combination of the two functions again provided a payoff curve that satisfied all properties of an ideal payoff curve as set out in Chapter 5.

The above example suggests that the incorporation of a logarithmic and linear function to generate ideal payoff functions is not an absolute requirement. Other functions that approximate the shape of these functions can also be used. For example, the Weibull function (Jarvstad et al., in press) captured two critical aspects of the logarithmic function: early high gain rates and later diminishing returns.

Note that the tasks that I used in Chapter 4 (tracking-while-typing) also approximated the "ideal combination of task". The typing task was a linear step function, and the tracking task can be described using a logarithmic function (see also Chapter 5). However, in comparison with the models described in this Chapter, the detailed consideration of further constraints (e.g., switch costs) introduced more factors that influenced the ability to create ideal payoff curves (see Chapter 5).

Note also that in the experimental studies reported in Chapter 4, the performance of the tasks approximated a linear step and a logarithmic function. However, to generate the payoff function, the performance curve was transformed using an exponential function before reporting the payoff to the participant (e.g., see Equations 4.1-4.3). As the analysis in this Chapter pointed out, using exponential performance curves might push performance more to the extremes.

Other combinations of tasks

The finding that a combination of (approximately) linear and (approximately) logarithmic tasks is worthwhile for exploring optimal decision making and task interleaving is not to say that explorations of other combinations of tasks is not useful. However, it does suggest that in such cases, no ideal performance curves can be generated that are in line with my definition of an ideal curve (see Chapter 5). This might then make it difficult to identify an objective criterion of "optimum performance" to which human performance can be compared.

However, with such functions, other predictions about performance might be given if other assumptions are made about how performance is constrained. For example, when two or more logarithmic tasks are combined, a prediction might be that people divide their time such that they are at the point where inhibition of returns starts (e.g., Payne et al., 2007). Such an assumption is useful, as it gives constrained predictions of performance: optimal performance should be at the point of inhibition of return, which is a unique point on the curve.

6.4.3. Contribution to the literature: an exploration of the usefulness of different task types for studying interleaving with ideal payoff manipulations

This Chapter has contributed to the literature by exploring what payoff curves might look like for a variety of general task types, and by exploring under what conditions ideal payoff manipulations are possible. It was found that the most well pronounced peaks in the payoff curve that require a participant to make trade-offs are those that combine a linear task with a logarithmic task. Such task combinations therefore seem most worthwhile to use in future studies that want to explore interleaving patterns using payoff manipulations. Moreover, the findings of this Chapter provided more insight into why other experimenters have been successful in generating ideal payoff functions in their studies: because they used tasks of which the performance function approximated these functions (Howes et al., 2009; Jarvstad et al., in press; Juni et al., 2011; Warren et al., 2012).

6.4.4. Limitations

The work in this Chapter has five limitations. First, no comparison was made with experimental data to verify whether the predictions for optimal performance align with human performance. The ability to understand under what conditions ideal payoff manipulations can be generated is a precondition to later test whether people can adapt their performance to achieve the optimal performance that was predicted by the model. This empirical question has been the focus of recent studies

(Howes et al., 2009; Jarvstad et al., in press; Juni et al., 2011; Warren et al., 2012).

A second limitation of this work was that the effect that moment-to-moment patterns of interleaving had on performance was not investigated. Rather, I studied how performance was affected by the overall proportion of time that was spent on one task at the cost of the time spent on another task (i.e., 80:20, 60:40, or 50:50). These general divisions of time can be achieved through various strategies for interleaving attention. For example, at one extreme, a 50:50 division of time can be achieved by spending the complete first half of a time frame on one task, and then dedicating the second half completely to another task. At the other extreme, this performance can be achieved by rapidly interleaving tasks. In between these two extreme strategies, there are many alternative strategies. An investigation of how these moment-to-moment interleaving strategies affect performance and payoff requires further specification of the constraints of a task (e.g., the switch costs). Specifying those for the studies in this Chapter would move the attention away from the general patterns that underlie the different performance functions and was therefore not considered here (but see Chapter 5 for a mathematical analysis of a tracking-while-typing task).

A third limitation of the work in this Chapter is that I have assumed that the penalty function of losses on a task (for not attending the task) followed the same function as the function of gains. For example, it was assumed that when gains followed a logarithmic function, the losses also followed a logarithmic function (e.g., as approximated in the tracking task, see Chapter 5). However, this might not always be the case. Take the example of writing a paper for a conference. Performance here might approximate a logarithmic function: early writing will significantly increase the quality of the document, whereas later "polishing" of the work often only makes smaller contributions to the quality. Not spending time on writing might also follow a logarithmic function. For example, the longer one does not look at the document, the more one might forget the details of the study being written up. At the same time, the more important deadline (submission) follows a step-function: you either submit in time or not. Not spending enough time on the paper, and missing the deadline due to distraction, might therefore have a steep loss as a result: not submitting.

A fourth limitation of this work is that only a limited set of parameter combinations for gains and losses was explored for all functions (see Figures 6.1-6.6). Although small, the various combinations of parameters formed useful contrasts to pursue, and allowed for identification of general patterns (e.g., that combinations of linear functions always favor one tasks completely, or favor no task at all). I explored more parameter combinations for one particular setting that seemed particularly promising (combining logarithmic and linear tasks). To be confident that no subtle patterns were missed, more explorations can be made of the influence of parameter values for other task combinations.

One final limitation is, that I have not looked at situations where the performance on a task is transformed with another function before applying a payoff function. For example, in Chapter 4 the performance on the typing task (a linear step function) was used as input to an exponential function before reporting the payoff value (see Equations 4.1-4.3). Such complex combinations of functions were not considered here, as they expand the parameter space enormously (i.e., all parameters a, b, c, and d can then be applied to every function). Moreover, in many cases, performance of such manipulations will result in performance that can be approximated by the functions that were explored here. In particular, all transformations of a linear function by other functions will be similar to the predictions when directly applying the non-linear function.

6.5. Conclusion

This analysis showed that explorations of interleaving patterns in a multitasking setting through ideal payoff manipulations can be best carried out with a combination of an (approximately) linear and a logarithmic task. With such functions, the location of the maximum strategy can be manipulated through a manipulation of the gains and losses associated with (in)attendance to the task. However, there are limitations in the set of strategies that can be made optimal in this way and further work is required to understand what is "special" about this combination of tasks.

Chapter 7. Conclusion and General Discussion

Abstract

In this thesis I explored when people interleave attention in dual-task settings. The hypothesis is that people try to perform in a cognitively bounded rational way. Performance is limited by constraints that come from the task environment and cognition. If, given these constraints, multiple strategies for interleaving tasks are available, then people will interleave tasks in a way that aligns with their local priority objective (Chapter 3), or which maximizes the value of an objective payoff function that evaluates performance (Chapter 4). This hypothesis was tested using a combination of experimental studies and computational cognitive models. Across a series of studies, the interplay between different constraints was investigated. In Chapters 5 and 6, I developed mathematical models to study what task combinations in general allowed for "ideal payoff manipulations" to study task interleaving. The work contributed to the existing literature in four ways: (1) it provided an overarching theory of skilled human dual-task performance and tested this in relatively applied settings, (2) the theory was formalized in computational cognitive models that can predict performance of unobserved strategies and that can bracket the (optimal) performance space, (3) linear and logarithmic tasks were identified as an ideal combination for achieving ideal payoff manipulations, and (4) results demonstrated that in multitasking situations attention is not necessarily interleaved solely at chunk boundaries and other "natural breakpoints", but that this depends on a person's priorities. The work has implications for driver distraction

research, in that it helps in systematically understanding the performance trade-offs that people face when multitasking. Moreover, the modeling framework could be used for model-based evaluation of new mobile interfaces. Finally, the demonstration that priorities can strongly influence multitasking performance highlights the importance of public safety campaigns that emphasize awareness of driver safety. Limitations and further implications are discussed.

7.1. Summary of research

The core research question of this thesis is to better understand when people interleave their attention between two tasks and to offer an explanation for this behavior. Following work by Howes, Lewis, and Vera (2009) on skilled behavior, I put forward the hypothesis that dual-task performance is cognitively bounded rational. Figure 7.1 (a redraw of Figure 2.2) schematically captures this theory. It is proposed that a set of constraints (grey boxes at the top of the Figure) systematically narrows down the space of possibilities for interleaving two tasks (blue boxes at the bottom of the Figure). Characteristics of the task and cognition constrain a set of strategies that is possible for a specific individual and a specific task. If an explicit priority is set, or if performance is evaluated using an objective payoff function, then a set of locally optimal strategies can be identified. These are the strategies that comply with the priority or maximize the score from the payoff function. The hypothesis is that participants will apply these locally optimal strategies. This can then explain why participants interleave at certain moments in the task and not others.

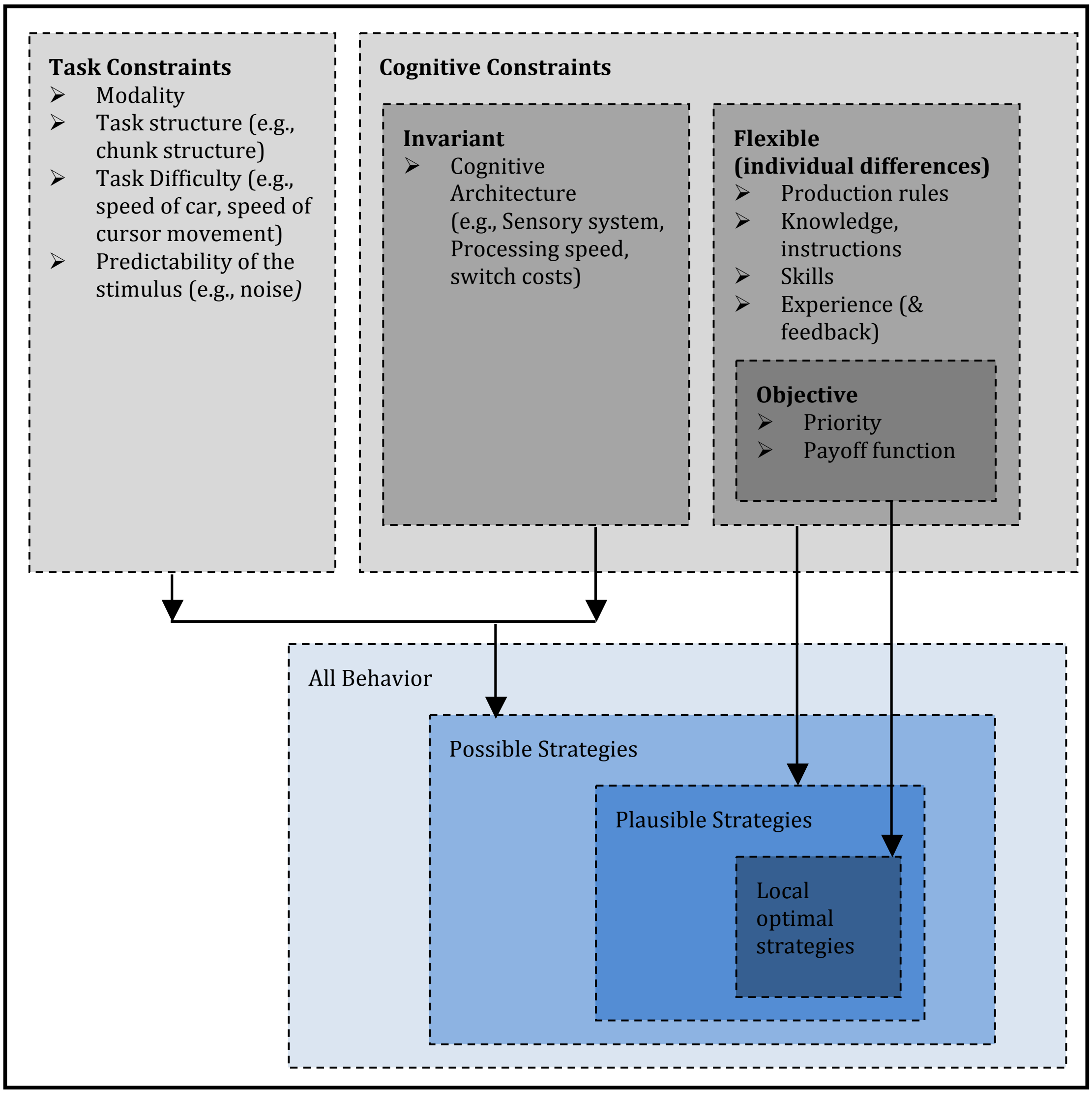

Figure 7.1: The task constraints and cognitive constraints that narrow down the space of all possible behavior to those strategies that are locally optimal. (reprint of Figure 2.2)

To test this hypothesis, I systematically manipulated the nature of these constraints in a series of experiments. In a dialing-while-driving task (Chapter 3), I manipulated the structure of a rehearsed phone number (task and cognitive constraint) and the priority of the driver. Participants adapted their performance to these manipulations. A subsequent modeling analysis suggested that participants adapted their performance in a way that achieved an optimal performance trade-off: they tried to comply with the priority objective while at the same time balancing performance on both tasks (by applying a strategy that was on the dual-task performance trade-off curve). In contrast to preceding work (Salvucci, 2005), my studies showed that participants did not interleave attention if and only if they reached a 'natural breakpoint' in the dialing task (see Chapter 3 for more details).

In Chapter 4, I also manipulated the task constraints and the payoff function for evaluating performance. Participants adapted their performance to the task constraints and to the payoff function. In addition, their performance was adapted to their own typing skills (i.e., to measurable individual differences in performance). A subsequent modeling analysis suggested that participants adapted their performance in such a way as to try and maximize the payoff value. This is not to say that performance was optimal on every trial. Indeed, it was found that multiple strategies could achieve reasonably good performance. Unfortunately, this made the set of optimal strategies less constrained than originally intended.

A mathematical model of the tracking-while-typing task was developed in Chapter 5, to see whether "ideal payoff manipulations" were possible in this task environment. In an ideal payoff manipulation, different strategies can be made the global optimal solely through a manipulation of payoff. In addition, in such settings, each payoff function only has one strategy that is the global maximum and that is well defined within the set of available strategies (see Chapter 5 for a more detailed definition). It was found that ideal payoff manipulations are difficult to achieve in a tracking-while-typing environment.

In Chapter 6, I then investigated what task environments were suitable for such ideal payoff manipulations. This was done by comparing the payoff curves of different classes of tasks. Tasks were distinguished based on the mathematical form of their underlying performance function (e.g., linear, linear step, exponential, power, or logarithmic). It was found that ideal payoff manipulations are easiest to conduct in settings where a (approximately) logarithmic performance function (or task) is combined with a (approximately) linear performance function (or task), and in which the linear function has the strongest weight in the payoff function. Examples were found in the literature of where such combination of functions were used or approximated to generate ideal payoff curves (e.g., Jarvstad et al., in press; Juni et al., 2011; Warren et al., 2012), or in which the payoff curve approximated a similar shape as would arise when combining these two functions (Howes et al., 2009).

Taken together, the work suggests that people are very flexible in their decisions of when they interleave two tasks. They adapt their performance to both external factors (e.g., the task structure) and internal factors (e.g., their priorities).

7.2. Contributions to the literature

In the general discussion of the individual Chapters I have listed the specific contributions to the literature that were made in those Chapters. Here I list four more general contributions that were made across the thesis. All contributions (specific and general) are also listed separately at the beginning of this thesis (see index), with a reference to the relevant Chapter.

Contribution 7.1: A theory of skilled dual-task behavior, tested in two relatively applied settings

As summarized above, I proposed that dual-task performance can be understood as cognitively bounded rational behavior (Howes et al., 2009). This theory has been tested before in the psychological refractory (PRP) paradigm (Howes et al., 2009; Howes et al., 2004; Lewis et al., 2004; Vera et al., 2004). However, in some respects the PRP task is simple: stimuli appear at their own pace and single responses need to be made. Slightly more complex are dynamic discretionary task interleaving scenarios, where participants need to decide themselves when to switch attention from one dynamic task to another. In this thesis I contributed to the literature by testing the theory of cognitively bounded rational analysis in such settings: a dialing-while-steering task (Chapter 3) and a tracking-while-typing task (Chapter 4). Independent of the model assessment, the experiments demonstrated that human dual-task performance adapts to characteristics of the task, cognition, individual differences in skill, and objectives.

Contribution 7.2: Refinement of a methodology to predict human multitasking performance and application to two dual-task scenarios

This thesis also made a methodological contribution through refinement and development of computational cognitive models for understanding dual-task strategic adaptation. In Chapter 3 I refined an existing model of driver distraction (Brumby, Howes et al., 2007; Brumby, Salvucci et al., 2007; Brumby et al., 2009). The model was applied to investigate dual-task performance with phone numbers with various internal structures. The model was then refined to incorporate components for modeling the motor cost of moving the fingers between keys on the phone.

In Chapter 4, I developed a new model that could be used to assess the utility of dual-task strategies using payoff functions. Payoff functions have been used in experimental psychology before (e.g., Neth et al., 2006; Schumacher et al., 1999; Wang et al., 2007, 2009) and are being applied more and more in computational frameworks. For example, in studies on the Psychological Refractory Period (Howes et al., 2009; Howes et al., 2007; Howes et al., 2004; Lewis et al., 2004; Vera et al., 2004), dynamic task interleaving (Hornof & Zhang, 2010; Neth et al., 2008), motor movement (Jarvstad et al., in press; Juni et al., 2011; Maloney & Mamassian, 2009; Maloney & Zhang, 2010; Trommershauser et al., 2003a, 2003b, 2008; Warren et al., 2012; Wu, Delgado, & Maloney, 2011; Zhang et al., 2011), and vision (Ballard & Sprague, 2007; Reichle & Laurent, 2006; Tatler et al., 2011). The work in this thesis built on this preceding work and extended it by applying payoff functions to a dynamic task, while at the same time using the payoff function to make predictions about locally optimal performance.

The model in Chapter 4 was also refined to take individual differences in skill into account to predict multitasking performance. There has been some focus in the process-oriented cognitive modeling community on explaining individual differences in performance. However, this focus has mostly been on bracketing performance for different types of individuals, such as fast and slow users (Card et al., 1983), or for old and young participants (Meyer et al., 2001). More graded models of differences in skill have been applied, but are few in number. Most notable are the model of working memory by Lovett, Daily, and Reder (2000), in which performance on a new task was predicted based on an independent calibration of working memory skill, and the models of individual strategic differences in performance of the PRP task by Howes, Lewis, and Vera (2009).

In Chapter 5 and 6, I developed a mathematical model to investigate whether ideal payoff manipulations are in principle possible. The model in Chapter 5 revealed limitations to the method of using payoff functions to identify optimum performance in the tracking-while-typing task. The model in Chapter 6 provided insights into what types of tasks are useful for ideal payoff manipulations in general (see contribution 7.3).

Contribution 7.3: Identification of linear and logarithmic tasks as an ideal task combination for further study of ideal payoff manipulations

Given the difficulty with generating ideal payoff functions in a tracking-while-typing task (see Chapter 5), in Chapter 6, I studied more generally what task environments allowed for ideal payoff manipulations. It was found that settings that combine (approximately) logarithmic tasks with (approximately) linear tasks are most worthwhile, especially when the linear task has the strongest weight in the payoff function. Examples were found in the literature of where such combination of functions were used to generate ideal payoff curves (e.g., Jarvstad et al., in press; Juni et al., 2011; Warren et al., 2012), or in which the performance curve approximated a similar shape as would arise when combining these two functions (Howes et al., 2009).

Contribution 7.4: An investigation of interleaving at natural breakpoints

In Chapter 3, I investigated a specific theory about interleaving: that people tend to interleave tasks when they reach a natural breakpoint in the task structure (Salvucci, 2005). The experiments demonstrated that participants do not necessarily interleave if and only if they reach a natural breakpoint. Rather, natural breakpoints are taken as cues to interleave. Whether people follow up on these cues depends on their priority and on the number of available natural breakpoints. The results of the third study in Chapter 3 suggested that motor cues can also be considered to be natural breakpoints. The modeling analysis suggested a new advantage that interleaving at natural breakpoints gives: interleaving here, rather than at other points, offers beneficial speed-accuracy performance trade-offs.

7.3. Practical implications

Besides contributions to the scientific literature, the work presented in this thesis also has implications for driver distraction research. Research in this area is important, as drivers (1) show a tendency to continue to perform distracting side tasks in the car, despite legislation forbidding it (e.g., Crowd Science, 2009 ; Diels et al., 2009; The Economist, 2011), (2) will not always choose the most safe point to perform side tasks while they are driving (Horrey & Lesch, 2009), and (3) are not always aware of the distracting effects that performing side-task can have on driving (Horrey et al., 2009). Engineering solutions that can alleviate the distracting effects of multitasking in the car, or can encourage appropriate interleaving (i.e., trigger a driver to switch attention back to driving) are therefore valuable. This will allow a driver to multitask as desired, while at the same time minimizing risks. The modeling analysis that I presented in this work can help in identifying engineering solutions and, in future work, could be used to test the possible success of those solutions.

Practical implication 1: Proposing consideration of natural breakpoints and priorities in the design, prototyping, and evaluation of technology for multitasking

In Chapter 3, I demonstrated how "natural breakpoints" in a task can work as a cue to interleave attention between tasks. This implies that it is worthwhile to consider natural breakpoints when designing, prototyping, or evaluating technology for multitasking situations. Incorporating more breakpoints in a task can encourage more interleaving, for example to ensure that a user repeatedly focuses on the more important task (e.g., driving).

This approach has two challenges. First, in order to design natural breakpoints into a task, a detailed understanding of the (hierarchical) task structure is required. This might not always be feasible. A second challenge is that even if the interface is designed to incorporate an appropriate set of breakpoints, whether people use them depends on their priorities (see Chapter 3).

These challenges put forth two further implications. First, dangerous behavior is hard to design out, because priorities can outdo design efforts for safety. Second, it is worthwhile to keep on promoting driver safety, as it might encourage drivers to set their priorities appropriately, this will then have a beneficial effect on their safety.

Practical implication 2: Using cognitive models to bracket performance

Throughout this thesis I have demonstrated how cognitive models can be used to explore the performance of alternative strategies – even unobserved strategies – in a dual-tasking context. The models that I used were developed around a relatively small explicit set of parameters, which was based on literature and measurements of performance (mostly in single-task settings). In Chapter 3 specifically, I highlighted examples of model parameters and their associated real-world counterparts and how these can be changed to explore performance trade-offs in novel driver distraction settings. Even in cases where a developer is not certain about which priorities a user will have when using the device, the methodology in this thesis can be used to bracket human performance (Kieras & Meyer, 2000). That is, it can be used to identify what the space of possible interactions with the interface looks like.

7.4. Limitations

Within each Chapter I have listed specific limitations of the work reported there. Here I focus on four general limitations of the work. Future research is needed to address each of these limitations and is beyond the scope of this thesis.

Limitation 1: Assumptions of the model

I used predictions made by computational cognitive models to assess whether performance by participants was optimal. Any inferences made about whether participants performed optimally therefore underlie the assumption that the model and its predictions are correct. In all Chapters (in particular, in Chapters 3 and 4) I have highlighted aspects of the model that could be improved. Some of these aspects are also discussed later in the section on "important areas for further research".

There are two general limitations to the model. First, in most cases, the model only approximated psychological (moment-to-moment) processes, and more detailed theories of these processes can be developed. For example, typing was described only as a fixed time interval. However, it can be decomposed into more fine-grained processes, such as reading a digit, memorizing it, moving the finger to an appropriate key, and then typing in the key.

A second general limitation is that only a limited strategy space was explored. This limitation will be discussed in more detail below for general limitation 2 ("More adaptation is possible outside of the lab") and limitation 4 ("forced interleaving").

In this context it is worth to highlight the ways in which I tried to validate the models and attempted to prevent overfitting of the models to the data (see also general discussions in each Chapter). Foremost, the models consisted of a series of assumptions that were explicitly stated, and that were either based on direct observations of performance, on findings in the literature, or on assumptions that were tested in preceding models. For example, the driving model was directly taken from a previous modeling effort and only required minor modifications for the current setting (Brumby, Howes et al., 2007; Brumby, Salvucci et al., 2007; Brumby et al., 2009). As such, the generalizability of the model improved – the model functioned in various contexts (cf. Pitt & Myung, 2002; S. Roberts & Pashler, 2000).

For the tracking-while-typing task a novel model was developed. It followed the same general principles of the driver distraction models: strategies were varied using the time on task and actions were grounded in measurements. As no separate benchmark and calibration data was available for the model development, the model structure and underlying assumptions were tested rigorously (especially for model 4B) to avoid overfitting. Human and model data were compared on a variety of measures using various techniques. Model fit was also assessed in relationship to model complexity.

As is the case with any quantified theory, a good fit can never prove that the model is correct (Pitt & Myung, 2002; S. Roberts & Pashler, 2000). A good fit and correct predictions of performance for novel settings can only support the theory. I used the driving model to test predictions for novel settings, and also laid out how the model can be used to predict performance in other settings (see general discussion in Chapter 3).

Limitation 2: More adaptation is possible outside of the lab

A core argument in this thesis is that people adapt their multitasking strategies to the local context. In the models only two dimension of adaptation were considered: the amount of time spent on task A and the amount of time spent on task B. I consider these to be fundamental dimensions of the tasks studied here and representative of the options available to participants in the controlled experiments.

In the real world, however, adaptation might take other, additional forms. For example, in the literature review I discussed how people sometimes change their task environment to better suit their goals (Kirlik, 1998). As people gain more control over their environment, they can use this control for further (strategic) adaptation. For example, in studies that used high fidelity driving simulators, people have been observed to compensate for unsafe multitasking behavior by for example reducing their speed (e.g., Cnossen et al., 2004; Iqbal et al., 2010).

Limitation 3: Task environments were without risks

As the studies were conducted in a lab setting, there were no real risks for the participant. They might therefore have felt more comfortable to multitask and to adopt more "risky" strategies, such as spending prolonged times away from steering in the simulated driving environment that was used in Chapter 3. The real world imposes more risks, and therefore might discourage people from performing in such ways.

At the same time, observations of unsafe performance have also been made in more realistic settings. For example, Horrey and Lesch (2009) conducted a study on a closed test-track where participants had to perform secondary tasks while driving a vehicle. The participants knew the track, which had a variety of sections that differed in demands, including a shoulder where participants could park the car. Despite this knowledge, the participants did not delay the secondary tasks until they arrived at these safer points. Moreover, even in real traffic, mobile phone usage is not decreasing (Diels et al., 2009), despite increasing knowledge of the associated risks.

Limitation 4: Forced interleaving

In the experiments in this thesis I forced (Chapters 4, 5) or assumed (Chapter 3) that participants could only perform one task at a time. This made it easier to study attention interleaving. Although in many settings people will not be able to perform two tasks fully in parallel, they might be able to dedicate part of their resources (e.g., eyes, hands, memory) to one task and part of their resources to other tasks (Salvucci & Taatgen, 2008, 2011; Wickens, 2002, 2008). Allowing for such "shared attention" will of course change the strategy space. For example, in the simulated driving environment, participants might quickly glance at the road while they type in a digit.

The performance of such more fine-grained strategies was not considered in my analyses. Using the models, I did explore performance of extreme strategies (no interleaving and interleaving after every digit) and many strategies in between those extremes. It is therefore expected that performance of more "complex" strategies lies somewhere within the performance brackets of these extreme strategies (similar to bracketing approaches in e.g., Card et al., 1983; Kieras & Meyer, 2000).

7.5. Important areas for further research

The work in this thesis has paved the way for three broad areas of further research. Each of these avenues would be interesting to pursue.

Area 1: Adaptation over time and learning

The work in this thesis has contributed to our understanding that people adapt their performance in multitasking situations to several factors to try and achieve optimal performance. An important open question is how people adapt their performance over time to achieve optimal performance. The question of how multitasking performance changes over time is interesting for theory formation. For example, because it requires an understanding of how feedback signals from multiple tasks are taken into account for learning. Having a better understanding of how people adapt their interleaving over time is also of practical importance. For example, it can inform the design of training methods for learning to multitask appropriately and it can be used to design appropriate feedback signals (i.e., those that encourage fast learning and put emphasis on the right tasks).

Area 2: Individual differences

In Chapter 4, individual differences in skills emerged as an important factor that influences multitasking performance. This also aligns with recent other findings of individual differences in multitasking performance (e.g., Ophir et al., 2009; Watson & Strayer, 2010). My studies were not designed to manipulate or provide a detailed explanation of individual differences. As such, the findings here are only the tip of the iceberg. To get a more detailed understanding of individual differences in multitasking skill, future studies need to provide more control over the source of individual differences. This can be achieved in at least two ways. First, participants could be preselected based on their different characteristics (e.g., highly-trained versus novice users, old versus young participants). Second, the experiment can provide more benchmark and calibration (single-task) tests to study the skills in which people differ before studying their effect on multitasking performance.

Area 3: Model and theory refinement

The models and theories that I developed in this thesis can be expanded and refined. In each Chapter I have addressed specific ways for doing this. Here, I use a more general description, borrowed from Peter Pirolli (see Chapter 10, in Pirolli, 2007). The development of the model and the associated theory can be upward, downward, inward, and outward.

Upward and downward extension means that the model can be extended to provide more insight in performance at different levels of abstraction (Newell, 1990) (see also Chapter 2 for a discussion of the abstraction continuum and the position of my work within it). Upward extension means that the model could be further developed to provide more insight about phenomena at higher levels of abstraction. For example, to provide insight into social aspects of multitasking and how these aspects (for example: motivation, or a desire to stay connected to friends all day long) influence performance.

Downward extension means that the model can be extended to provide more insight into more detailed aspects of the cognitive and biological band. For example, in all models the typing of a digit was modeled only as a fixed time cost. Yet, in reality many other fine-grained processes underlie this. Similarly, all the strategies that I modeled assume that only one task can be performed at a time. In practice, people might be able to dedicate some of their resources to additional tasks (Salvucci & Taatgen, 2008, 2011; Wickens, 2002, 2008), which might then influence performance. To investigate this, a different level of abstraction is needed, as well as collection of more fine-grained data (e.g., eye-tracking data).

Inward extension means that more phenomena could be explained within the current level of abstraction and the current task set. As noted earlier, a theory of learning and a more detailed theory of individual differences in performance would seem particularly fruitful avenues to explore.

Outward extensions means that the model can be extended to explain phenomena in different tasks and different contexts. In particular, the theory could be applied to tasks that require interleaving over different time scales. That is, to tasks that are on different sections of the multitasking continuum (Salvucci & Taatgen, 2011; Salvucci et al., 2009), see Chapter 2. The theory can also be applied to different levels of application, see Chapter 2 (Salvucci & Taatgen, 2011).

7.6. Conclusion

Human multitasking performance is highly adaptive. In this thesis I described how this adaptive performance is systematically influenced by characteristics of the task and cognition (see Figure 7.1). By systematically specifying these factors, the work has helped to understand how these factors interact and constrain human performance.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrox & F. Caski (Eds.), *Proceedings of the International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Alm, H., & Nilsson, L. (1995). The Effects of a Mobile Telephone Task on Driver Behaviour in a Car Following Situation. *Accident Analysis & Prevention, 27*, 707-715.

Altmann, E. M., & Trafton, J. G. (2002). Memory for Goals: An Activation-Based Model. *Cognitive Science, 26*, 39-83.

Anderson, J., R. Taatgen, N. A., & Byrne, M. (2005). Learning to Achieve Perfect, Timesharing: Architectural Implications of Hazeltine, Teague, and Ivry (2002). *Journal of Experimental Psychology: Human Perception and Performance, 31*, 749-761.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review, 111*, 1036-1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An Integrated Theory of List Memory. *Journal of Memory and Language 38*, 341-380.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science, 2*, 396-408.

Atchley, P., & Chan, M. (2011). Potential Benefits and Costs of Concurrent Task Engagement to Maintain Vigilance: A Driving Simulator Investigation. *Human Factors, 53*, 3-12.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding Changes in Mental Workload During Execution of Goal-Directed Tasks and Its Application for Interruption Management. *ACM Transactions on Computer-Human Interaction, 14*, 1-28.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience, 7*, 66-80.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic Codes for the Embodiment of Cognition. *Behavioral and Brain Sciences, 20*, 723-742.

Ballard, D. H., & Sprague, N. (2007). On the Role of Embodiment in Modeling Natural Behaviors. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 283-296). New York, NY: Oxford University Press.

Ballas, J., Heitmeyer, C., & Pérez-Quiñones, M. (1992). Evaluating Two Aspects of Direct Manipulation in Advanced Cockpits. In P. Bauersfeld, J. Bennett & G. Lynch (Eds.), *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 127-134). New York, NY: ACM Press.

Bogacz, R. (2007). Optimal Decision-Making Theories: Linking Neurobiology with Behaviour. *Trends in Cognitive Sciences, 11*, 118-125.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks. *Psychological Review, 113*, 700-765.

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do Humans Produce the Speed-Accuracy Trade-Off That Maximizes Reward Rate? *Quarterly Journal of Experimental Psychology, 63*, 863-891.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The Neural Basis of the Speed-Accuracy Tradeoff. *Trends in Neuroscience, 33*, 10-16.

Bogunovich, P., & Salvucci, D. D. (2010). Inferring Multitasking Breakpoints from Single-Task Data. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual meeting of the Cognitive Science Society* (pp. 1732-1737). Austin, TX: Cognitive Science Society.

Borst, J. P., & Taatgen, N. A. (2007). The Costs of Multitasking in Threaded Cognition. In R. L. Lewis, T. A. Polk & J. E. Laird (Eds.), *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 133-138). Ann Arbor, MI: University of Michigan.

Borst, J. P., Taatgen, N. A., & Van Rijn, H. (2010). The Problem State: A Cognitive Bottleneck in Multitasking. *Journal of Experimental Psychology: Learning, memory, and cognition, 36*, 363-382.

Botvinick, M. M. (2008). Hierarchical Models of Behavior and Prefrontal Function. *Trends in Cognitive Sciences, 12*, 201-208.

Brookhuis, K., De Vries, G., & De Waard, D. (1991). The Effects of Mobile Telephoning on Driving Performance. *Accident Analysis & Prevention, 23*, 309-316.

Brumby, D. P., Davies, S., Janssen, C. P., & Grace, J. J. (2011). Fast or Safe? How Performance Objectives Determine Modality Output Choices While Interacting on the Move In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole & W. Kellogg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 473-482). New York, NY: ACM Press.

Brumby, D. P., del Rosario, N., & Janssen, C. P. (2010). When to Switch? Understanding How Performance Tradeoffs Shape Dual-Task Strategy. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the International Conference on Cognitive Modeling* (pp. 19-24). Philadelphia, PA: Drexel University.

Brumby, D. P., Howes, A., & Salvucci, D. D. (2007). A Cognitive Constraint Model of Dual-Task Trade-Offs in a Highly Dynamic Driving Task. In B. Begole, S. Payne, E. Churchill, R. St. Armant, D. Gilmore & M. B. Rosson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 233-242). New York, NY: ACM Press.

Brumby, D. P., Salvucci, D. D., & Howes, A. (2007). Dialing While Driving? A Bounded Rational Analysis of Concurrent Multi-Task Behavior. In R. L. Lewis, T. A. Polk & J. E. Laird (Eds.), *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 121-126). Ann Arbor, MI: University of Michigan.

Brumby, D. P., Salvucci, D. D., & Howes, A. (2009). Focus on Driving: How Cognitive Constraints Shape the Adaptation of Strategy When Dialing While Driving. In S. Greenberg, S. E. Hudson, K. Hinkley, M. R. Morris & D. R. Olsen Jr. (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1629-1638). New York, NY: ACM Press.

Byrne, M. D., & Anderson, J. R. (2001). Serial Modules in Parallel: The Psychological Refractory Period and Perfect Time-Sharing. *Psychological Review, 108*, 847-869.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Castellar, E. N., Kühn, S., Fias, W., & Notebaert, W. (2010). Outcome Expectancy and Not Accuracy Determines Posterror Slowing: ERP

Support. *Cognitive, Affective & Behavioral Neuroscience, 10*, 270-278.

Chater, N., & Vlaev, I. (2011). The Instability of Value In M. R. Delgado, E. A. Phelps & T. W. Robbins (Eds.), *Attention and Performance XXVI: Decision Making, Affect, and Learning* (pp. 81-100). Oxford: Oxford University Press.

Chong, R. S. (1998). *Modeling Dual-Task Performance Improvement: Casting Executive Process Knowledge Acquisition as Strategy Refinement (Tech. Rep. No. Cse-Tr-378-98).* PhD Thesis, University of Michigan, Ann Arbor, MI.

Chong, R. S., & Laird, J. (1997). Towards Learning Dual-Task Executive Process Knowledge Using Epic-Soar. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the cognitive science society* (pp. 107-112). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cnossen, F., Meijman, T., & Rothengatter, T. (2004). Adaptive Strategy Changes as a Function of Task Demands: A Study of Car Drivers. *Ergonomics, 47*, 218-236.

Cohen, M. X. (2008). Neurocomputational Mechanisms of Reinforcement-Guided Learning in Humans: A Review. *Cognitive, Affective & Behavioral Neuroscience, 8*, 113-125.

Cooper, R. (2010). Cognitive Control: Componential or Emergent? *Topics in Cognitive Science, 2*, 598–613.

Crowd Science (2009). Social Media Insights - Twitter.

Danielmeier, C., & Ullsperger, M. (2011). Post-Error Adjustments. *Frontiers in Psychology, 2*, 1-10.

Daw, N. D., & Frank, M. J. (2009). Reinforcement Learning and Higher Level Cognition: Introduction to Special Issue. *Cognition, 113*, 259-261.

Diels, C., Reed, N., & Weaver, L. (2009). *Drivers' Attitudes to Distraction and Other Motorists' Behaviour: A Focus Group and Observational Study*. Wokingham, UK: TRL Limited.

Duggan, G. B., & Payne, S. J. (2009). Text Skimming: The Process and Effectiveness of Foraging through Text under Time Pressure. *Journal of Experimental Psychology: Applied, 15*, 228-242.

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning among Cognitive Strategies. *Psychological Review, 112*, 912-931.

Erev, I., & Gopher, D. (1999). A Cognitive Game-Theoretic Analysis of Attention Strategies, Ability, and Incentives. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application* (pp. 343–371). Cambridge, MA: MIT Press.

Farmer, G. D. (2010). *Optimal Performance in Multitasking*. MSc thesis, University College London, London.

Farmer, G. D., Janssen, C. P., & Brumby, D. P. (2011). How Long Have I Got? Making Optimal Visit Durations in a Dual-Task Setting. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 862-867). Austin, TX: Cognitive Science Society.

Fitts, P. M. (1954). The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology, 47*, 381-391.

Foehr, U. G. (2006). *Media Multitasking among American Youth: Prevalence, Predictors and Pairings*. Menlo Park, CA: HJKF Foundation.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-Based Learning in Dynamic Decision Making. *Cognitive Science, 27*, 591-635.

González, V., & Mark, G. (2004). "Constant, Constant, Multi-Tasking Craziness": Managing Multiple Working Spheres. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 113-120). New York, NY: ACM Press.

Gopher, D. (1993). The Skill of Attention Control: Acquisition and Execution of Attention Strategies. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (pp. 299–322). Cambridge, MA: MIT Press.

Gopher, D., Brickner, M., & Navon, D. (1982). Different Difficulty Manipulations Interact Differently with Task Emphasis: Evidence for Multiple Resources. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 146-157.

Gray, W. D. (2007). *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The Soft Constraints Hypothesis: A Rational Analysis Approach to Resource Allocation for Interactive Behavior. *Psychological Review, 113*, 461-482.

Green, P. (2003). Motor Vehicle Driver Interfaces. In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook* (pp. 844-860). Hillsdale, NJ: Laurence Erlbaum Associates.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic Models of Cognition: Exploring Representations and Inductive Biases. *Trends in Cognitive Sciences, 14*, 357-364.

Guiard, Y., Olafsdottir, H. B., & Perrault, S. T. (2011). Fitts' Law as an Explicit Time/Error Trade-Off. *SIGCHI Conference on Human Factors in Computing Systems*.

Hildreth, E., Beusmans, J., Boer, E., & Royden, C. (2000). From Vision to Action: Experiments and Models of Steering Control During Driving. *Journal of experimental psychology: Human perception and performance, 26*, 1106-1132.

Hornof, A. J., & Zhang, Y. (2010). Task-Constrained Interleaving of Perceptual and Motor Processes in a Time-Critical Dual Task as Revealed through Eye Tracking. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the International Conference of Cognitive Modeling* (pp. 97-102). Philadelphia, PA: Drexel University.

Hornof, A. J., Zhang, Y., & Halverson, T. (2010). Knowing Where and When to Look in a Time-Critical Multimodel Dual Task. In S. E. Hudson, G. Fitzpatrick, W. K. Edwards, T. Rodden & E. Mynatt (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2103-2112). New York, NY: ACM Press.

Horrey, W. J., & Lesch, M. F. (2009). Driver-Initiated Distractions: Examining Strategic Adaptation for in-Vehicle Task Initiation. *Accident Analysis and Prevention, 41*, 115-122.

Horrey, W. J., Lesch, M. F., & Garabet, A. (2009). Dissociation between Driving Performance and Drivers' Subjective Estimates of Performance and Workload in Dual-Task Conditions. *Journal of Safety Research, 40*, 7-12.

Horrey, W. J., & Wickens, C. D. (2006). Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques. *Human Factors, 48*, 196-205.

Horrey, W. J., Wickens, C. D., & Consalus, K. P. (2006). Modeling Drivers' Visual Attention Allocation While Interacting with in-Vehicle

Technologies. *Journal of Experimental Psychology: Applied, 12*, 67-78.

Howes, A., Duggan, G. B., Kalidindi, K., Tseng, Y.-C., & Lewis, R. L. (in prep.). Bounded Optimal Strategies for Short-Term Remembering

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational Adaptation under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action. *Psychological Review, 116*, 717-751

Howes, A., Vera, A., & Lewis, R. L. (2007). Bounding Rational Analysis: Constraints on Asymptotic Performance. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 403–413). New York, NY: Oxford University Press.

Howes, A., Vera, A., Lewis, R. L., & McCurdy, M. (2004). Cognitive Constraint Modeling: A Formal Approach to Supporting Reasoning About Behavior. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the Cognitive Science Society* (pp. 595-600). Mahwah, NJ: Lawrence Erlbaum Associates.

Howes, A., & Young, R. M. (1997). The Role of Cognitive Architecture in Modeling the User: Soar's Learning Mechanism. *Human-Computer Interaction, 12*, 311-343.

Iqbal, S. T., & Bailey, B. P. (2007). Understanding and Developing Models for Detecting and Differentiating Breakpoints During Interactive Tasks. In B. Begole, S. Payne, E. Churchill, R. St. Armant, D. Gilmore & M. B. Rosson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 697-706). New York, NY: ACM Press.

Iqbal, S. T., & Bailey, B. P. (2010). Oasis: A Framework for Linking Notification Delivery to the Perceptual Structure of Goal-Directed Tasks. *ACM Transactions on Computer-Human Interaction, 17*, 15:11-28.

Iqbal, S. T., Ju, Y.-C., & Horvitz, E. (2010). Cars, Calls, and Cognition: Investigating Driving and Divided Attention. In S. E. Hudson, G. Fitzpatrick, W. K. Edwards, T. Rodden & E. Mynatt (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1281-1290). New York, NY: ACM Press.

Janssen, C. P. (2008). *Cognitive Models of Strategy Shifts in Interactive Behavior.* MSc Thesis, University of Groningen, Groningen.

Janssen, C. P. (2011). Understanding Multitasking as an Adaptive Strategy Selection Process. In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole & W. Kellogg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1041-1044). New York, NY: ACM Press.

Janssen, C. P. (2012). The Quest for a Unified Theory of Multitasking. The Multitasking Mind, Salvucci, Taatgen. Oxford University Press (2011). *Cognitive Systems Research, 13*, 95-97.

Janssen, C. P., & Brumby, D. P. (2010). Strategic Adaptation to Performance Objectives in a Dual-Task Setting. *Cognitive Science, 34*, 1548-1560.

Janssen, C. P., & Brumby, D. P. (2011). Design and Make Aware: Virtues and Limitations of Designing for Natural Breakpoints in Multitasking Settings. *Proceedings of Automotive User Interfaces 2011: Workshop on Cognitive Load and In-Vehicle Human-Machine Interaction*.

Janssen, C. P., Brumby, D. P., Dowell, J., & Chater, N. (2010). A Cognitively Bounded Rational Analysis Model of Dual-Task Performance Trade-Offs. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 103-108). Philadelphia, PA: Drexel University.

Janssen, C. P., Brumby, D. P., Dowell, J., Chater, N., & Howes, A. (2011). Identifying Optimum Performance Trade-Offs Using a Cognitively Bounded Rational Analysis Model of Discretionary Task Interleaving. *Topics in Cognitive Science, 3*, 123-139.

Janssen, C. P., Brumby, D. P., & Garnett, R. (2010). Natural Break Points: Utilizing Motor Cues When Multitasking. In *Proceedings of the 54th annual meeting of the Human Factors and Ergonomics Society* (pp. 482-486). San Francisco, CA: Human Factors and Ergonomics Society.

Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural Break Points: The Influence of Priorities, and Cognitive and Motor Cues on Dual-Task Interleaving. *Journal of Cognitive Engineering and Decision Making, 6*, 5-29.

Janssen, C. P., Brumby, D. P., & Howes, A. (2012). Towards a Better Understanding of Adaptive Multitasking by Individuals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts*. New York, NY: ACM Press.

Janssen, C. P., & Gray, W. D. (2012). When, What, and How Much to Reward in Reinforcement Learning Based Models of Cognition. *Cognitive Science, 36*, 333-358.

Janssen, C. P., Howes, A., & Brumby, D. P. (2012). Towards Optimal Payoff Manipulations*11th International Conference on Cognitive Modeling*. Berlin: Universitaetsverlag der TU Berlin.

Jarvstad, A., Rushton, S., Hahn, U., & Warren, P. (in press). Knowing When to Move On: Cognitive and Perceptual Decisions in Time. *Psychological Science*.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical

Contributions of Bayesian Models of Cognition *Behavioral and Brain Sciences, 34*, 169-188.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Don't Stop 'Til You Get Enough: Adaptive Information Sampling in a Visuomotor Estimation Task. In Laura Carlson, C. Hőlscher & T. F. Shipley (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2854-2859). Austin, TX: Cognitive Science Society.

Kieras, D. E. (in press). Model-Based Evaluation. In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook* (3rd ed.). London, UK: Taylor & Francis.

Kieras, D. E., & Meyer, D. E. (1997). An Overview of the Epic Architecture for Cognition and Performance with Application to Human-Computer Interaction. *Human-Computer Interaction, 12*, 391-438.

Kieras, D. E., & Meyer, D. E. (2000). The Role of Cognitive Task Analysis in the Application of Predictive Models of Human Performance. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive Task Analysis* (pp. 237-260). Mahwah, NJ: Erlbaum.

Kieras, D. E., Meyer, D. E., Ballas, J. A., & Lauber, E. J. (2000). Modern Computational Perspectives on Executive Mental Processes and Cognitive Control: Where to from Here? In S. Monsell & J. Driver (Eds.), *Attention and Performance XVIII: Control of Cognitive Processes* (pp. 681-712). Cambridge, MA: MIT Press.

Kirlik, A. (1998). The Ecological Expert: Acting to Create Information to Guide Action.*Fourth symposium on human interaction with complex systems* (pp. 15-27). Dayton, OH: IEEE Computer Society Press.

Kopecky, J. J. (2008). *Individual Strategic Differences in Human High-Speed Performance.* PhD Thesis, University of Michigan, Ann Arbor, MI.

Kurniawan, I. T., Seymour, B., Vlaev, I., Trommershauser, J., Dolan, R. J., & Chater, N. (2010). Pain Relativity in Motor Control. *Psychological Science, 21*, 840 - 847.

Lallement, Y., & John, B. (1998). Cognitive Architecture and Modeling Idiom: An Examination of Three Models of the Wickens Task. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth annual conference of the Cognitive Science Society* (pp. 597-602). Mahwah, NJ: Lawrence Erlbaum Associates.

Land, M. F., & Lee, D. N. (1994). Where We Look When We Steer. *Nature, 369*, 742-744.

Lee, J. D., & Strayer, D. L. (2004). Prefrace to the Special Section on Driver Distraction. *Human Factors, 46*, 583-586.

Levy, J., & Pashler, H. (2008). Task Prioritisation in Multitasking During Driving: Opportunity to Abort a Concurrent Task Does Not Insulate Braking Responses from Dual-Task Slowing. *Applied Cognitive Psychology, 22*, 507-525.

Lewis, R., Vera, A., & Howes, A. (2004). A Constraint-Based Approach to Understanding the Composition of Skill. In M. Lovett, C. Schunn, C. Lebiere & P. Munro (Eds.), *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 148-153). Mahwah, NJ: Lawrence Erlbaum.

Lovett, M. C. (1998). Choice. In J. R. Anderson & C. Lebiere (Eds.), *The Atomic Components of Thought* (pp. 255–296). Mahwah, NJ: Erlbaum.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A Source Activation Theory of Working Memory: Cross-Task Prediction of Performance in ACT-R. *Cognitive Systems Research, 1*, 99-118.

MacLean, A., Barnard, P., & Wilson, M. (1985). Evaluating the Human Interface of a Data Entry System: User Choice and Performance

Measures Yield Different Tradeoff Functions. In P. Johnson & S. Cook (Eds.), *People and Computers: Designing the Interface* (pp. 172-185). Cambridge, UK: Cambridge University Press.

Maloney, L. T., & Mamassian, P. (2009). Bayesian Decision Theory as a Model of Human Visual Perception: Testing Bayesian Transfer. *Visual Neuroscience, 26*, 147-155.

Maloney, L. T., Trommershauser, J., & Landy, M. S. (2007). Questions without Words: A Comparison between Decision Making under Risk and Movement Planning under Risk. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 297–313). New York, NY: Oxford University Press.

Maloney, L. T., & Zhang, H. (2010). Decision-Theoretic Models of Visual Perception and Action. *Vision Research, 50*, 2362-2374.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive Niches: An Ecological Model of Strategy Selection. *Psychological Review, 118*, 393-437.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.

Martin-Emerson, R., & Wickens, C. D. (1992). The Vertical Visual Field and Implications for the Head-up Display. In *Proceedings of the Thirty-sixth Annual Symposium of the Human Factors Society* (pp. 1408 –1412). Santa Monica, CA: Human Factors Society.

McKnight, A., & McKnight, A. (1993). The Effect of Cellular Phone Use Upon Driver Attention. *Accident Analysis & Prevention, 25*, 259-265.

Messac, A., Ismail-Yahaya, A., & Mattson, C. A. (2003). The Normalized Normal Constraint Method for Generating the Pareto Frontier. *Structural and Multidisciplinary Optimization, 25*, 86-98.

Meyer, D. E., Glass, J. M., Mueller, S. T., Seymour, T. L., & Kieras, D. E. (2001). Executive-Process Interactive Control: A Unified

Computational Theory for Answering 20 Questions (and More) About Cognitive Ageing. *European Journal of Cognitive Psychology, 13*, 123-164.

Meyer, D. E., & Kieras, D. E. (1997a). A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: Part 1. Basic Mechanisms. *Psychological Review, 104*, 3-65.

Meyer, D. E., & Kieras, D. E. (1997b). A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: Part 2. Accounts of Psychological Refractory-Period Phenomena. *Psychological Review, 104*, 749-791.

Meyer, D. E., & Kieras, D. E. (1999). Precis to a Practical Unified Theory of Cognition and Action: Some Lessons from Epic Computational Models of Human Multiple-Task Performance. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application* (pp. 17-88). Cambridge, MA: MIT Press.

Miyata, Y., & Norman, D. A. (1986). Psychological Issues in Support of Multiple Activities. In D. A. Norman & S. Draper (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 265-284). Hillsdale, NJ: Lawrence Erlbaum Associates.

Monk, C. (2004). The Effect of Frequent Versus Infrequent Interruptions on Primary Task Resumption. In *Proceedings of the 48th Human Factors and Ergonomics Society Annual Meeting* (pp. 295-299). New Orleans, LA: Human Factors and Ergonomics Society.

Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic Behavior, Workload, and Performance in Task Scheduling. *Human Factors, 33*, 607-629.

Nah, F. F.-H. (2004). A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait? *Behaviour & Information Technology, 23*, 153-163.

Navon, D., & Gopher, D. (1979). On the Economy of the Human-Processing System. *Psychological Review, 86*, 214-255.

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience Matters: Information Acquisition Optimizes Probability Gain. *Psychological Science, 21*, 960-969.

Neth, H., Khemlani, S. S., & Gray, W. D. (2008). Feedback Design for the Control of a Dynamic Multitasking System: Dissociating Outcome Feedback from Control Feedback. *Human Factors, 50*, 643-651.

Neth, H., Khemlani, S. S., Oppermann, B., & Gray, W. D. (2006). Juggling Multiple Tasks: A Rational Analysis of Multitasking in a Synthetic Task Environment. In *Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1142-1146). Santa Monica, CA: Human Factors and Ergonomics Society.

Newell, A. (1973). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. In W. G. Chase (Ed.), (pp. 283-308): Visual Information Processing.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Norman, D. A., & Bobrow, D. G. (1975). On Data-Limited and Resource-Limited Processes. *Cognitive Psychology, 7*, 44-64.

Norman, D. A., & Shallice, T. (1986). Attention to Action: Willed and Automatic Control of Behavior. In R. Davidson, G. Schwartz & D. Shapiro (Eds.), *Consciousness and Self Regulation: Advances in Research and Theory* (Vol. 4, pp. 1-18). New York, NY: Plenum Press.

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-Error Slowing: An Orienting Account. *Cognition, 111*, 275-279.

O'Hara, K., & Payne, S. (1999). Planning and the User Interface: The Effects of Lockout Time and Error Recovery Cost. *International Journal of Human Computer Studies, 50*, 41-60.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational Models of Cognition*. Oxford: Oxford University Press.

Ophir, E., Nass, C., & Wagner, A. (2009). Cognitive Control in Media Multitaskers. *Proceedings of the National Academy of Sciences, 106*, 15583-15587.

Payne, S. J., & Duggan, G. B. (2011). Giving up Problem Solving. *Memory & Cognition, 39*, 902-913.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary Task Interleaving: Heuristics for Time Allocation in Cognitive Foraging. *Journal of Experimental Psychology: General, 136*, 370-388.

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction with Information*. New York, NY: Oxford University Press.

Pirolli, P., & Card, S. K. (1999). Information Foraging. *Psychological Review, 106*, 643-675.

Pitt, M. A., & Myung, I. J. (2002). When a Good Fit Can Be Bad. *Trends in Cognitive Sciences, 6*, 421-425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a Method of Selecting among Computational Models of Cognition. *Psychological Review, 109*, 472-491.

Poulton, E. (1982). Influential Companions: Effects of One Strategy on Another in the within-Subjects Designs of Cognitive Psychology. *Psychological Bulletin, 91*, 673-690.

Rabbitt, P. (1966). Errors and Error Correction in Choice-Response Tasks. *Journal of Experimental Psychology, 71*, 264-272.

Ralph, J., Gray, W. D., & Schoelles, M. J. (submitted). Cognitive Workload and the Motor Component of Visual Attention. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Proceedings of the Annual meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Rattenbury, T., Nafus, D., & Anderson, K. (2008). Plastic: A Metaphor for Integrated Technologies. In H. Y. Youn, W.-D. Cho, J. McCarthy, J. Scott & W. Woo (Eds.), *Proceedings of the 10th international conference on Ubiquitous computing* (pp. 232-241). New York, NY: ACM Press.

Reed, M., & Green, P. (1999). Comparison of Driving Performance on-Road and in a Low-Cost Simulator Using a Concurrent Telephone Dialling Task. *Ergonomics, 42*, 1015-1037.

Reichle, E. D., & Laurent, P. A. (2006). Using Reinforcement Learning to Understand the Emergence Of "Intelligent" Eye-Movement Behavior During Reading. *Psychological Review, 113*, 390-408.

Richtel, M. (2011, Dec 15). As Doctors Use More Devices, Potential for Distraction Grows*. The New York Times,* p. A1,

Rieskamp, J., & Otto, P. E. (2006). Ssl: A Theory of How People Learn to Select Strategies. *Journal of Experimental Psychology: General, 135*, 207-236.

Roberts, M. J., & Newton, E. J. (2001). Understanding Strategy Selection. *International Journal of Human-Computer Studies, 54*, 137-154.

Roberts, S., & Pashler, H. (2000). How Persuasive Is a Good Fit? A Comment on Theory Testing. *Psychological Review, 107*, 358-367.

Rosenbaum, D. A. (1991). *Human Motor Control*. New York, NY: Academic.

Russell, S., & Subramanian, D. (1995). Provably Bounded-Optimal Agents. *Journal of Artificial Intelligence Research, 2*, 575-609.

Salvucci, D. D. (2001). Predicting the Effects of in-Car Interface Use on Driver Performance: An Integrated Model Approach. *International Journal of Human Computer Studies, 55*, 85-107.

Salvucci, D. D. (2005). A Multitasking General Executive for Compound Continuous Tasks. *Cognitive Science, 29*, 457-492.

Salvucci, D. D., & Beltowska, J. (2008). Effects of Memory Rehearsal on Driver Performance: Experiment and Theoretical Account. *Human Factors, 50*, 834-844.

Salvucci, D. D., & Bogunovich, P. (2010). Multitasking and Monotasking: The Effects of Mental Workload on Deferred Task Interruptions. In S. E. Hudson, G. Fitzpatrick, W. K. Edwards, T. Rodden & E. Mynatt (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Salvucci, D. D., & Gray, R. (2004). A Two-Point Visual Control Model of Steering. *Perception, 33*, 1233-1248.

Salvucci, D. D., & Macuga, K. L. (2002). Predicting the Effects of Cellular-Phone Dialing on Driver Performance. *Cognitive Systems Research, 3*, 95-102.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded Cognition: An Integrated Theory of Concurrent Multitasking. *Psychological Review, 115*, 101-130.

Salvucci, D. D., & Taatgen, N. A. (2011). *The Multitasking Mind*. New York, NY: Oxford University Press.

Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a Unified Theory of the Multitasking Continuum: From Concurrent Performance to Task Switching, Interruption, and Resumption. In S. Greenberg, S. E. Hudson, K. Hinkley, M. R. Morris & D. R. Olsen Jr.

(Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1819-1828). New York, NY: ACM Press.

Schultz, W. (2006). Behavioral Theories and the Neurophysiology of Reward. *Annual Review of Psychology, 57*, 87-115.

Schumacher, E., Lauber, E., Glass, J., Zurbriggen, E., Gmeindl, L., Kieras, D. E., & Meyer, D. E. (1999). Concurrent Response-Selection Processes in Dual-Task Performance: Evidence for Adaptive Executive Control of Task Scheduling. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 791-814.

Schwartz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*, 461-464.

Siegler, R. S. (1991). Strategy Choice and Strategy Discovery. *Learning and Instruction, 1*, 89-102.

Siegler, R. S., & Lemaire, P. (1997). Older and Younger Adults' Strategy Choices in Multiplication: Testing Predictions of Ascm Using the Choice/No-Choice Method. *Journal of Experimental Psychology: General, 126*, 71-92.

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review, 63*, 129-138.

Smith, M. R., Lewis, R. L., Howes, A., Chu, A., & Green, C. (2008). More Than 8,192 Ways to Skin a Cat: Modeling Behavior in Multidimensional Strategy Spaces. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1441-1446). Austin, TX: Cognitive Science Society.

Sperling, G., & Dosher, B. A. (1986). Strategy and Optimization in Human Information Processing. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance:*

Volume 1. Sensory Processes and Perception (pp. 2-1 - 2-65). New York, NY: Wiley-Interscience.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*. Princeton, NJ: Princeton University Press.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by Sampling. *Cognitive Psychology, 53*, 1-26.

Strayer, D. L., Drews, F., & Johnston, W. (2003). Cell Phone-Induced Failures of Visual Attention During Simulated Driving. *Journal of Experimental Psychology: Applied, 9*, 23-32.

Strayer, D. L., & Johnston, W. A. (2001). Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychological Science, 12*, 462-466.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Swets, J. A., Tanner, W., & Birdsall, T. (1961). Decision Processes in Perception. *Psychological Review, 68*, 301-340.

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The Acquisition of Robust and Flexible Cognitive Skills. *Journal of Experimental Psychology: General, 137*, 548-565.

Taatgen, N. A., & Van Rijn, H. (2010). Nice Graphs, Good R2, but Still a Poor Fit? How to Be More Sure Your Model Explains Your Data In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 247-252). Philadelphia, PA: Drexel University.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye Guidance in Natural Vision: Reinterpreting Salience. *Journal of Vision, 11*, 1-23.

The Economist (2011, Apr 13). Think before You Speak: Distracted Driving Is the New Drunk Driving.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2003a). Statistical Decision Theory and the Selection of Rapid, Goal-Directed Movements. *Journal of the Optical Society A, 20*, 1419-1433.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2003b). Statistical Decision Theory and Trade-Offs in the Control of Motor Response. *Spatial Vision, 16*, 255-275.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2008). Decision Making, Movement Planning and Statistical Decision Theory. *Trends in Cognitive Sciences, 12*, 291-297.

Tseng, Y.-C., Howes, A., & Lewis, R. L. (submitted). Bounded Optimal Reward Rate in the Multiple-Alternative Adaptive Decision Task.

Vera, A., Howes, A., McCurdy, M., & Lewis, R. (2004). A Constraint Satisfaction Approach to Predicting Skilled Interactive Cognition. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 121-128). New York, NY: ACM Press.

Vlaev, I., Chater, N., & Stewart, N. (2007). Financial Prospect Relativity: Context Effects in Financial Decision-Making under Risk. *Journal of Behavioral Decision Making, 20*, 273-304.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. A. (2011). Does the Brain Calculate Value? *Trends in Cognitive Sciences, 15*, 546-554.

Walsh, M. M., & Anderson, J. R. (2009). The Strategic Nature of Changing Your Mind. *Cognitive Psychology, 58*, 416-440.

Wang, D. D., Proctor, R. W., & Pick, D. F. (2007). Acquisition and Transfer of Attention Allocation Strategies in a Multiple-Task Work Environment. *Human Factors, 49*, 995-1004.

Wang, D. D., Proctor, R. W., & Pick, D. F. (2009). Allocation of Effort as a Function of Payoffs for Individual Tasks in a Multitasking Environment. *Behavior Research Methods, 41*, 705-716.

Warren, P., Graf, E., Champion, R., & Maloney, L. T. (2012). Visual Extrapolation under Risk: Human Observers Estimate and Compensate for Exogenous Uncertainty. *Proceedings of the Royal Society Biological Sciences, 279*, 2171-2179.

Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in Extraordinary Multitasking Ability. *Psychonomic Bulletin & Review, 17*, 479-485.

Wickens, C. D. (2002). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science, 3*, 159-177.

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors, 50*, 449-455.

Wobbrock, J., Cutrell, E., Harada, S., & MacKenzie, I. (2008). An Error Model for Pointing Based on Fitts' Law. In M. Burnett, M. F. Costabile, T. Catarci, B. De Ruyter, D. Tan, M. Czerwinski & A. Lund (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1613-1622). New York, NY: ACM Press.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2011). The Neural Correlates of Subjective Utility of Monetary Outcome and Probability Weight in Economic and in Motor Decision under Risk. *Journal of Neuroscience, 31*, 8822-8831.

Young, R. M. (1978). Strategies and the Structure of Cognitive Skill. In G. Underwood (Ed.), *Strategies of Information Processing* (pp. 357-401). London: Academic Press.

Yu, S. (2011). *Sub-Optimal Multitasking: Do People Interleave Tasks at a Natural Boundary Even When It Is Not Optimal to Do So?* BSc Thesis, University College London, London.

Zhang, H., Daw, N., & Maloney, L. T. (2011). Testing Whether Humans Have an Accurate Model of Their Own Motor Uncertainty in a Speeded Reaching Task. *Journal of Vision, 11*, 953-953.