ECONOMIC BEHAVIOUR & PSYCHOLOGICAL BIASES *in* HUMAN–COMPUTER INTERACTION

A Thesis

Submitted in Partial Fulfilment

of the Requirements for the Degree

of

Doctor of Philosophy

in the

University of Canterbury

by

Philip Quinn

Department of Computer Science and Software Engineering

University of Canterbury

2016

‘I can’t help but feel I’ve been ripped off . . .
I’m sure you’re feeling something similar.’

—— Guybrush Threepwood
The Secret of Monkey Island

CONTENTS

PUBLICATIONS

A portion of the work in this thesis (specifically, Experiment 5.1) was published in a peer-reviewed publication:

- Quinn, P., & Cockburn, A. (2016). When bad feels good: Assistance failures and interactive preferences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4005–4010). New York, NY: ACM. doi: 10.1145/2858036.2858074.

The model and other experimental work in this thesis (Chapters 2–4) are currently under review.

Two other publications contain ideas and experiments that were formative to the ideas in this thesis. They embody many of the concepts presented herein, but were conducted apart. Several of the ideas presented in the first are carried in Experiment 6.1, and the second is reproduced in Appendix B.

- Cockburn, A., Quinn, P., & Gutwin, C. (2015). Examining the peak-end effects of subjective experience. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 357–366). New York, NY: ACM. doi: 10.1145/2702123.2702139.
- Quinn, P., & Zhai, S. (2016). A cost–benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 83–88). New York, NY: ACM. doi: 10.1145/2858036.2858305.

I began my doctoral studies with research on issues associated with low-level mechanical interactions (in particular, those associated with scrolling), resulting in the following peer-reviewed publications:

- Cockburn, A., Quinn, P., Gutwin, C., & Fitchett, S. (2012). Improving scrolling devices with document-length-dependent gain. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 267–276). New York, NY: ACM. doi: 10.1145/2207676.2207714.
- Quinn, P., Cockburn, A., Casiez, G., Roussel, N., & Gutwin, C. (2012). Exposing and understanding scrolling transfer functions. In *Proceedings of the 25th annual ACM symposium on user interface software and technology* (pp. 341–350). New York, NY: ACM. doi: 10.1145/2380116.2380161.
- Quinn, P., Malacria, S., & Cockburn, A. (2013). Touch scrolling transfer functions. In *Proceedings of the 26th annual ACM symposium on user interface software and technology* (pp. 61–70). New York, NY: ACM. doi: 10.1145/2501988.2501995.
- Malacria, S., Aceituno, J., Quinn, P., Casiez, G., Cockburn, A., & Roussel, N. (2015). Push-edge and slide-edge: Scrolling by pushing against

the viewport edge. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2773–2776). New York, NY: ACM. doi: 10.1145/2702123.2702132.

This evolved into research on the costs and benefits of low-level interactions that have cognitive trade-offs, resulting in the following publications:

- Quinn, P., Cockburn, A., Räihä, K.-J., & Delamarche, J. (2011) On the costs of multiple trajectory pointing methods. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 859–862). New York, NY: ACM. doi: 10.1145/1978942.1979067
- Quinn, P., Cockburn, A., & Delamarche, J. (2013). Examining the costs of multiple trajectory pointing techniques. *International Journal of Human-Computer Studies, 71*(4), 492–509. doi: 10.1016/j.ijhcs.2012.12.001.

Finally, my research settled on investigation of the general costs and benefits of interactive techniques in an economic sense (the present work).

DECLARATION

The work contained in this thesis is my own. The inspiration for the work's direction was provided by my senior supervisor, Professor Andy Cockburn. His guidance and collaboration are fundamental to the substance of this thesis.

FUNDING

Portions of this thesis were funded by a Royal Society of New Zealand Marsden Grant (10-UOC-020) and a UC Doctoral Scholarship.

ETHICS

All experiments were conducted at the University of Canterbury under Human Ethics Committee programme code HEC 2015/11/LR-PS. Subjects completed a written consent form, reproduced in Appendix A.

ABSTRACT

Research on human judgement and decision making has documented a wide-range of psychological biases in how people perceive, evaluate, and ultimately decide between alternatives. These biases are understood in a contrast with normative economic principles of utility and risk, which reduce alternatives to quantitative variables and define axiomatic frameworks to identify the optimal choice. Understanding and modelling the incongruity between normative economic principles and empirical human behaviour has been the focus of behavioural economics research, which aims to align economic models with actual behaviour. Such research has been productive in developing an understanding of human decision-making behaviour across investment decisions, consumer purchasing behaviour, job-offer negotiation, and academic performance – but there has been little investigation in its application to human–computer interaction problems.

Users of interactive systems encounter decisions that share many properties with those studied in the behavioural economics literature. In particular, users often choose to invest actions and effort into using an interface, anticipating that they will receive a return of at least commensurate value in productivity. Models from behavioural economics have the potential to describe how users evaluate such interfaces, and how they decide between them. For example, when faced with a text-entry interface that automatically replaces incorrect words, such models could predict the utility of instances where a correct replacement happens, where an incorrect replacement happens, and the influence of these events on the preferences held by the user. These are important issues for the designers of such interfaces, but are currently poorly understood and under-investigated.

This thesis adapts a model of reference-dependent preferences from the behavioural economics literature to interaction. The model predicts the relationship between an outcome that a user experiences and their evaluation of that outcome via measures of utility. In particular, the model emphasises the importance of salient positive progress towards a task goal.

Tests of the model's predictions are presented in two experiments that involved simple text-selection tasks using either a conventional letter-by-letter selection technique, or a technique that attempted to assist the user by snapping their selection to word boundaries. The experiment found a negativity bias: small components of negative progress (when the attempted assistance failed) overwhelmed subjects' assessment of overall utility, and substantial objective performance (time) gains were required to overcome this assessment.

The second experiment found that this effect was neutralised by manipulating the interface's behaviour to appear more helpful – even though it contained the same performance disadvantages. A new task (drag-and-drop) was developed for a third experiment that extended this manipulation, and found a positivity bias: subjects preferred small elements of positive progress, despite substantial objective performance losses.

The methodology developed for these experiments has subjects make binary choices between an experimental interface and a neutral reference condition. From a series of these binary choices, a utility/preference scale can be constructed to identify biases in the subjects' preferences with respect to a manipulated objective variable. The flexibility of this method for analysing interactive choices is demonstrated in a fourth experiment which examined two psychological biases that are not predicted by the model – the peak-end rule and duration neglect – and found some support for their presence during interactive experiences.

The main contribution of this thesis is the demonstration of a connection between human–computer interaction research and the behavioural economics literature: that the psychological biases and economic principles described in prior work are present and applicable to interactive tasks. The model and experimental methodology also provide a robust foundation for improving practice in human–computer interaction research on understanding user behaviour and experience.

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:

Parts of Chapters 2–4 and 7.

Quinn, P., & Cockburn, A. (2015). Loss aversion and perceived value in interaction. Unpublished manuscript.

Please detail the nature and extent (%) of contribution by the candidate:

Mr Quinn implemented all programs and data analysis software. He conducted the entire literature search and wrote it up. He devised the data analysis method. Final paper authorship was a collaborative effort. Total estimation of Mr Quinn's contribution: 80%.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all.
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work.
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text.

Name: *Andy Cockburn*
Signature: *Andy Cockburn*
Date: *21-1-2016*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:

Parts of Section 2.5 and Chapter 6.

Cockburn, A., Quinn, P., & Gutwin, C. (2015). The peak-end rule in user preferences. Unpublished manuscript.

Please detail the nature and extent (%) of contribution by the candidate:

Mr Quinn wrote all software for implementing and analysing Experiment 2. The design of the experiment was a collaborative effort. Mr Quinn wrote the majority of the related work section, and all authors collaborated on producing the final paper. Estimated contribution: 40%

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all.
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work.
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text.

Name: *Andy Cockburn*
Signature: *Andy Cockburn*
Date: *21-1-2016*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:

Parts of Chapter 5.

Quinn, P., & Cockburn, A. (2016). When bad feels good: Assistance failures and interactive preferences. In Proceedings of the SIGCHI conference on human factors in computing systems. New York, NY: ACM. doi: 10.1145/2858036.2858074.

Please detail the nature and extent (%) of contribution by the candidate:

Mr Quinn wrote all software for conducting the experiment and analysing the data. He designed nearly all aspects of the experiment. He drafted the entire paper. Estimated contribution: 80%.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all.
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work.
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text.

Name: *Andy Cockburn*
Signature: *Andy Cockburn*
Date: *21-1-2016*

INTRODUCTION 1

INTERACTION with user interfaces can be expressed as a chain of user actions and system responses: a user presses a key and a character appears, they double-click on a file and it opens in an editor, they drag an icon and it follows their cursor. Compositions of these primitive actions are used to complete surprisingly complex tasks: from word processing and financial accounting, to photo manipulation and film editing. Computer systems also have the opportunity to actively *assist* users in completing their tasks – for example, a text entry system might automatically replace a mistyped *teh* with *the*, a drawing tool might facilitate object alignment by snapping dragged objects to neighbouring ones, and a web browser might aid URL entry by searching the user's browsing history and suggesting addresses.

User interface and interaction designs are often analysed and assessed by researchers and practitioners in objective terms: the number of steps required to complete a task, the time for the system to respond, the speed of users in performing actions, and the frequency with which they make errors. Assistive interfaces are often designed and deployed on the basis that they provide opportunities to improve these metrics and make user interactions more productive. However, a user's experience of interaction is rarely so disinterested: users are annoyed when a system takes a long time to respond, they distrust systems that impair their productivity, and they form preferences that cannot be articulated with objective qualities. Although designers intend that a system's assistance will increase a user's efficiency and productivity on balance – inevitably, a system's attempts to assist will periodically *impede*. Despite positive average objective performance, these impediments can have a negative impact on subjective experience in excess of the productivity loss. For example, an automatic correction of *harnomy* to *harmony* is clearly satisfying, but an incorrect replacement of *tome* with *time* can be unexpected and aggravating – the frustration of an incorrect replacement is not always balanced by the satisfaction brought by a comparable correct replacement.[1]

1. See http://theoatmeal.com/comics/autocorrect for a humorous personification of this autocorrection user experience.

This asymmetry in experience is known more generally in psychology as *negativity bias* or in economics as *loss aversion*, where it has been found that negative outcomes are consistently more potent than their positive counterparts. With few exceptions, negative experiences dominate positive experiences of equal measure across domains such as learning, investment, personal relationships, memory, and impression formation (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). The sentiment is that

'losses loom larger than gains' (Kahneman & Tversky, 1979, p. 279). For example, if offered a gamble on a coin-flip that wins $150 on tails and loses $100 on heads, most people refuse because the *psychological value* of potentially losing $100 dominates the potential gain of $150 (Tversky & Kahneman, 1992). Such research has also uncovered other psychological biases in the ways that people perceive, evaluate, and choose between alternatives. The findings have been applied to understand behaviour across varied choice domains: whether it be monetary gambles, consumer goods, medical treatments, or job prospects. The conflict these biases create between objective analyses of decisions and the often-illogical behaviour when people actually encounter them are captured in models of utility that describe the relationship between the objective and subjective factors of decision making.

1.1 RESEARCH HYPOTHESIS

> *Certain cognitive biases – revealed in the judgement and decision-making literature – apply to users during interaction, and their effects can be captured in economic models of utility. Establishing a relationship between these fields of research can lead to improved design and research practice in human–computer interaction.*

This thesis contends that the psychological biases found in people's monetary, cognitive, and interpersonal (etc.) evaluations are equally applicable to how users perceive, evaluate, and choose to use interactive systems. For instance, the psychological loss aversion a person feels when faced with the prospect of losing money is also felt when a system's assistance can cost them time and productivity. Furthermore, economic models of utility and reference-dependent preferences can be operationalised in human–computer interaction contexts to understand and explain decisions that result from the influence of such biases. This has implications for both the design of interactive systems and their evaluation: in particular, although it is tempting to focus on the positive outcomes that a system can provide, it is the negative outcomes that are potentially more formative of overall user experience. However, this biased aversion to negative outcomes can be tempered by crafting experiences to balance the most salient aspects that influence user satisfaction.

1.2 RESEARCH CONTRIBUTIONS

The breadth and scope of the prior psychological and economic work – spanning over half a century of productive research – makes it unrealistic to comprehensively establish the presence, nature, and extent of its relationship with

human–computer interaction in a single research agenda. Therefore, this thesis focusses on establishing parallels with several fundamental principles from the literature that together form a compelling argument for its broad applicability in human–computer interaction and a robust basis for future work. This is done through the following specific contributions:

- A model of reference-dependent preferences is adapted from the behavioural economics literature to the domain of interactive actions and system responses (Chapter 3). This utility model is based on a formalisation of several key psychological biases (namely, loss aversion and reference-dependence), and describes (a) the components of an interaction and (b) how the relationship between those components informs preferences between alternatives.
- Methodological considerations for assessing utility models and psychological biases for interactive tasks are discussed (Chapter 3). These include: (a) how to construct experimental manipulations that test utility, (b) how to expose subjects to those manipulations and elicit their preferences, and (c) how to analyse those preferences to assess utility models. These issues are not specific to the model outlined above but are generally applicable to work that investigates decision-making behaviour and subjective preferences in interaction.
- To establish support for features of the model, two experiments are presented (Experiments 4.1 and 4.2). The experiments explored subjects' loss aversion towards an interface that factually assisted their performance but included an element of interactive progress loss (i.e. appeared to contain a loss of productivity). The first experiment establishes the presence of a loss aversion, while the second finds that it can be neutralised by manipulating the perceived progress loss (while maintaining objective outcomes).
- A third experiment (5.1) further studied this effect using a new experimental task that appeared to contain progress advantages, but had an objective performance disadvantage. The results find subjects exhibited a bias in favour of the interface's perceived progress gains – despite experiencing significant objective performance losses.

  These three experiments can be understood within the parameters of the model. Table 1.1 summarises their designs (the objective outcome given to subjects) and their results (the subjective preference observed).
- Finally, the generality of the choice-based methodology is demonstrated with an experiment that tested psychological biases that are not captured by the model: the peak-end rule and duration neglect (Experiment 6.1). The experiment demonstrates the methodological principles

TABLE 1.1
Summary of experimental designs and results.

		Objective Outcome		
		Negative	Neutral	Positive
Subjective Preference	Negative	—		4.1 (Negativity Bias)
	Neutral		4.2 (Unbiased)	
	Positive	5.1 (Positivity Bias)		—

in a context beyond the scope of the model, and finds some support for the existence of other psychological biases in interaction.

1.3 THESIS OUTLINE

The remainder of this thesis is organised as follows. Chapter 2 reviews the extensive economic and psychological research on judgement and decision making. This includes perspectives from normative economic models that aim to calculate correct and rational behaviour, and the psychological biases and heuristics that permeate people's intuitions and empirical behaviour. Chapter 3 adapts a model from the literature to the context of human–computer interaction. This involves an analysis of how interactive tasks can be decomposed and user choices examined. The model describes the relationships between user actions, system responses, and preferences. A choice-based methodology for testing hypotheses about this model (and psychological biases in general) is also described.

Chapter 4 examines the loss aversion and reference-dependence components of the model in two experiments that demonstrate their presence and isolate their cause – in particular, that losses in perceived progress towards a task goal weigh heavily on people's subjective choices (regardless of objective performance). Chapter 5 presents an experiment that uses a different interactive task to further test the model in a design that finds subjective preferences in favour of an interface that is objectively poor, but perceptually positive.

The new task and the choice-based methodology are used in Chapter 6 to examine two psychological biases that are not captured by the model: the peak-end rule and duration neglect effects.

Finally, Chapters 7 and 8 reflect on this work and its implications. Chapter 7 places this research into broader literature and discusses its applications, limitations, and suggests a number of directions for continuing to explore these issues. Chapter 8 summarises the main findings and contributions.

JUDGEMENT & DECISION MAKING 2

This chapter reviews economic and psychological work on judgement & decision making relevant for building and testing a model of choice for human–computer interaction decisions.

UNDERSTANDING how people perceive, evaluate, and ultimately decide between alternative choices (e.g. which product to purchase in a store or where to go on vacation) is the focus of psychological research on *judgement and decision making*. Early work was influenced by behavioural economics research that sought to understand similar issues for economic choices (e.g. investment or gambling decisions), and the discrepancies economists observed between how they thought people *ought* to behave and how people *actually* do (Edwards, 1954c). For example, people exhibit an *aversion to losses* that is inconsistent with the simple economic model of expected value, but is consistent with the idea that losses are psychologically more potent than gains. Later research has generalised these findings to other choice domains and continues to investigate the processes that elicit them.

Theories of judgement and decision making consider two components of a choice between alternatives: (a) how people assess the *value* of an outcome (its attractiveness), and (b) how people assess the *risk* of obtaining that value. Enquiry into these components has its origins in seventeenth century philosophical and mathematical treatises on utility and probability, with modern academic research stimulated by John von Neumann (1903–57) and Oskar Morgenstern's (1902–77) *game theory*, and the psychological paradoxes highlighted by Maurice Allais (1911–2010) in the 1950s. The economics side of the literature has sought to find functional relationships between the objective properties of these components (such as monetary value and probability) and the responses people should make to maximise their returns (reviewed by Arrow, 1958; Camerer, 1995, 1998; Fehr-Duda & Epper, 2012; Machina, 1987; Rabin, 1998; Starmer, 2000). The psychological side has sought to understand the perceptual and cognitive processes that underly the choices observed from human subjects (Edwards, 1954c, 1961; Pitz & Sachs, 1984; Slovic, Fischhoff, & Lichtenstein, 1977; E. U. Weber & Johnson, 2009). In concert, this body of work has developed rich insights into vagaries of human decision-making behaviour (DellaVigna, 2009; Gigerenzer & Gaissmaier, 2011; Hastie, 2001; Kahneman, Knetsch, & Thaler, 1991; Tversky & Kahneman, 1974).[1]

1. See Wakker (2015) for an extensive annotated bibliography.

However there is also a tension between these fields (Bell, Raiffa, & Tversky, 1988; Hogarth & Reder, 1986a; Lopes, 1994; Rabin, 1998; Simon, 1959). Economists typically aspire for people to make stable and cogent choices that maximise some definite return, and develop normative rules and models that prescribe people to these ideals. Behaviour that does not conform is deemed

to be *irrational*. On the other side, psychologists do not diagnose violations of economic rules as abnormal, and develop models to account for whatever behaviour is observed. Behaviour that does not conform renders the model *invalid*. This dualism presents a dilemma for practitioners who want to apply these models to improve people's success and performance: what is the balance between telling people the choices they *should* be making, or adapting to the choices they *want* to make?

Understanding the history of this divergence in economic and psychological approaches is important for reading current theory and practice, and for applying it to new kinds of decision problems. Compared with the field of judgement and decision making,[2] human–computer interaction is a fledgling field that is only beginning to enquire into the nature of its user's choices: how users make interactive choices, and how interaction can be used to improve and support decision-making behaviour. The breadth and depth of judgement and decision making research can be a significant aid to understanding these questions if the right parallels between the fields can be established. In discussing how to improve the synergy between economics and psychology, Rabin (2013) argued that the most useful theories are those that (a) extend existing models with new parameters and (b) define them across domains with measurable variables. This thesis adopts a similar approach in order to place human–computer interaction research within this literature.

2. These areas of research are named 'judgement and decision making' in psychology and 'behavioural economics' in economics. Here these terms are all-encompassing, and differences between the two fields are noted explicitly.

This chapter reviews the development of judgement and decision making research with a focus on the theories and debates that have shaped the research agenda and exposed pitfalls that may impede or confound new applications. This begins with a discussion of the judgement of human behaviour as either rational or irrational (§2.1), and then reviews the development of economic theories of utility and the influence of psychological biases on people's preferences. This includes the quantification of how people interpret and assess the value and risk of outcomes (§2.2 and §2.3), the psychological factors that induce systematic biases (§2.5 and §2.6), and the broader decision-making process (§2.4). Finally, the theorised and applied uses of this work in human–computer interaction research are reviewed (§2.7). Much of this work examines monetary decisions due to their convenient units that are easy to experimentally manipulate (e.g. it is obvious that a reward of $2 is better than $1), but the principles behind those decisions are believed to be general – and in particular, the following chapter develops an interactive context for those principles and reviews the methodological issues for evaluating them.

2.1 RATIONAL BEHAVIOUR

Models of human behaviour can be framed as either *normative* or *descriptive*. Normative models explain how people *ought* to behave under a set of logical

axioms that can be used to deduce which alternative is the optimal choice; whereas descriptive models aim to capture how people *actually* behave by observing their decisions and postulating about the decision-making process that led to them (Becker & McClintock, 1967; Bell et al., 1988; March, 1978). This is not a dichotomy, however, and models often straddle both sides: normative models do not want to be ignorant of actual human behaviour, and descriptive models want to be able to distinguish good judgement from bad. Rather, the division is in how researchers apply these models, interpret experimental data, and define *rational behaviour*: the criteria for labelling a person's decisions as either *sensible* or *foolish*.[3]

3. There is also a significant philosophical discussion of this question, reviewed by Manktelow and Over (1993), Stigler (1950b), and Briggs (2014).

2.1.1 Normative Models

Normative models structure decision-making behaviour around a set of axioms that define a formal system for expressing outcomes (reviewed by Becker & McClintock, 1967; Camerer, 1995; Fishburn, 1979; Luce & Suppes, 1965). Using quantifiable variables that describe alternative outcomes (such as monetary values and probability scores), a decision can be expressed with a set of well-formed formulae, and an optimal choice can be discovered through the model's rules of inference.

Notation. Outcomes are labelled A, B, C, and so forth, from some set of mutually exclusive alternatives $\mathcal{S}$. Each outcome is conditional on a probability $p \in [0,1]$, which sums to 1 across $\mathcal{S}$. The binary relation $\succ$ denotes preference between outcomes: $A \succ B$ indicates A is strictly preferred to B; whereas $\sim$ denotes the absence of a preference: $A \sim B$ indicates an indifference between A and B. These relations can be combined as $A \succcurlyeq B$ if either $A \succ B$ or $A \sim B$ is the case, but are otherwise incompatible (i.e. $A \succ B$ only if $A \not\sim B$). Only $\sim$ is symmetric ($A \sim B$ if and only if $B \sim A$) and reflexive ($A \sim A$). Other properties that are sometimes ascribed to these objects and relations are reviewed by Fishburn (1986), and Hansson & Grüne-Yanoff (2012).

Some of the axioms that are generally pursued by normative models include (see also Kreps, 1988):

- *Comparability* or *Completeness*. For any two outcomes A and B, where $A \neq B$, either $A \succ B$, $B \succ A$, or $A \sim B$ (if $A = B$, then $A \sim B$).
- *Transitivity*. If $A \succcurlyeq B$ and $B \succcurlyeq C$, then $A \succcurlyeq C$.[4,5]
- *Continuity*. For any $A \succcurlyeq B \succcurlyeq C$, there is a unique probability p such that $pA + (1-p)C \sim B$.
- *Independence*. If $A \succcurlyeq B$, then $pA + (1-p)C \succcurlyeq pB + (1-p)C$ for any C.
- *Dominance* or *Sure-thing*. If A is sure to yield more than B,[6] then $A \succ B$.

6. e.g. if for some utility (§2.2) function u, $u(A) > u(B)$.

4 This is known as the transitivity of *weak preference* to distinguish it from the transitivity of *indifference* or *strict preference* when applied with the $\sim$ and $\succ$ operators, respectively – although the transitivity of $\sim$ is debated (Schumm, 1987).

5 It also follows that $[(A \sim B) \wedge (B \succ C)] \Rightarrow (A \succ C)$; and similarly $[(A \succ B) \wedge (B \sim C)] \Rightarrow (A \succ C)$.

- *Irrelevance* or *Invariance*. If two actions result in the same outcome, preference should not be influenced by any factors that do not alter the outcomes, and desires for or against an outcome should not affect beliefs about its likelihood.

Models that accept such axioms (reviewed by Fishburn, 1981) discuss alternatives in terms of quantitative scores (i.e. their utility and probability). Those alternatives that maximise these scores are the *rational* choices for an individual to select (the so-called *economic man* [Edwards, 1954c][7]).

7. And perhaps *economic pigeons*, too (Battalio, Kagel, Rachlin, & Green, 1981).

Normative models are attractive to economists because they define a set of quantifiable variables that should influence choices, and an algebra for discriminating and comparing alternatives impartially. The choices made under these models can be deductively proven to satisfy certain properties that are argued to epitomise good decision making (such as maximising investment returns). Violating these axioms would create logical inconsistencies or arbitrage opportunities that would lead a person to ruin (reviewed by Hansson & Grüne-Yanoff, 2012).

Although there is some support for the application of these axioms (see Plott, 1986), empirical data finds that actual decision-making behaviour frequently violates them (reviewed by Anand, 1993; Becker & McClintock, 1967; MacCrimmon & Larsson, 1979; Shafir & LeBoeuf, 2002; Tversky & Kahneman, 1986, and with examples throughout this chapter). This is not surprising to psychologists, who readily accept that 'human beings are neither perfectly consistent nor perfectly sensitive' (Edwards, 1954c, p. 388) and 'make mistakes, have remorse, suffer anxieties, and cannot make up their minds' (Bell et al., 1988, p. 9). But such data raise questions about the validity and interpretation of normative models: are violations a failure of people to behave correctly, or a failure of a model's axiomatic construction?

Tversky (1969) presented subjects with pairs of gambles that were constructed to test the transitivity of preferences by gradually increasing the probability of winning, while decreasing the pay-off – but at a rate that increased the overall expected value (reviewed in §2.2.1). He found that when the difference between gambles was small, subjects preferred the one with the higher pay-off;[8] but when the difference was large, subjects preferred the gamble with the higher expected value[9] – a violation of transitivity once all of the gambles were sorted by these choices.

8. e.g. between a 7/24 chance at $5 and a 8/24 chance at $4.75, subjects preferred the former.
9. e.g. between a 7/24 chance at $5 and a 11/24 chance at $4, subjects preferred the latter.

From a normative perspective these violations are poor choices that ought to be corrected; but from a descriptive perspective they are simply the random errors of human fallibility (Einhorn & Hogarth, 1981; March, 1978), or the model's neglect for some other aspect of the decision (Pitz & Sachs, 1984; Regenwetter, Dana, & Davis-Stober, 2011). However, normative models are at least inspired by observed human behaviour, and if a person's actual behaviour

deviated too far from a model's behavioural axioms, then their choices could be exploited by people who follow them (Bell et al., 1988). In their account of game theory – the most influential axiomatic model – von Neumann and Morgenstern (1947) argued that 'each axiom should have an immediate intuitive meaning by which its appropriateness may be judged directly' (p. 25).

However, even when subjects are instructed on the rationale behind normative axioms, they reject the axioms' validity and application. For example, Slovic and Tversky (1974) asked subjects for their preferences in two gambling problems designed to expose a violation of the independence axiom. One of the problems is known as the *Ellsberg paradox* (Ellsberg, 1961):

> An urn contains 90 balls: 30 are red, and the remaining 60 are either black or yellow in some unknown proportion. One ball will be drawn, and you may bet according to I or II below. Then imagine the scenario again, under the same circumstances, but according to bets III or IV.

	30	60	
	Red	Black	Yellow
I	$100	$0	$0
II	$0	$100	$0
III	$100	$0	$100
IV	$0	$100	$100

The experimenters explained the arguments for both conforming to[10] and violating[11] the independence axiom,[12] and collected subjects' evaluations of the arguments and their choices to the gambles. They found opinions on each argument's logical persuasiveness varied, and actual choices generally did not conform to the axiom. That is, even when educated on the axioms that would lead to consistent behaviour, subjects were not persuaded that they would lead to the best choice.

10. Choosing either I and III, or II and IV.

11. Choosing either I and IV, or II and III.

These conflicts arise because normative models lack variables to account for the subjective tastes, preferences, goals, and desires that have strong intuitive effects on human decision-making behaviour. For proponents of normative models, these are intentional omissions: introducing subjective factors that cannot be measured or reasoned about only vindicates errors that need to be corrected, and risks formulating models that are tautological. However, responding to frequent empirical violations, economists have explored

12 Note that the outcomes for a yellow ball are the same between I/II and III/IV, thereby making this outcome irrelevant to the decision (no matter which bet is chosen, the drawing of a yellow ball results in the same payout). Once the outcomes for a yellow ball are removed from the table, I and II are identical to III and IV, respectively, and therefore should elicit consistent responses.

diverse approaches for interpreting and redressing axiomatic models – while retaining the ability to rationalise decisions (reviewed by Anand, 1993):

- Savage (1972) admitted that humans make mistakes in their application of logic, but insisted that they could – and should – be corrected when brought to light. That is, such models rely on their logical consistency and not empirical verification (see also Raiffa, 1961).
- Some models relaxed, weakened, or removed dependence on troublesome axioms (e.g. Chew, 1983; Machina, 1982); and others attempted to include psychological factors, such as disappointment (Bell, 1982; Gul, 1991), or interpret the axioms stochastically (Luce, 1958).
- Conversely, Anand (1987) argued that the completeness, transitivity, and independence axioms are unnecessary and do not imply rationality at all (see also Fishburn, 1988).
- Machina (1989) proposed a compromise: consistency with the axioms is desirable at some deep level of decision making, but this is unobservable through economic methods. Therefore, there is value in descriptive models that explain the factors of choice from observable behaviour.

However, revised models eventually have to justify why the empirical data that contradicts their predictions should be labelled *irrational*. In a review of the cat-and-mouse game between normative models and empirical data, Tversky and Kahneman (1986) remarked that 'the dream of constructing a theory that is acceptable both descriptively and normatively appears unrealizable' (p. S272).

2.1.2 *Bounded Rationality*

Simon (1957) argued that the demands of reliable and unlimited information processing that normative models ask of people are too great, and any notion of rationality must be bounded by the psychological limits of human beings. Given limited computational and predictive abilities, 'actual human rationality-striving can at best be an extremely crude and simplified approximation to the kind of global rationality that is implied' (p. 243) – but still, 'people have reasons for what they do. They have motivations, and they use reason (well or badly) to respond to these motivations and reach their goals' (Simon, 1986, p. S209). He contended that rationality – in economics and psychology – is fundamentally about the proper allocation of *scarce resources*. Normative models consider factors of a choice to be made (such as monetary value and the probability of occurrence), but not factors of the *process* that is used to decide between choices: the costs of thought and attention that must be expended to examine, understand, and discriminate – the scarce resources of the mind (Simon, 1978).

When these factors are considered, the result is a *bounded rationality* that leads to a *satisficing* behaviour (March, 1978; Simon, 1978). That is, people make generalisations and take shortcuts in their decision-making processes to make the resulting behaviour as effective as possible within the constraints of time and mental power – even if it is known to be occasionally faulty. This can be observed in the tendency for people to occasionally follow normative axioms, but sometimes fall back to heuristics and generalisations (Brandstätter, Gigerenzer, & Hertwig, 2006; Shugan, 1980; Tversky & Kahneman, 1974).

2.1.3 Prescribing Behaviour

Descriptive models do not make judgements about what is or is not rational behaviour. Rather, they describe the relationships between the contributing factors of observed decision-making behaviour – without regard for whether these relationships are logically robust. However, there is still a desire to be able to critique choices, and descriptive models are open to criticism for a lack of explanatory power if they go no further than fitting the data (Simon, 1959).

Bell et al. (1988) argued that the divide between normative and descriptive models can be bridged by *prescriptive* models that are evaluated on their pragmatic value to 'help people make better decisions' (p. 18). For instance: normative models demand a transitivity over choices, and descriptive models capture the fact that people often hold intransitive preferences – but prescriptive models should expose the 'psychological nuances' (p. 14) that lead to the divergence between real and ideal behaviour so that it can be reduced or eliminated (or at least, better assessed by the decision maker).

Another approach is to view rationality and the application of normative axioms as a personal ideal, rather than a global rule. Allais (1979a) argued that rational behaviour only needs to (a) be a pursuit of ends that are not contradictory and (b) use means that are appropriate to those ends. This would still imply that choices are ordered, the dominance axiom applies, and objective probabilities should be considered – but it would not enforce a particular set of principles (see also Allais, 1979b). A person may subscribe to any model of behaviour that allows them to meet those goals, thereby turning a descriptive model into a personal normative model. That is, a person must have the desire to behave optimally, and normativity is a property of their acceptance of a set of axioms to achieve that. In the same sense that there are opposing and varied schools of economic thought, rational behaviour is a raison d'être to be adopted by an individual (Morgenstern, 1972).

Although there is little agreement about how to define rational behaviour or how to prescriptively correct people's choices, the debate and experimental evidence in pursuit of these questions has provided significant insight into

the complexity of human behaviour. In particular, it is difficult to criticise a decision as *foolish* or *irrational* solely on the basis of presupposed objective criteria or axioms, as there are many factors of an individual's disposition that can influence what is sensible for them to do.[13] March (1978) argued that these influences are 'not necessarily a fault in human choice to be corrected but often a form of intelligence to be refined by the technology of choice rather than ignored by it' (p. 598). For researchers designing experiments and interpreting subject behaviour, Mellers, Schwartz, and Cooke (1998) summarised three faulty assumptions that we should be cognisant of: (a) that there exists a single, correct choice; (b) that people should be internally coherent and logically consistent; and (c) that rationality is the same for subjects and experimenters.

13. Although people are often willing to criticise others and attribute responsibility for poor outcomes on the basis of supposed irrational behaviour (S. G. B. Johnson & Rips, 2015).

2.2 ECONOMIC UTILITY

Models in the economic literature are typically normative, but are based on an idea of *utility* that is inspired by empirical behaviour (Stigler, 1950a, 1950b). Utility models take an objective outcome (such as a payout) and use a *utility function* to transform it into a *utility score*, which better represents the value of that outcome to the person receiving it. A utility score is a quantitative value with constraints that allow it to be algebraically manipulated. When paired with a set of axioms in a *model of choice* (the conditions for when $A \succ B$), outcomes can be compared and decisions analysed. The focus of utility theories is not so much on the formulation of a particular utility function (as it varies between people and is largely unmeasurable), but on the shape and constraints of the curve produced by the function across the domain of outcomes – which predicts people's attitudes towards prospects.

Notation. A *prospect* is a set of outcomes $X = (x_1, ..., x_n)$ and associated probabilities $P = (p_1, ..., p_n)$ that a contract may yield (e.g. a flip of a coin, an insurance policy, or a financial derivative), where $\sum p_i = 1$ and an outcome x_i is obtained with corresponding probability p_i. For example, a fair coin that wins \$100 on heads and loses \$75 on tails is represented as: $X = (100, -75)$, $P = (.5, .5)$.

The *value* of a prospect is some measure of its worth. Measures of value depend on the nature of the outcomes, but in general the value of a prospect is what a person expects to receive from it, how much they would pay to receive its outcomes, or how much they would need to be paid to give them up. For example, is it worth gambling on the above coin flip, or is it worth offering the gamble to someone else?

The value of a prospect with nothing but a single, certain outcome (i.e. $X = (x)$, $P = (1)$) is simply its outcome (i.e. x). However, when there are multiple probabilistic outcomes – and especially when some of the outcomes

are negative – the question of assessing value is the subject of considerable debate (reviewed by Kreps, 1988; Starmer, 2000).

2.2.1 *Expected Value*

Originally discussed through a series of letters in the summer of 1654, Blaise Pascal (1623–62) and Pierre de Fermat (1601–65) proposed that the total value of a prospect should be divided by the odds with which its outcomes could be obtained (de Fermat, 1654/1894).[14] This was formalised by Christiaan Huygens (1629–95) as the *value of his expectation* (Huygens, 1657/1714):[15]

14. The question at the time concerned the inexplicable losses of the gambler Chevalier de Méré (1607–84) in his aleatory games.

15. For a historical review, see Daston (1988).

$$E(X,P) = \sum_{i=1}^{n} p_i x_i. \tag{2.1}$$

For example, the expected value of the fair coin described above is:

$$\begin{aligned} E(X,P) &= (.5 \cdot 100) + (.5 \cdot (-75)) \\ &= 12.50. \end{aligned}$$

Therefore it is worth playing the gamble for less than \$12.50 as it has a positive expected value (and it is a poor gamble to be offering someone else to play for more than \$12.50).[16]

16. This is irrespective of whether it is played once, or 100 times – even though repetition influences choice (Wedell & Böckenholt, 1990). The following chapter discusses this as a methodological concern.

Although this disinterested approach to assessing value is arguably fair, it is an extremely poor predictor of how people actually behave (reviewed below). Even worse, as a prescriptive model for rational behaviour, gambles can be easily constructed where it encourages ostensively irrational choices.

2.2.2 *Expected Utility*

In 1738, Daniel Bernoulli (1700–82) contemplated a problem given to him by his cousin Nicolas Bernoulli (1687–1759) with Gabriel Cramer (1704–52) that demonstrated a reductio ad absurdum problem with using the expected value analysis for decision making – since known as the *St. Petersburg paradox*:

> You are invited to play a game where a fair coin is repeatedly tossed. The pot starts at \$1 and doubles with each head that appears. Upon a tail, the game ends and you win whatever is in the pot. How much would you pay to play this game?

Under an expected value analysis:

$$\begin{aligned} X &= (1, 2, 4, \ldots), \\ P &= (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \ldots), \\ E(X,P) &= (\tfrac{1}{2} \cdot 1) + (\tfrac{1}{4} \cdot 2) + (\tfrac{1}{8} \cdot 4) + \ldots \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \ldots \end{aligned}$$

$$= \infty.$$

An infinite expected value suggests that it is reasonable to wager everything you own for an opportunity to play this game – but this is intuitively absurd!

Bernoulli (1738/1954) observed that (a) the value of money is not linear in its price and (b) the value of a prospect is influenced by the circumstances of the person considering it. For example, an increase of \$100 in weekly income has a higher utility to someone whose current weekly income is \$500 than to one whose current weekly income is \$5,000. He called this subjective value of a prospect its *utility* (cf. §2.5), and developed an *expected utility* measure of a prospect's value:

$$U(X,P) = \sum_{i=1}^{n} p_i u(W + x_i). \tag{2.2}$$

Where W is a person's current asset position and u is an increasing, concave utility function that reflects the diminishing marginal utility of wealth (Figure 2.1). The concavity of u describes three important properties:

- The diminishing marginal utility of wealth as amounts increase:[17]

$$0 < \frac{[u_{+2}] - [u_+]}{[+2x] - [+x]} < 1.$$

- A strong aversion to decreases in W:

$$|u_-| > |u_+| \text{ when } |x_-| = |x_+|.$$

- A preference for more certain outcomes over riskier ones – *risk aversion*:

$$p \cdot u(x) > \frac{p}{\lambda} u(\lambda x), \text{ for } \lambda > 1.$$

17. e.g. \$100 has more utility when added to a payout of \$50 than a payout of \$10,000.

Bernoulli used his expected utility theory to explain the existence of insurance policies and answer the St. Petersburg paradox: if the utility of the rapidly increasing payouts in X diminishes with a concave function,[18] the expected utility of the gamble $U(X,P)$ converges and is fairly small.[19]

18. Such as $u(x) = \ln(x)$.

19. If W is \$100,000, its utility would be \$11.51.

Coombs and Komorita (1958) developed experimental support for utility as a measure of value by asking subjects to choose between pairs of fair coin-flip gambles (X or Y). Each gamble had a winning outcome (x_1 or y_1) and a losing outcome (x_2 or y_2) such that:

$$u(y_1) > u(x_1) \geq u(x_2) > u(y_2).$$

Therefore, $X \succ Y$ if and only if:

$$u(x_2) - u(y_2) > u(y_1) - u(x_1).$$

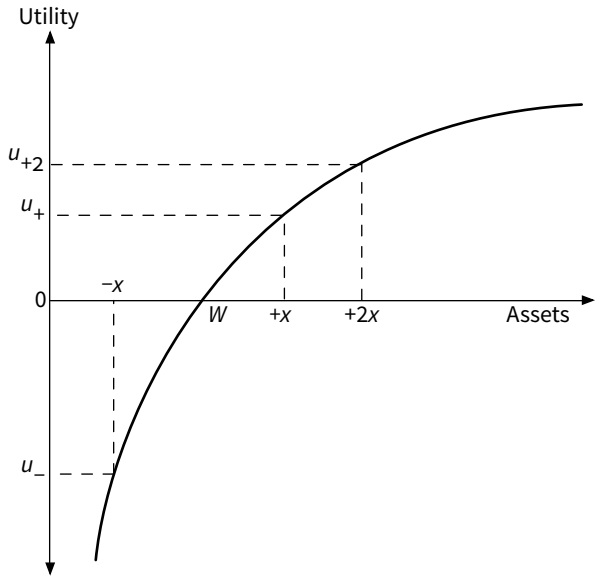


FIGURE 2.1 Bernoulli's (1738/1954) utility function u (Equation 2.2). The function intersects the abscissa at a person's total current assets W from which the utility of gains (e.g. $+x \rightarrow u_+$ and $+2x \rightarrow u_{+2}$) and losses (e.g. $-x \rightarrow u_-$) are measured.

For example, between either $X = (\$5, -\$5)$ or $Y = (\$6, -\$6)$, $X > Y$ if and only if the loss of utility in losing a dollar from -$5 to -$6 is greater than the gain of winning a dollar from $5 to $6 (and vice versa; Siegel, 1956). The influence of risk in the decision can be examined with a pair like $X = (\$0, \$0)$ and $Y = (\$6, -\$6)$ – that is, whether or not to gamble at all. Results generally supported a concave utility function for the ordering of preferences (although among only three subjects), but with mixed results for riskless gambles. Hurst and Siegel (1956) found similar results when using this procedure to test the expected value versus the expected utility of cigarettes for prisoners. For simple gambles, a utility analysis with an independence axiom[20] appears to model choices with fairly consistent utility functions between subjects (e.g. Davidson, Suppes, & Siegel, 1957; Tversky, 1967a, 1967b).

20. Usually *additivity*: if $u(a) > u(b)$ and $u(c) > u(d)$, then $u(a) + u(c) > u(b) + u(d)$.

However, a strictly concave utility function fails to describe some known behaviours that are difficult to observe in a laboratory – for example, the tendency for poor people to purchase insurance policies but also buy lottery tickets (a contradiction of risk aversion and risk seeking, respectively). Friedman and Savage (1948) proposed a modification to the utility function's shape: convex at high wealth levels (along the abscissa of Figure 2.1) but concave at low wealth levels – dividing a risk averse area (concave) from a risk seeking area (convex). Markowitz (1952) argued that a third inflection point was necessary

to capture behaviour close to a person's current level of wealth, with the utility curve alternating between concavity and convexity as wealth increases.

Despite a lull of 200 years, the ideas of Bernoulli's expected utility theory became the basis for normative models in modern economics (Arrow, 1958; Schoemaker, 1982). This was led by game theory (von Neumann & Morgenstern, 1947), which derived expected utility from a set of axioms that modelled choices when U is maximised (reviewed by Fishburn, 1979). That is, for any A and $B \in \mathcal{S}$, $U(A) > U(B)$ implies $A \succ B$, and preferences subsequently follow axioms like those described in Section 2.1.1.

However, as a descriptive model of behaviour a large set of experimental data has found that many of the decisions people make imply contradictions – particularly when varying the probabilities of outcomes (reviewed by Luce & Suppes, 1965; Machina, 1982, 1987; Tversky & Kahneman, 1986, and the axiom violations in §2.1). For example, if asked to choose between:

$3,000 with certainty or an 80% chance at $4,000,

most people prefer the $3,000 – implying $u(3000) > .8u(4000)$. But if the probabilities are divided by four:

a 25% chance at $3,000 or a 20% chance at $4,000,

most people prefer the $4,000 gamble – implying $.2u(4000) > .25u(3000)$. Rearranging the inequalities, the first implies:

$$u(3000)/u(4000) > .8,$$

while the second implies the opposite (Kahneman & Tversky, 1979):[21]

$$u(3000)/u(4000) < .8.$$

21. First presented by Allais (1953, 1979a) as a violation of independence (§2.1.1).

This behaviour is known as the *common ratio effect* (Cubitt, Starmer, & Sugden, 1998; Starmer & Sugden, 1989): as probabilities decrease, preferences for small (but more certain) outcomes reverse and begin to favour large (but less certain) outcomes.

Along similar lines, the most influential (and controversial; Allais, 1979b) counterexample to utility theories is the *Allais paradox* (Allais, 1953, 1979a):

First choose which of these two gambles you would rather play:

I-A		I-B	
Probability	Winnings	Probability	Winnings
100%	$1,000,000	89%	$1,000,000
		10%	$5,000,000
		1%	$0

And then which of these two gambles you would rather play:

II-A		II-B	
Probability	Winnings	Probability	Winnings
11%	$1,000,000	10%	$5,000,000
89%	$0	90%	$0

Most people select gambles I-A and II-B: implying opposite utility inequalities (as above). The only stated difference between the games in I and those in II is the removal of an 89% chance of $1,000,000 from both sides. Thus, expected utility theories analyse both pairs of gambles identically. However, empirical behaviour exhibits the *common consequence effect* (Wu & Gonzalez, 1998): when a common outcome is either added or removed from a pair of prospects, preferences reverse (also a violation of independence – §2.1.1).[22]

22. Another analysis of this behaviour is that people prefer the certainty of the large sum in I-A, and view the 1% chance of nothing in I-B as too risky; but somehow find the extra 1% chance of nothing in II-B acceptable for the chance at a larger sum.

2.2.3 *Subjective Expected Utility*

Although the value of outcomes (i.e. X alone) appears to follow simple utility curves, violations of expected utility models consistently appear under manipulations of risk – that is, when the probabilities of obtaining the outcomes are manipulated. This behaviour could not be accounted for solely by adjusting the shape of u (Rabin, 2000a), and so attention turned to the probability term p_i (in Equation 2.2): if the value of an outcome x_i can be non-linearly transformed by a utility function u, then perhaps a similar treatment should be given to the probability p_i.

There is substantial psychological evidence that, unlike what Equation 2.2 describes, the probability of an outcome does not linearly influence the value of a prospect. In another example of the common ratio effect Kahneman and Tversky (1979) found that people prefer a 0.1% chance at $6,000 over a 0.2% chance at $3,000 – but when the probabilities are multiplied by 450 (a 45% chance at $6,000 or a 90% chance at $3,000), the preference reverses (see also MacCrimmon & Larsson, 1979).

To examine the shape of this influence, Preston and Baratta (1948) had subjects bid (using a pre-allocated sum of points that were exchanged for a tangible reward at the end of the experiment) for cards that offered them varied chances of winning more points (e.g. a 5% chance of 100 points). They found subjects consistently overbid for cards with low probability scores, and underbid for those with high probability scores (relative to the expected value of the cards) – there was no relationship between the size of the outcome and their relative bids.

Through a series of experiments, Edwards (1953; 1954a; 1954b) found that subjects did not necessarily prefer outcomes with a higher expected value to

those with lower, and they did not seek to maximise winnings or minimise losses in their gambles – rather there was a systematic preference for gambles that offered certain probabilities, and to either take big risks or avoid them. This culminated in a *subjective expected utility* (SEU) theory of a prospect's value (Edwards, 1955):[23]

$$SEU(X,P) = \sum_{i=1}^{n} w(p_i)u(x_i). \tag{2.3}$$

Where w is a non-linear transformation of an objective probability p_i into a *subjective probability* (in $[0,1]$).[24] The shape of this function was not tightly defined but has been regularly found to resemble Figure 2.4 – with an overweighting of low probabilities and an underweighting of high probabilities.

The SEU model was given extensive treatment in the psychology literature (reviewed by Becker & McClintock, 1967; Edwards, 1961, 1962; Luce & Suppes, 1965), but largely ignored in economics[25] as it is repugnant with the dominance axiom and jeopardises the concavity of u (Machina, 1982; cf. rejoinder in Tversky & Kahneman, 1986). Introducing a new degree of freedom also increases the difficulty of testing the theory as it requires separating the probability and utility components of choice,[26] and necessarily guarantees a better fit the with the data if the shape of w is not constrained (Edwards, 1961). However, subsequent work has demonstrated its descriptive success through methodologies that isolate the probability component of a choice and discover constraints for w (e.g. Coombs et al., 1967; Tversky, 1967a, 1967b; Tversky & Kahneman, 1992; Wallsten, 1970; Wu & Gonzalez, 1996).

25. Until *prospect theory* (§2.2.4).

26. To the extent that they are actually separable (Irwin, 1953; Slovic, 1966).

Prelec (1998) reviewed four properties that have come to characterise w:

- *Regressive*. At low levels of p, $w(p) > p$; but as p increases, $w(p) < p$ after some point, which begets *subcertainty*:

$$w(p) + w(1-p) \leq 1. \tag{2.4}$$

- *Asymmetric*. It is not mirrored along the diagonal axis, and has a fixed point at around $p = ⅓$.
- *Inverse S-shaped*. It is initially concave, and then convex.
- *Reflective*. It is invariant on the value (i.e. x_i) being weighted.

Critically, $w(0) = 0$, $w(1) = 1$, and must obey *subproportionality*: for any $\lambda \in (0,1)$ and $1 \geq p > q > 0$,

$$\frac{w(p)}{w(q)} > \frac{w(\lambda p)}{w(\lambda q)}. \tag{2.5}$$

23 In the economics literature, 'subjective expected utility' or 'subjective probability' usually refers to a component of a normative model by Savage (1972), which concerns the assessment of probability when there is uncertainty about the objective probabilities (reviewed in §2.2.5). For economists, if an objective probability is known, the only defensible action is to use it directly.

24 The subjective probability function w can also be applied to an expected value model: $SEV(X,P) = \sum w(p_i)x_i$. However, this model shares the same problems that expected value does and was not given serious consideration (Coombs, Bezembinder, & Goode, 1967; Edwards, 1961).

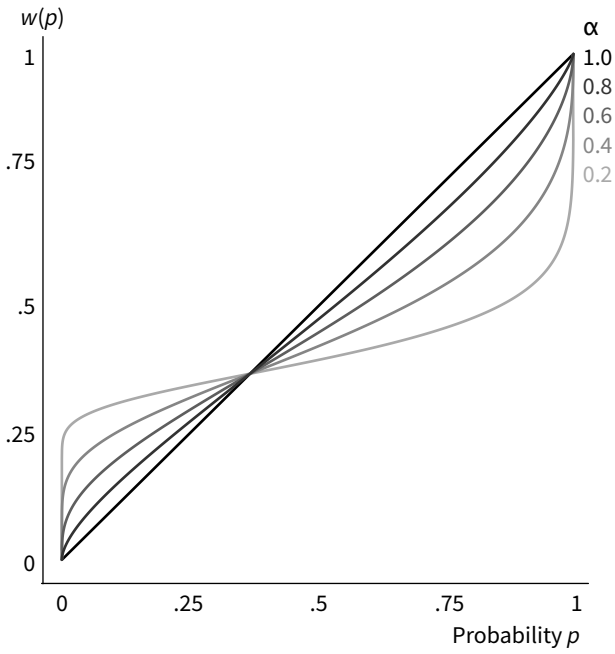


FIGURE 2.2 The weighting function *w* of Prelec (1998) – Equation 2.6 – at several levels of α, with β = 1.

That is, for any ratio of probabilities p/q, the corresponding $w(p)/w(q)$ is closer to 1 for small probabilities than for large – making them less distinguishable (Fehr-Duda & Epper, 2012; Kahneman & Tversky, 1979). Prelec also developed an axiomatic construction of w and proposed a formulation that fitted much of the empirical evidence (Gonzalez & Wu, 1999; Prelec, 2000):

$$w(p) = \exp(-\beta[-\ln(p)]^{\alpha}). \tag{2.6}$$

Where $\alpha \in [0,1]$ is an index of the subproportionality[27] (illustrated in Figure 2.2) and $\beta > 0$ independently increases the convexity by shifting the inflection point vertically. Fehr-Duda and Epper (2012) reviewed the substantial theoretical and empirical evidence for Prelec's (and other) proposed formulations for the shape of w. However, while the general shape of w is robust, there is evidence that factors other than p influence its parameters (reviewed later in §2.6).

27. The 'Allais Paradox Index' (Prelec, 2000, p. 78).

The Allais paradox (and other effects) can be understood with such a function w that is first concave and then convex. The *certainty effect* that subproportionality implies also disproportionately devalues certain outcomes that become risky (cf. theories of disappointment aversion in §2.6.3).

2.2.4 Prospect Theory

Kahneman and Tversky (1979) developed an SEU account of the seemingly contradictory behaviours observed under expected utility models in *prospect theory*, which models the *psychological value* of outcomes. Prospect theory combined three major advances over expected utility theories:

- Outcomes are measured as *gains* and *losses* relative to a neutral *reference point*. That is, prospect theory considers the changes that are brought about by an outcome instead of the final state that results. This reference point may be the current asset position of the person considering the outcome (as in expected utility theory), but crucially, it can also be altered by their expectations or the formulation of the prospect (reviewed in §2.3.1).
- The shape of the utility function (known as the *value function*) in the domain of losses is different from that in the domain of gains: it is concave above the reference point, convex below it, and steeper for losses than it is for gains.
- Probabilities are replaced with *decision weights* that incorporate information about the subjective desirability of a prospect, and perceptions of its likelihood.

Under prospect theory, the value of a prospect is:

$$V(X,P) = \sum_{i=1}^{n} \pi(p_i)v(x_i). \tag{2.7}$$

Where v is a value function for which $v(0) = 0$, and π is a decision weight function for which $\pi(0) = 0$ and $\pi(1) = 1$. The general shape of the value function v is shown in Figure 2.3 and must be concave for gains, convex for losses, and steeper for losses than for gains (akin to Friedman & Savage, 1948; Markowitz, 1952). The function v also transforms its argument to be measured as a gain or loss relative to some reference point. The general shape of the decision weight function π is shown in Figure 2.4, and must satisfy subcertainty and subproportionality (§2.2.3).[28]

The shape of v encapsulates three common empirical observations in how people evaluate the outcomes of prospects (with examples from Kahneman & Tversky, 1979, and Kahneman & Tversky, 1984):

- *Risk aversion* when all outcomes of a prospect are in the domain of gains: people prefer a certain \$3,000 to an 80% chance at \$4,000.

28 A revised *cumulative prospect theory* (Tversky & Kahneman, 1992; Wakker & Tversky, 1993) changed the weighting function to transform the cumulative probability distribution function, rather than individual probabilities – which can lead to violations of dominance (Fishburn, 1978; Starmer, 2000). This update brings it into the class of *rank-dependent expected-utility* theories, which make SEU theories more acceptable to economists (Diecidue & Wakker, 2001; Quiggin, 1982). Cumulative prospect theory also split π into a function for losses ($v(x) < 0$) and a function for gains ($v(x) \geq 0$). These changes are immaterial to the discussion in this thesis.

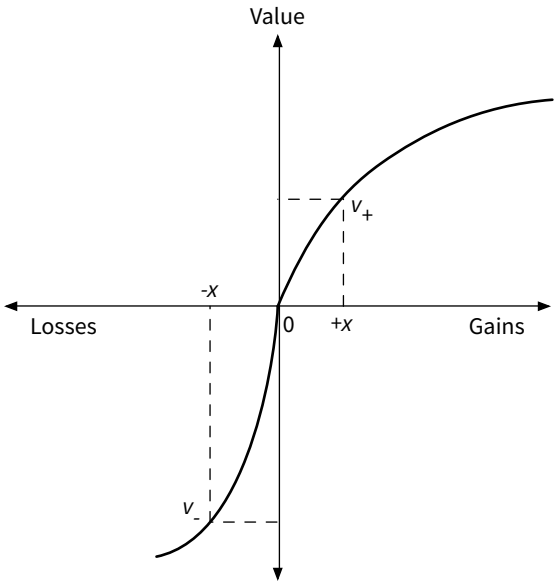


FIGURE 2.3 The value function of prospect theory (Kahneman & Tversky, 1979): the psychological value of an outcome's losses or gains: concave for gains, convex for losses, and steeper for losses (e.g. $-x \to v_-$) than for gains (e.g. $+x \to v_+$).

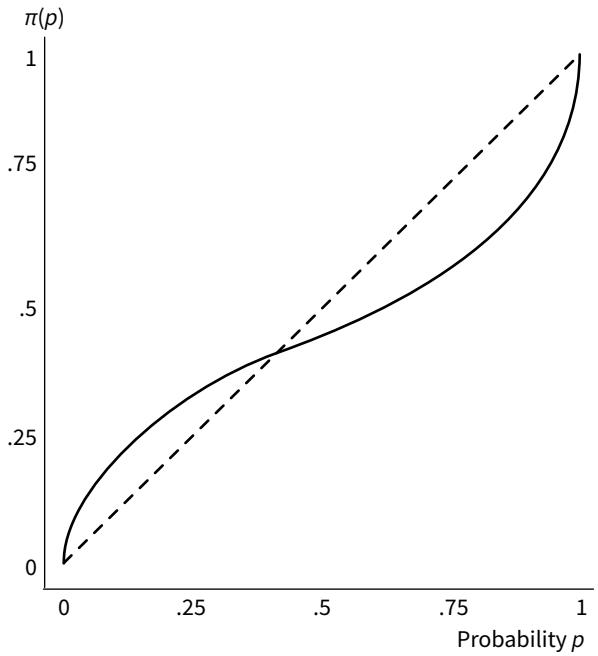


FIGURE 2.4 An example of prospect theory's decision weight function (cf. Figure 2.2). The dashed line shows an objective $\pi(p) = p$ function, whereas the solid line shows an overweighting of low probabilities and underweighting of high probabilities – similar to the functions experimentally found by Tversky and Kahneman (1992) and other investigations of SEU subjective probability functions w (Fehr-Duda & Epper, 2012).

- *Risk seeking* when all outcomes of a prospect are in the domain of losses: people prefer to gamble on an 80% chance at losing $4,000 than to accept a loss of $3,000 with certainty.
- *Loss aversion* when the outcomes of a prospect are mixed: people reject a gamble with a 50% chance of winning less than $30 and a 50% chance of losing $10 (i.e. a stronger preference for avoiding losses than for acquiring gains: $v(x) < -v(-x)$).

The sentiment is that 'losses loom larger than gains' (Kahneman & Tversky, 1979, p. 279).[29]

Loss aversion is a key contribution of prospect theory, and is part of a broader *negativity bias* observed in psychological research (reviewed in §2.6.1). In support of it, Tversky and Kahneman (1991) reviewed several experiments where subjects made choices between alternatives that combined positive and negative elements (e.g. a job that was socially isolated, but had a short commute – or vice versa). Results showed that the negative outcomes dominated preferences, and in general the magnitude of a loss was weighted approximately two times more than a comparable gain – the *coefficient of loss aversion* (see also Köbberling & Wakker, 2005).

An *editing phase* occurs prior to the evaluation of V that captures the perceptual process of interpreting a prospect, and prepares it for evaluation: (a) a reference point is established to define the domain of outcomes for gains and those for losses, (b) identical outcomes are combined or cancelled, (c) riskless components are separated from risky ones, (d) outcomes and probabilities are simplified, and (e) dominated prospects are removed from consideration.

Prospect theory makes no normative claims or comment on the rationality of the choices it describes, but exists to capture the influences of psychological factors on choice in an economic context – which may be overcome through training or experience (Barberis, Huang, & Santos, 2001; List, 2003, 2004). It has been remarkably successful as a descriptive model to explain why many of the choices people make conflict with prior economic utility models (reviewed by Barberis, 2013; Camerer, 1998; Kahneman & Tversky, 2000; Shafir & Tversky, 1995; Starmer, 1999), and has been applied to behaviour across a number of fields (e.g. Camerer, 2004; Chou, Chou, & Ko, 2009; Mercer, 2005; M. Weber & Camerer, 1998; Wu, Heath, & Larrick, 2008). However, prospect theory also has its own paradoxes – experimental data that do not conform to its hypotheses (reviewed by Birnbaum, 2008; Gonzalez & Wu, 2003); and there are challenges in applying it to new domains (Barberis, 2013) – reviewed for interaction in the following chapter.

29 Cf. 'We suffer more . . . when we fall from a better to a worse situation, than we ever enjoy when we rise from a worse to a better' (Smith, 1759/1976, p. 213).

2.2.5 *Uncertainty and Risk*

In the choice problems discussed so far, the exact probability of an outcome has always been given with the definition of a prospect – and many experimental tests of economic utility theories make objective probabilities explicit to subjects. However, rarely is probability so explicit: most decisions outside of the laboratory feature not only *risk* in obtaining their probabilistic outcomes, but also *uncertainty* in what those probabilities are. Einhorn and Hogarth (1985) distinguished between three types of this precariousness:

- *Risk*: the exact probability distribution is known (e.g. $P = (.3, .5, .2)$).
- *Ignorance*: there is no information about the probability distribution for the outcomes (e.g. $P = (?, ?, ?)$).
- *Ambiguity*: there is some information about the probability distribution for the outcomes (e.g. $P = (.3, ?, ?)$ or $P = (.3, \approx.5, \approx.2)$).

For example, the Ellsberg paradox in Section 2.1.1 asks people to make a bet about an urn that contains balls of three different colours for which

$$P(\text{Red}, \text{Black}, \text{Yellow}) = (1/3, ?, ?).$$

That is, there is ambiguity throughout the bets: risk about the drawing of a red ball, and ignorance about the drawing of a black or yellow ball. The ignorance could be resolved into ambiguity if a sample of several balls was drawn from the urn before the bet is made: for example, if yellow, yellow, black, yellow, and red balls were drawn (without replacement), then $P = (1/3, \approx 1/6, \approx 1/2)$.[30]

30. Ambiguity remains in the *estimated* black and yellow proportions.

In economics, exposition of the division between risk and uncertainty is attributed to Knight (1921), and its influence on preferences to Ramsey (1926/1931), de Finetti (1937/1980), and Savage (1962, 1972). They use the term *subjective probability*[31] to refer to a person's opinion on the probability of an uncertain outcome. As with models of choice, their research focussed on how to measure uncertainty and resolve it towards risk as relevant information is obtained (reviewed by Pratt, Raiffa, & Schlaifer, 1964). For example, consider a one-urn version of the Ellsberg paradox (1961; also given by Keynes, 1921):

31. Distinct from that reviewed in §2.2.3.

> You are confronted with two urns:
>
> **URN-I**: contains a total of 100 red and black balls in some unknown proportion.
>
> **URN-II**: contains exactly 50 red and 50 black balls.
>
> For each of the following gambles, either choose the bet you would prefer to make (i.e. $A \succ B$) or state that you are indifferent between them (i.e. $A \sim B$). Bets are labelled with a colour and an urn number – for example, 'Red-I' is a bet (of say, \$100) on a red ball being drawn from URN-I.
>
> (a) Red-I or Black-I, (b) Red-II or Black-II,

(c) Red-I or Red-II, and (d) Black-I or Black-II.

From the responses subjects give to questions like these, a probability distribution for the proportion of red or black balls in URN-I can be estimated. As more information is obtained (such as from watching other people play, drawing samples from the urns, or consulting an oracle), these distributions should be updated following Bayesian principles (Savage, 1962):

$$P(\alpha|x) \propto P(x|\alpha)P(\alpha). \tag{2.8}$$

That is, the probability of an unknown α given some information x is proportional to the product of the probability of observing x supposing α is true, and the initial probability of α.

However, most people are indifferent to the first two bets and have a definite preference for URN-II in the others. This implies under Savage's (1972) axioms that the probability of Red-II is greater than Red-I, and at the same time not-Red-II (i.e. Black-II) is greater than not-Red-I (i.e. Black-I) – a contradiction (Roberts, 1963):

(a) $P(\text{Red-I}) = P(\text{Black-I}) = 0.5$, (b) $P(\text{Red-II}) = P(\text{Black-II}) = 0.5$,

(c) $P(\text{Red-I}) > P(\text{Red-II})$, and (d) $P(\neg\text{Red-I}) > P(\neg\text{Red-II})$.

That is, people's preferences for prospects with uncertainty do not behave as if they are assigning and manipulating numeric probability scores. Ellsberg (1961) and Roberts (1963) argued that the uncertainty in the contents of URN-I is itself a factor in the choice – which has subsequently been refined into a general *uncertainty aversion* (Einhorn & Hogarth, 1986; Fox & Tversky, 1995; Schmeidler, 1989).[32]

Uncertainty aversion is distinct from the *risk aversion* modelled by utility theories (§2.2.2 and reviewed by Camerer & Weber, 1992). Risk aversion describes people's preferences for low risk outcomes (i.e. high values of p_i) over high risk outcomes (i.e. low values of p_i), whereas uncertainty aversion describes people's preferences for more certain outcomes (i.e. a narrower distribution of possible p_i values) over less certain outcomes (i.e. a wider distribution of possible p_i values). Risk aversion is also matched by a corresponding *risk seeking* behaviour when outcomes are uniformly in the domain of losses (§2.2.4), but there is no matching 'uncertainty seeking' behaviour.

Tversky and Fox (1995) developed a modification of prospect theory's decision weight function π to capture the effects of uncertainty. Unlike expected utility theories, which expect probabilities to be additive (i.e. $P(A \vee B) =$

32 Subsequent work has also found that the application of Bayesian logic (Equation 2.8) to updating probability estimates is likewise faulty (Grether, 1980): people underweight base rates (Bar-Hillel, 1980), place too much belief in small samples (Tversky & Kahneman, 1971), and display a bias towards representativeness over likelihood (Kahneman & Tversky, 1973; Tversky & Kahneman, 1983) – further reviewed in §2.6.

$P(A)+P(B)$ for any disjoint A and $B \in \mathcal{S}$), Tversky and Fox permitted π to have *subadditivity*:

$$\pi(A) \geq \pi(A \vee B) - \pi(B), \text{ and} \tag{2.9}$$

$$\pi(\mathcal{S}) - \pi(\mathcal{S} - A) \geq \pi(A \vee B) - \pi(B). \tag{2.10}$$

Where $\pi(\mathcal{S}) = 1$. This captured their empirical findings that 'an event has more impact when it turns impossibility into possibility [Equation 2.9], or possibility into certainty [Equation 2.10], than when it merely makes a possibility more likely' (p. 281). A measure for the degree of subadditivity – the influence of uncertainty when compared against a certainty equivalent[33] – can be found by rearranging the inequalities:

33. The minimum value a person would prefer with certainty over taking a gamble.

$$D(A,B) = \pi(A) + \pi(B) - \pi(A \vee B), \text{ and} \tag{2.11}$$

$$D'(A,B) = \pi(\mathcal{S}) - \pi(\mathcal{S} - A) - \pi(A \vee B) + \pi(B). \tag{2.12}$$

Similarly, if risk aversion is solely a property of the shape of u (as assumed by expected utility models), then it too can be measured by examining the function's curvature (Arrow, 1970; Pratt, 1964):

$$r(x) = -\frac{u''(x)}{u'(x)}. \tag{2.13}$$

Where, u' and u'' are the first and second derivatives of u, respectively. That is, for some asset position x, r measures the local risk aversion around that area of the utility function u – essentially, a measure of the concavity at x.[34] Under an expected utility model, as total assets W increase (Equation 2.2), an individual is more willing to accept risk if r is decreasing. However, as highlighted in SEU and prospect theories, risk attitudes are expressed with a combination of the utility or value function, and a non-linear probability or decision weighting function (Fehr-Duda & Epper, 2012; Kahneman & Tversky, 1979; Rabin, 2000b). In these cases, r is not a complete measure of risk aversion, but it is still a useful measure for the other properties of a utility curve (diminishing sensitivity and loss aversion).

34. When multiplied by x (i.e. $r'(x) = x \cdot r(x)$), it is a measure of aversion in proportion to total assets x.

2.3 PREFERENCES

Prospect theory expands the scope of economic models of utility with the influences of a decision-maker's beliefs, desires, and expectations – that is, their *preferences*. These psychological factors drive the non-linearity of v and π. However, the original formulation of prospect theory was vague about the source of these subjective qualities – for example, how a reference point is established (Barberis, 2013). Subsequent research has refined these properties through extensive empirical research (reviewed in Lichtenstein & Slovic, 2006): resulting in more complete models of *reference-dependent preferences*.

2.3.1 Reference-Dependence

Prospect theory's *reference point* is based on a hypothesis that people respond to the changes that an outcome delivers from an adapted neutral state (Tversky & Bar-Hillel, 1983; Tversky & Kahneman, 1991), rather than the resultant state (assumed by expected utility). In simple monetary prospects this reference point may be the current wealth of the person considering the prospect – but importantly, even for simple prospects, the reference point can be shifted by a person's expectations of what they want to receive, or by unadapted changes (such as a recent windfall or loss).

Experiments with riskless prospects (i.e. with a single outcome $X = (x)$ and $P = (1)$) have supported the hypothesised shape of prospect theory's value function v and its reference-dependence. For example, Kahneman, Knetsch, and Thaler (1990) gave a group of subjects a mug that retailed for \$5 and asked them how much they would sell it for (a *willingness-to-accept* measure), while another group of subjects were asked how much they would buy it for (a *willingness-to-pay* measure). They found subjects were willing to accept around \$7, but willing to pay only \$3.50. This reluctance to trade[35] – and in particular, the stronger aversion to selling – indicates that possession of the mug established a reference point for subjects that weighted their evaluation of losing it in excess of its fair value (see also Bateman, Munro, Rhodes, Starmer, & Sugden, 1997; Tversky & Kahneman, 1991).

35. Known generally as the *endowment effect* (§2.6.2).

Outside the laboratory, evidence for reference-dependent loss-averse behaviour has been found in across a range of products. For example, Putler (1992) found changes in egg sales were 2.4 times greater when egg prices increased (a relative loss) than when they decreased (a relative gain). Similar effects have been found for orange juice sales (Hardie, Johnson, & Fader, 1993), Boston apartment listing prices (Genesove & Mayer, 2001), the evaluation of job candidates (Highhouse & Johnson, 1996), and New York City taxicab driver wage elasticities (Camerer, Babcock, Loewenstein, & Thaler, 1997).

Although it is often convenient to assume that the reference point for choices is the status quo state (whatever a person currently has or recently experienced), there is a large body of experimental work showing that reference points can be manipulated without any material change to outcomes or the status quo – these are reviewed later for the endowment effect (§2.6.2), framing effects (§2.6.3), and sunk costs (§2.6.5). The method for establishing a reference point experimentally or identifying one ethnographically is also dependent on the nature of the outcomes being evaluated (money, mugs, eggs, etc.), and demands experimental controls that validate it – discussed in the following chapter.

2.3.2 *Expectations*

A key influence on a person's reference point is their *expectation* of a prospect's outcome: their beliefs about an outcome prior to its occurrence. As with empirical reference points (e.g. current wealth or possessions), expectations can provide a benchmark against which an outcome can be measured: is an outcome better or worse than expected? Expectations may be developed through a careful analysis of knowledge and experience, or wishful thinking about ideal outcomes; they may be swayed by anxiety or suspense in anticipation of an event (Caplin & Leahy, 2001); and they may be held with strong conviction or with relative disinterest.

Kahneman and Tversky (1982) distinguished between three main types of 'perceptual expectancies' (p. 144):

- *Active* expectancies occupy conscious thought and draw on attention and consideration.
- *Passive-temporary* expectancies are automatic and effortless, and are driven by the effects of context and priming.
- *Passive-permanent* expectancies are similarly automatic, but are built on long-lasting experiences and models of reality.

These expectancies determine not only what an outcome is anticipated to be, but also influence how it is perceived and the strength with which it is held. When there is limited knowledge about the cause–effect relationship of an action and its outcome, there is *uncertainty* about the belief (§2.2.5).

It is desirable that expectations be based on robust forecasts made from reliable data (Muth, 1961). For example, businesses can forecast future sales from prior data and economic models, and sports betting odds are usually calculated from past performance. However, such forecasts for economic data – such as sales, price inflation, and wages – are frequently found to contain errors (Lovell, 1986; Williams, 1987). These errors are not random (i.e. their mean is not zero), but contain systematic biases. The psychological literature calls this the 'illusion of validity' (reviewed by Einhorn & Hogarth, 1978, p. 395), wherein expert clinicians are found to frequently make confident predictions that contain large systematic errors. For example, the judgement of university student performance by academic admissions faculty is often worse than predictions from a simple linear model of the relevant variables (reviewed by Dawes, 1979; Meehl, 1954; Simon, 1959). If these forecasts and judgements are used as the basis for reference points, then choices about alternatives will be similarly biased.

Heath, Larrick, and Wu (1999) studied the effect of goals on the evaluation of outcomes in the context of prospect theory's predictions about loss

aversion and diminishing sensitivity – hypothesising that goals establish active expectations that set a reference point for v (formalised into a model by Wu et al., 2008). For example, if a person establishes a goal of scoring 90% on an exam but only achieves 87%, they would feel negatively about their performance (as a loss) – even if their typical level of performance was only 80%. In a series of such scenarios described to subjects, Heath et al. found goals had a systematic effect on the evaluation of outcomes that was consistent with prospect theory's model. They argued that prospect theory's value function could be used to explain the effects observed in the goal literature of effort, persistence, and performance (reviewed by Krantz & Kunreuther, 2007; Locke & Latham, 2002).

The influence of expectations on perceived value also received attention in the marketing literature, where researchers considered how expectations influence consumer purchasing decisions and satisfaction (reviewed by Oliver & Winer, 1987).[36] This literature uses models of *disconfirmation* (e.g. Oliver, 1980): a consumer's expectation of a product or experience can either be (a) *confirmed* if it is met by the outcome, (b) *positively disconfirmed* if the outcome is better than expected, or (c) *negatively disconfirmed* if the outcome is worse than expected. Yi (1990) reviewed these models, the theories of resolving disconfirmation,[37] and studies of their effects. In general, negative disconfirmation has a stronger impact on overall experience than either of the others (see also Rozin & Royzman, 2001). However, the proposed resolution theories predict conflicting responses, and observing the subtle effects and relationships that may distinguish them has been difficult (Yi, 1990). Recent attention in the marketing literature has turned towards the behavioural economics models discussed in this chapter (Goldstein, Johnson, & Sharpe, 2006; Ho, 2006).

36. Usually without prospect theory (cf. Thaler, 1980).

2.3.3 *A Reference-Dependent Preferences Model*

Kőszegi and Rabin (2006) developed a general model of reference-dependent preferences based on a person's probabilistic beliefs about outcomes – combining much of the prior theoretical and empirical work into a unified model. In their model, the utility of a prospect is analysed using a *consumption bundle* $\mathbf{c}$ of what the prospect yields, and an endogenous *reference bundle* $\mathbf{r}$ of what a person expects it to yield. These bundles are K-dimensional vectors (in $\mathbb{R}^K$) that contain a collection of objective outcomes. This model, with an account

37 For example, in *contrast theory* (Spector, 1956) the surprise of disconfirmation causes perceptions to be exaggerated (an outcome that is positively disconfirmed is perceived as more positive than is objectively the case, and vice versa), while in *dissonance theory* (Festinger, 1957) disconfirmation is psychologically uncomfortable and is resolved by distorting the perception of performance to bring it closer to what was expected.

of the consumption and reference bundles that express interactive prospects, is further detailed in the following chapter.

The consumption bundle **c** is derived from the specification of a prospect's outcomes, and the reference bundle **r** is based on a person's 'rational expectations held in the recent past about outcomes' (p. 1133). By using bundles of outcomes, multiple factors of a decision can be given consideration. For example, if a consumer goes shopping for shoes, they may assess outcomes (e.g. potential product purchases) along at least two dimensions:

1. Their wealth adjusted by the price of the product (in $\mathbb{R}$).
2. Whether or not they will posess new shoes (in $\{0,1\}$).

Their reference bundle **r** will contain:

1. Their wealth adjusted by the amount they expect to spend (i.e. their budget).
2. 1 (indicating a purchase of some shoes).

And the consumption bundles they encounter during their excursion **c** (i.e. various goods being sold) will contain:

1. Their wealth adjusted by the amount they need to spend.
2. Whether or not they will obtain new shoes from the purchase.

Using consumption bundles allowed Kőszegi and Rabin to separately evaluate the influence of the multiple, independent dimensions of an outcome that bear upon a decision. For example, when the consumer visits a store with their expectation of purchasing shoes, the expectation of making a purchase sets a belief about future consumption that may make them less sensitive to the prices they encounter. To avoid experiencing a sensation of loss if no purchase is made (an unrealised expectation of consumption) they will prefer to buy at higher prices than if they had no prior expectation of making a purchase.

Using these consumption bundles, the utility of a prospect is:

$$u(\mathbf{c}|\mathbf{r}) = m(\mathbf{c}) + n(\mathbf{c}|\mathbf{r}), \tag{2.14}$$

$$m(\mathbf{c}) = \sum_{k=1}^{K} m_k(c_k), \tag{2.15}$$

$$n(\mathbf{c}|\mathbf{r}) = \sum_{k=1}^{K} \mu\big(m_k(c_k) - m_k(r_k)\big). \tag{2.16}$$

Where—

- $u(\mathbf{c}|\mathbf{r})$ is the *overall utility* of a consumption bundle **c**, given reference expectations **r**.[38]
- m is the *consumption utility* of obtaining a particular bundle. This is a purely objective utility measure associated with those outcomes (analogous to u in Equation 2.2). Consumption utility is evaluated separately for each dimension of a bundle, and summed across them.

38. This is distinct from the u in Equation 2.2.

- n is the *gain–loss utility* function that weights gains and losses according to the difference between consumption utility for an actual outcome $\mathbf{c}$ and the reference expectations $\mathbf{r}$.
- μ is the *universal gain–loss function* that satisfies the properties of prospect theory's value function (essentially v from Equation 2.7), and models the subjective aspects of the gain or loss in marginal utility.

In general, people prefer prospects that (a) have high consumption utility $m(\mathbf{c})$ and (b) are aligned with or exceed reference expectations $n(\mathbf{c}|\mathbf{r})$. Using the properties of prospect theory's value function in μ exaggerates the effect of failing to meet reference expectations (loss aversion), and dampens the effect of exceeding them (diminishing sensitivity).

When outcomes are drawn from a probability distribution F (i.e. there is risk), the utility is integrated across that distribution:

$$U(F|\mathbf{r}) = \int u(\mathbf{c}|\mathbf{r})\,\mathrm{d}F(\mathbf{c}). \tag{2.17}$$

And similarly if expectations are drawn from a probability distribution G (i.e. there is uncertainty):[39]

$$U(F|G) = \iint u(\mathbf{c}|\mathbf{r})\,\mathrm{d}G(\mathbf{r})\,\mathrm{d}F(\mathbf{c}). \tag{2.18}$$

39. Kőszegi and Rabin note that for simplicity these formulations are wilfully ignorant of the non-linearity of prospect theory's π.

This formulation captures several psychological properties of reference-dependent preferences in an axiomatic formulation:

- If an outcome $\mathbf{c}$ is fixed, lowering the reference point $\mathbf{r}$ increases satisfaction with the outcome.
- Preferences have a status-quo bias: if a reference point $\mathbf{r}$ is better than an outcome $\mathbf{c}$ along some dimensions, then for $\mathbf{c} > \mathbf{r}$, the bundle $\mathbf{c}$ must offer larger gains in the remaining dimensions.
- Small changes in utility share the properties of prospect theory's value function, but when there are dramatic differences between outcomes (such as life-and-death scenarios), loss aversion may not be observed.

This model is built on the experimental findings of prospect theory and reference-dependence, and has subsequently found support in modelling and explaining labour supply choices (e.g. Abeler, Falk, Goette, & Huffman, 2011; Crawford & Meng, 2011; Kőszegi & Rabin, 2006), consumer behaviour in the marketing literature (Ho, 2006), and other psychological effects (Barberis, 2013; Heidhues & Kőszegi, 2008; Herweg, Müller, & Weinschenk, 2010). The following chapter will develop this model for interactive contexts and discuss approaches for applying it in human–computer interaction research.

2.4 DECISION-MAKING PROCESSES

As with the introduction of uncertainty (§2.2.5), the models of choice discussed so far have assumed that all of a prospect's outcomes are defined: the

complete set of outcomes, their values, and all relevant parameters are known. For instance, although the Ellsberg paradoxes (§2.1 and §2.2.5) have uncertain outcomes, all possible choices and their resultant outcome states are given. However, as the decisions under consideration shift from monetary gambles in a laboratory to more conventional questions outside of one – where to go on holiday, which breakfast cereal to buy, which text editor to use, etc. – it usually becomes less clear to people what the consequences of each alternative are, or even what the complete set of alternatives is. Understanding how people approach these uncertainties crosses multiple areas of cognitive psychology – attention, memory, learning, etc. – and are brought together in *information processing theories* of decision making (reviewed by Oppenheimer & Kelso, 2015; E. U. Weber & Johnson, 2009). Although applying these theories is beyond the scope of this thesis, it is useful to briefly review them as context for economic theories of choice.

Identifying a set of alternative choices depends upon the complexity of the decision, the significance of its consequences, and the will of the individual. For example, selecting a poor breakfast cereal is far less disastrous than selecting a poor holiday location, and therefore demands less scrutiny. However, even important decisions cannot be given an unlimited amount of consideration – for example, when thinking about a holiday location, people do not give conscious consideration to every location on Earth. As the number of factors that could contribute to a decision is unbounded, people apply heuristics and simplification strategies to reduce the set of alternatives – for example, favouring familiar brands or recommendations from friends (reviewed by Gigerenzer & Gaissmaier, 2011). Strategies such as *information foraging* describe how people balance the resource costs of seeking new information with the opportunity costs of attending to other decisions (Pirolli & Card, 1999). Kahneman (2003) argued for an overarching dual-system approach: decision making is divided between a fast, parallel, and emotional intuition system for quick judgements of little consequence, and a slow, serial, and neutral reasoning system for complex judgements. The system that is engaged to solve a decision-making problem depends on a person's available resources, training, and psychological priming (Ferreira, Garcia-Marques, Sherman, & Sherman, 2006; Stanovich & West, 2000).

Oppenheimer and Kelso (2015) reviewed the major approaches to understanding this cognitive architecture of decision making. Unlike expected utility models (which assume preferences are stable) and heuristic models (which are only sporadically applicable), such theories approach decision making as a dynamic system: information is collected under the constraints of a person's memory and attention, and is stochastically sampled based on current goals and preferences. Importantly, there is also a temporal element that allows

feedback to adjust preferences over time (either from observing the consequences of a choice, or deliberating on a problem for longer).

A particularly compelling theory is *decision by sampling* (Stewart, Chater, & Brown, 2006), which explicitly rejects the idea that people calculate utility scores for outcomes. Instead, Stewart et al. argued that people sample the available outcomes from memory and choose between them with a series of binary comparisons. That is, the value of an outcome is its rank among the sampled alternatives – and not a global utility score. This hypothesis was based on findings that people are good at making comparative judgements (i.e. that *A* is better or worse than *B*), but are poor at estimating the magnitude of that relationship (reviewed by Stewart, Brown, & Chater, 2005).

Evidence against the use of utility scores for judgment was found in experiments that asked subjects to value a gamble from a set of options (e.g. Stewart, Chater, Stott, & Reimers, 2003). For example, when subjects were asked how much the following gamble was worth:

a 50% chance of winning £200;
£40, £50, £60, or £70?

Most chose £60. But when the options were:

£90, £100, £110, or £120?

Most chose £100 (reviewed by Vlaev, Chater, Stewart, & Brown, 2011). Expected utility theories assert that prospects have a stable utility that is independent of any options presented – people should choose the option that is closest to the utility of the prospect. That is, if the gamble is worth £80, people should select the £70 and £90 options. However, the experimental procedure demonstrated that the options presented can skew a person's perceptions of a prospect's value – suggesting a comparative approach to judgement that is based on the sample of options presented.

The implications of decision by sampling are consistent with the risk aversion, risk seeking, and loss averse predictions of utility models (in particular, prospect theory; §2.2.4), but are derived from accumulated comparisons from samples of outcomes rather than an assumed utility function curvature (Stewart et al., 2006). The sampling approach is also better equipped to explain temporal aspects of a choice: how the order of decisions or incidental priming can affect preferences (e.g. Ungemach, Stewart, & Reimers, 2011).

However, the rejection of utility scoring as a cognitive model is not detrimental to the use of utility models for understanding preferences. Most utility models do not depend on utility scores actually being calculated,[40] but rather

40. It is also unclear if they were ever intended to be cognitively admissible.

that the relative order of outcome utilities implies a certain ordering of preferences (discussed further in the following chapter). This only requires that outcomes be ordinally comparable to be able to construct a utility scale. Modern utility models – such as Kőszegi and Rabin's (2006, §2.3.3) – include parameters for psychological influences on this ordering through comparisons with malleable reference points and expectations. However, the assumption that utility scales are temporally stable is thoroughly refuted (reviewed by Stewart, Reimers, & Harris, 2015; Vlaev et al., 2011).

2.5 EXPERIENCED UTILITY

The idea of utility maximisation has its roots in Jeremy Bentham's (1748–1832) *utilitarianism* theory of normative ethics.[41] Bentham (1789/1948) defined utility as a measure of *pleasure* and *pain*: our two 'sovereign masters [that] point out what we ought to do, as well as determine what we shall do' (p. 1). He classified 14 kinds of pleasure[42] and 12 kinds of pain[43] that could be measured by their intensity, duration, certainty, propinquity, fecundity, and purity. Actions that promote pleasure should be pursued, and those that promote pain should be avoided.

41. Except for this section, unqualified uses of the term *utility* in this thesis refer to the economic utility described in §2.2.

42. Sense, wealth, skill, amity, good name, etc.

43. Privation, awkwardness, enmity, ill name, etc.

Kahneman, Wakker, and Sarin (1997) reintroduced these ideas as *experienced utility* to understand how people retrospectively evaluate events, and why such evaluation often conflicts with objective accounts of them. Kahneman et al. argued that people experience events (such as psychological experiments, surgical procedures, and personal relationships) as a series of discrete *moments*, rather than as a continuous stream. The construction of these moments, and biases in one's memories of them, introduces discrepancies between the recollection of an experience and what was actually experienced.

2.5.1 *Experience by Moments*

Experiences are not stored in a person's memory as if they were a photographic recording – with each second stored in perfect experiential and temporal fidelity. Rather, memories are formed from snapshots of the representative moments in an experience: the most significant or interesting episodes (Fredrickson & Kahneman, 1993). Analyses of these moments – both at the time they occur and retrospectively after the fact – are used to characterise experienced utility.

Kahneman et al. (1997; and further refined by Kahneman, 2000a, 2000b) distinguished four types of this utility:

- *Decision utility*. The economic utility discussed in Section 2.2.

- *Instant utility*. A 'measure of hedonic and affective experience, which can be derived from immediate reports of current subjective experience' (p. 376). That is, an account of the pleasure or pain in an experienced moment at the instant it is experienced.
- *Total utility*. A temporal integral of the instant utilities for the moments that constitute an event: a measure of the total amount of pleasure or pain that was experienced.
- *Remembered utility*. A retrospective account of the total utility measure: what a person remembers and is able to recall about their experience of an event.

Total utility is considered an objective measure: it is an account of the total pleasure and pain experienced by a person during an event from real-time measures as it is experienced. That is, it is a *ground truth* measure of a person's experience that is unaffected by their memory. Kahneman et al. proposed two rules that govern total utility:

- *Separability*. The order in which moments are experienced does not affect total utility.
- *Time neutrality*. All moments are weighted equally. That is, the time gap between experience and choice does not influence total utility – utility does not diminish over time.

However, experimental evidence has found that these rules do not hold for remembered utility – the retrospective account of total utility (Varey & Kahneman, 1992). Remembered utility is vulnerable to the psychological biases and frailties of memory, including:

- *Peak-end*. The most intense (i.e. most pleasurable or most painful) and terminating moments of an event have a disproportionally high influence on its remembered utility.
- *Duration neglect*. The total duration of an event has little influence on its remembered utility.
- *Violations of temporal dominance*. The remembered utility of a strongly negative event can be increased by extending it with moments that reduce the average pain, even if they increase the overall (summed) pain.

2.5.2 *Peak-End*

Kahneman, Fredrickson, Schreiber, and Redelmeier (1993) conducted a 'cold-pressor' experiment, which had subjects submerge their hands in water under two conditions: short and long. Both began with the subject's hand in unpleasantly cold water (14°C) for one minute. Then: in the short condition subjects removed their hand from the water, but in the long condition they kept their hand submerged for an additional 30 seconds while the water was surreptitiously warmed to a less-unpleasant 15°C. When subjects were asked

which condition they would prefer to repeat, most chose the long condition: preferring the longer unpleasant experience that had a marginally less unpleasant ending – effectively choosing more pain over less.

Subjects were also asked to report a real-time measure of their discomfort using a potentiometer (their instant utility). The discomfort during the first minute was comparable between conditions, but the final 30 seconds of the long condition had a significant drop in discomfort for the majority of subjects (the main finding was robust under replication without this measure). That is, subjects were able to recognise the drop in discomfort at the end of the long condition, and despite also correctly identifying that they endured it for more time, they felt it had less overall discomfort (their remembered utility). This apparent conflict was attributed to the *peak-end rule*: reports of subjective experience are dominated by the average of the most intense moment and the terminating moment.

The peak-end rule was also observed in retrospective reports from patients of a colonoscopy procedure, whose experience was dominated by their peak discomfort and by their discomfort at the end of the procedure – it was unrelated to the procedure's duration, which varied between 4 and 69 minutes (Redelmeier & Kahneman, 1996; Redelmeier, Katz, & Kahneman, 2003). Ariely (1998) also observed the rule in two experiments that applied pain to participants using a heating element or vice: they found a peak-end effect in global retrospective evaluations of the experience, and mixed results for the influence of its duration.

Judgements about pleasurable experiences are also subject to the peak-end rule. For example, Do, Rupert, and Wolford (2008) examined the perceived pleasure in receiving gifts. In their first experiment, two lists of DVDs were produced (*A* and *B*) and given to participants as part of a raffle. The *A* list was populated with highly-rated movies, and the *B* list with poorly-rated movies. Participants reported their pleasure upon receiving either: (a) one *A* movie, (b) an *A* movie and a *B* movie later, (c) a *B* movie and an *A* movie later, or (d) one *B* movie. Receiving a *B* movie alone was rated positively, but receiving an *A* movie followed by a *B* movie was rated worse than receiving an *A* movie alone. These findings were further supported in a second experiment that gave sweets to children on Halloween. They concluded their paper with some advice for gift giving: 'you might consider giving only the best one – or at least making sure that you give the best one last' (p. 98).

2.5.3 *Duration Neglect & Violations of Temporal Dominance*

The effects of duration neglect are closely related to the peak-end rule – results of both the cold-pressor and colonoscopy studies found subjects ignored the duration of the events in their attention to the peak and end moments.

However, duration neglect and violations of temporal dominance are also independent effects.

Fredrickson and Kahneman (1993) exposed subjects to films containing either pleasant or aversive imagery, and recorded both their real-time and retrospective affect ratings. They found that the duration of the films had only a small effect on retrospective evaluation, and that evaluations appeared to be based on a weighted average of the moments in the experience – that is, adding less-negative imagery to a substantially negative experience improved the reported evaluation of the overall experience. This contrasts with the assumption of *temporal dominance*, which asserts that adding a moment of negativity to an experience should make the overall evaluation more negative (as it reduces the sum of the experienced utility).

Schreiber and Kahneman (2000) conducted a similar series of experiments using unpleasant sounds of varied loudness and duration. Although they found peak-end responses to the stimuli, they did not find duration neglect: the duration of the experience had an additive effect on remembered utility. However, they did observe violations of temporal dominance: extending an aversive sound with a less aversive one improved the remembered utility of the total experience.

Related effects have been observed in many other fields, including article pricing (Nasiry & Popescu, 2011) and effortful study (Finn, 2010). However, the effects are sometimes subtle: for example, experiments involving purchase payment sequences found that peak-end effects were not observed when subjects were focused on the experimental manipulation, but became significant when they were concurrently engaged in a distractor task (Langer, Sarin, & Weber, 2005). Experiments on gastronomic experiences validated duration neglect and demonstrated preference for increasing pleasure across courses, but failed to validate reliable effects of peak or end experience (Rode, Rozin, & Durlach, 2007).

2.5.4 *Time Perception*

These effects on experienced utility are part of a broader range of influences on the perception of time (reviewed by Grondin, 2008), including:

- *Vierordt's law* (Lejeune & Wearden, 2009; von Vierordt, 1868). The durations of short events are overestimated, and the durations of long events are underestimated (with a smooth transition in the amount of error). This includes both events of immediate interest (such as in a controlled experiment to reproduce a beat) and retrospective accounts (Fortin & Rousseau, 1998; Yarmey, 2000).

- *Interruption effects* (Weybrew, 1984). People tend to overestimate the duration of interrupted tasks, and underestimate the duration of uninterrupted ones.
- *Rhythm effects*. The rhythm and pacing of stimuli influences people's perception of time through interference with their internal clock (e.g. McAuley, 1995; Treisman, Faulkner, Naish, & Brogan, 1990).

Such effects also have implications for models of choice when outcomes are received over time (such as retirement savings) – in particular, how people discount the utility of an outcome that may be received in the future (reviewed by Frederick, Loewenstein, & O'Donoghue, 2002; Loewenstein, 1987, 1988; Loewenstein & Prelec, 1992). This issue will be revisited in Chapter 7.

2.6 PSYCHOLOGICAL INFLUENCES

Although descriptive models, such as prospect theory, are used to explain the psychological influences on decision making within an economic context, they are generally not predictors of psychological influences. For example, the endowment effect (§2.3.1 and §2.6.2) can be understood as a manipulation of prospect theory's reference point, but 'prospect endowment' is not a parameter of the theory nor a consequence of its construction. Awareness of these effects is important for not only understanding behaviour, but also ensuring adequate controls in choice experiments to avoid confounding and misinterpreting results.

2.6.1 *Negativity Bias*

Studies across multiple areas of psychology have found that, in general 'bad is stronger than good' (Baumeister et al., 2001, p. 355). That is, negative events (such as losing money or receiving criticism) have a greater impact on people than corresponding positive events (such as gaining money or receiving praise). A single negative event can overpower the psychological effects of many positive ones.

In a review across psychological disciplines, Baumeister et al. (2001) found that almost without exception, negative events

- have stronger effects on initial impressions,
- are remembered in more detail and for longer,
- are responded to faster,
- are processed more thoroughly, and
- have a larger impact on learning.

In a similar review of such work, Rozin and Royzman (2001) described four aspects that characterise this bias:

- *Potency*. Negative events are more potent and of higher salience than corresponding positive counterparts.
- *Steeper gradient*. The strength of negative events increases as they are approached in space or time.
- *Dominance*. The holistic perception of an object with both positive and negative attributes is more negative than the algebraic mean of those attributes.
- *Greater differentiation*. Negative events are construed to be more elaborate than corresponding positive ones.

For example, negative attributes can spike an otherwise positive evaluation – but the inverse does not apply (reviewed by Highhouse & Johnson, 1996; Peeters & Czapinski, 1990).[44] These effects have also received considerable attention in the marketing and consumer satisfaction literature (Bettman, Luce, & Payne, 1998; Yi, 1990, §2.3.1).

44. 'A spoonful of tar can spoil a barrel of honey, but a spoonful of honey does nothing for a barrel of tar' [ложка дёгтя портит бочку мёда] (Rozin & Royzman, 2001, p. 296).

There is also evidence for a converse *positivity bias* when past experiences are being recalled – sometimes referred to as the *Pollyanna Principle* (Matlin & Stang, 1978). Recalling experiences of an event appears to engender a bias towards its positive aspects or a minimisation of its negative aspects – especially as the time between experience and recollection increases (Peeters, 1971; Rozin & Royzman, 2001).

Both positivity and negativity biases are speculated to have evolutionary roots, but it is not clear if they are contradictory effects or parallel processes (Baumeister et al., 2001; Taylor, 1991).

2.6.2 *Endowment Effects*

People exhibit a bias towards maintaining the status quo – a feature of a broader *endowment effect* (Thaler, 1980). Under this effect, goods that are currently possessed are valued more than those that are not, and the loss from giving up a good is greater than the gain of receiving it. That is, a person may prefer A to B if they currently have A, and B to A if they currently have B. For example, Knetsch (1992) randomly awarded subjects with either a mug or a pen, and asked if they wanted to swap. The trade was rejected by most subjects, indicating a preference for whichever item they held (i.e. regardless of its objective value or their preference if they were not endowed). This effect is not restricted to tangible objects and has been observed in hypothetical situations involving investment decisions, job opportunities, and healthcare plans (W. Samuelson & Zeckhauser, 1988; Sprenger, 2010).

However, the prospect of giving up something does not always engender the endowment effect, and the limits of the effect have been debated (Bateman, Kahneman, Munro, Starmer, & Sugden, 2005; Bateman et al., 1997).

For example, Chapman (1998) found there was a relationship between the strength of the effect and the similarity of the goods being traded: the more similar goods were, the weaker the effect. Similarly, when the good being given up is intended to be used for exchange (such as money or something highly fungible), there are no feelings of endowment for that item (Novemsky & Kahneman, 2005).

2.6.3 *Description Invariance Violations* (*Framing Effects*)

Normative economic theories assume that decisions do not depend on the way that outcomes are described: manipulating the phrasing or constant factors of a prospect's outcomes should have no effect on preferences. However, there are many examples where this principle of *description invariance* is violated under manipulations of *framing effects* (reviewed by Levin, Schneider, & Gaeth, 1998; Tversky & Kahneman, 1986).

The canonical example of a framing effect is the *Asian disease* problem (Tversky & Kahneman, 1981):

> An unusual Asian disease is expected to kill 600 people. Two alternative programmes are available to combat the disease:
>
> **A**: 200 people will be saved.
>
> **B**: there is a ⅓ probability that 600 people will be saved, and a ⅔ probability that nobody will be saved.
>
> *Or* (to a second group of subjects):
>
> **C**: 400 people will die.
>
> **D**: there is a ⅓ probability that nobody will die, and a ⅔ probability that 600 people will die.

The outcomes of A and B are identical to C and D, but phrased in terms of either the people that 'will be saved' or 'will die', respectively. However, Tversky and Kahneman (1981) reported that subjects were willing to accept A over B, but rejected C for D.

Levin et al. (1998) reviewed three types of framing manipulations that have been found to effect choices:

- *Risky choice framing*. Outcomes that involve different levels of risk are described differently.
- *Attribute framing*. Some characteristic of an object or outcome is emphasised over another.
- *Goal framing*. The goal of an action is emphasised over the outcomes.

For example, people respond more positively to surgical procedures described in terms of survival rates than mortality rates (attribute framing), and are more willing to forego a gain than sustain an equivalent loss (goal framing). Reviewing studies across many different task domains, Levin et al. found

framing to have a significant and consistent impact on decisions (although the degree to which these are manipulations of the reference point or simply outcome salience has been debated [Bordalo, Gennaioli, & Shleifer, 2013; Kühberger, 1998; van Schie & van der Pligt, 1995]).

Description invariance violations are also found in prospects that stimulate *regret*: when the comparison between a selected outcome and what might have been received under another reduces the overall satisfaction of the selected outcome (Loomes & Sugden, 1982, 1987). Similarly, if outcomes have the potential to fall short of expectations (§2.3.2), then people exhibit an aversion to those that may trigger feelings of disappointment (Bell, 1982).[45] For example, if offered a 60% chance at £13 or a guaranteed £7, the latter is more appealing because of the sensation of regret that would be felt if the gamble at £13 resulted in nothing. However, this preference reverses when the odds are halved and the choice is between a 30% chance at £13 and a 50% chance at £7 (Loomes, 1988).

45. And vice versa for feelings of joy and elation at the possibility of receiving some outcomes (Bell, 1985; Loomes & Sugden, 1982).

Axiomatic models of disappointment and regret (e.g. Bell, 1985; Gul, 1991; Loomes & Sugden, 1982) have been used to explain the Allais paradox (§2.2.2; Loomes & Sugden, 1986), and are unique amongst expected utility models in that they permit outcomes within a prospect to influence the utility of each other (cf. decision by sampling, §2.4). There is also evidence that suggests the strength of this effect interacts with the framing of the prospect (Starmer & Sugden, 1993).

2.6.4 *Procedure Invariance Violations*

Decisions should (normatively) also be invariant to the procedure used to elicit them: for example, asking if someone would accept a particular outcome is logically equivalent to asking if they would reject the competing alternatives. However, Shafir (1993) found that compatibility between the scales used to describe prospects and the requested response can affect choices. In particular, positive aspects are weighted more when asked to *choose*, and negative aspects are weighted more when asked to *reject*. In an experiment that presented subjects with the pros and cons of two alternatives and asked them (between-subjects) to either choose or reject one (e.g. award/deny custody, choose/give up an ice-cream flavour, voting for/against a candidate), Shafir found that responses were not complements of each other. The options with more substantial positive and negative features were both chosen and rejected more often than the blander alternatives.

Similarly, Fischer, Carmon, Ariely, and Zauberman (1999) found that preferences amongst logically equivalent alternatives (job offers and consumer products) could be systematically manipulated by controlling the salience of

the attribute that subjects had to differentiate. For example, subjects were given a description of two job options:

A: $31,500 annual salary with 15 vacation days.

B: $36,500 annual salary with 5 vacation days.

In a *choice* condition subjects simply chose their preferred option, but in a *matching* condition (between-subjects) the salary of Option B was hidden and subjects specified a salary that would make the two options equally attractive. For subjects in the matching condition, their preference between A and B in a hypothetical choice condition can be inferred from their response – for example, a response that a salary of $40,000 would make B equivalent to A strictly dominates the $36,500 in the original option: indicating that A > B (Tversky, Sattath, & Slovic, 1988). Fischer et al. found that in choice conditions, most subjects preferred the job with the higher salary (the salient attribute), but this preference reversed in the matching conditions.

This example is part of a larger class of paradoxes known as *preference reversals* (reviewed by Camerer, 1995): people's choices are correlated with an outcome's probabilities, but the prices people assign to outcomes are correlated with its pay-offs (Slovic & Lichtenstein, 1968). For example, Lichtenstein and Slovic (1971) presented subjects with pairs of bets that had roughly equal expected values, such as:

P: a 99% chance to win $4, and a 1% chance to lose $1.

$: a 33% chance to win $16, and a 67% chance to lose $2.

That is, a bet with a high probability of winning a small amount (P) and a bet with a low probability of winning a large amount ($). When asked which bet they would rather play, most subjects chose the P bet, but when asked to price the bets (i.e. if they owned a ticket to play each bet, what is the minimum price they would sell it for), most priced the $ bet higher.[46]

46. Later replicated with casino patrons placing real bets (Lichtenstein & Slovic, 1973).

Evidence of these reversals, where a 'preference measured one way is the *reverse* of preference measured another and seemingly theoretically compatible way' (Grether & Plott, 1979, p. 623), generated significant debate about the possible economic, experimental, and psychological explanations (reviewed by Grether & Plott, 1979; Lichtenstein & Slovic, 2006; Slovic & Lichtenstein, 1983). Tversky, Slovic, and Kahneman (1990) examined several competing explanations, concluding that preference reversals are predominantly due to the *scale compatibility*: asking subjects to *choose* between gambles emphasises their probability components, while asking them to *price* the gambles emphasises their pay-offs. The more salient component is given more weight in the decision-making process (Tversky et al., 1988).

2.6.5 Sunk Costs & Mental Accounting

Investments that have been made and cannot be recovered are known as *sunk costs*. Individuals that experience sunk costs through a significant financial investment are found to have a tendency to invest further – even when the return is no longer worthwhile (reviewed by Thaler, 1980). For example:

> A man joins a tennis club and pays a $300 yearly membership fee. After two weeks of playing he develops a tennis elbow. He continues to play (in pain) saying 'I don't want to waste the $300!' (Thaler, 1980, p. 47)

Normative economics argues that only incremental costs (e.g. if the tennis club had a $25 monthly fee) should influence decisions (cf. McAfee, Mialon, & Mialon, 2010), but these sunk costs have strong psychological effects (Arkes & Blumer, 1985).

Tversky and Kahneman (1981) reported the results of an experiment which demonstrated an interaction between sunk costs and framing effects. Subjects were presented with the following hypothetical scenario (p. 457):

> Imagine that you have decided to see a play where admission is $10 per ticket. As you enter the theatre you discover that you have lost a $10 bill. Would you still pay $10 for a ticket for the play?

A majority of subjects said they would pay. But when the ticket was described as lost after it had been purchased and the question was whether or not to pay for another, their responses were evenly split. Both scenarios have the same monetary premise: $10 has been sunk; and the decision is financially the same: do you spend another $10? However, the framing of the circumstances under which the premises are faced affects how people choose.

Thaler (1980; 1985) argued that problems involving sunk costs are treated as *mental accounting* exercises: people mentally segregate resources into accounts, and respond to prospects as if their outcomes only manipulate the stated accounts (rather than the aggregate of all relevant accounts). For example, a '$10 bill' is not mentally fungible with a '$10 play ticket': losing the former has no bearing on the desire spend another $10 to see the play, but losing the latter appears to increase the cost of seeing the play to $20 and decreases its desirability (Tversky & Kahneman, 1981). Thaler (1999) reviewed observations of mental accounting in personal finance, gift giving, and myopic loss aversion.

2.7 HUMAN–COMPUTER INTERACTION

There are few direct applications of this literature within human–computer interaction despite interaction decisions sharing many features with those studied in the economic and psychological literature (detailed in the following

chapter). Interactive tasks introduce new aspects to decision making with a system's ability to adapt to a user's choices and customise the alternatives presented to them. That is, users do not experience computer systems passively – they are actively engaged in perceiving feedback from the system and deciding how to respond. For interaction designers, this engagement is also viewed from the system's perspective: processing the user's input and deciding upon the best feedback to give them (e.g. Bunt, Conati, & McGrenere, 2007; Findlater & McGrenere, 2010; Gajos, Czerwinski, Tan, & Weld, 2006). Understanding the relationship between system behaviour and user behaviour is an active area of research; however, only recently has the literature on psychological biases and economic models begun to receive attention.

2.7.1 Emotion & User Experience

Studies of emotion during interaction investigate the relationship between users and technology: how elements of interaction influence affective experience with computer interfaces (reviewed by Brave & Nass, 2008; Hassenzahl, 2005). Despite computers being deterministic machines, users often anthropomorphise them by projecting social qualities onto their behaviour: they can be perceived as co-operative (Nass, Fogg, & Moon, 1996), elicit courtesies (Nass, Moon, & Carney, 1999), and felt to have a personality (reviewed by Nass & Moon, 2000). While the connection to human social biases studied in psychology (such as negativity bias, §2.6.1) has not been well-studied, there is extensive research on emotional responses to interaction experiences.

Ceaparu, Lazar, Bessiere, Robinson, and Shneiderman (2004) used survey and diary studies to characterise the events that evoke user frustration with interfaces. These included flaky network connections, system crashes and freezes, unclear error messages, and other events that require time and effort to fix – particularly when they distract from an important primary task (Lazar, Jones, Hackley, & Shneiderman, 2006). An important moderator of this frustration is the user's knowledge, experience, and feelings of self-efficacy (Bessière, Newhagen, Robinson, & Shneiderman, 2006; Jokinen, 2015; Kay, 2008; Wilfong, 2006): users with higher self-efficacy beliefs have reduced feelings of frustration, anxiety, and anger. However, these studies are very broad in their objectives and do not experimentally manipulate these factors (they are based on post hoc surveys or diaries). Therefore it is difficult to model these findings in a causal relationship (such as whether self-efficacy reduces the incidence of frustrating experiences, or reduces the sensation of them through acclimatisation).

Studying more positive experiences, Hassenzahl, Diefenbach, and Göritz (2010) found a strong connection between need fulfilment and positive affect – in particular, the stimulation, relatedness, competence, and popularity of

an interface influenced affective responses from users and subsequent assessment of user experience (see also Tuch, Trusell, & Hornbæk, 2013; Venkatesh, 2000). Picard and Klien (2002) reviewed how computer systems could better support the emotional needs of users: including how to detect a user's emotional state (e.g. frustration or despair with the system) and intervening by modelling human social skills (e.g. active listening and empathy).

2.7.2 *Subjective Measures of Experience*

Quantitative analyses of user experience have focussed on how interface attributes, such as aesthetic quality and usability, influence user satisfaction and preferences (reviewed by Hartmann, Sutcliffe, & De Angeli, 2008). However, experimentally manipulating and measuring these factors is a difficult and unsettled problem, and studies often have conflicting results.

For example, Tractinsky, Katz, and Ikar (2000) found a strong relationship between the aesthetics of an Automated Teller Machine (ATM) user interface and perceptions of its usability – concluding that 'what is beautiful is usable' (p. 127). This was challenged by Hassenzahl (2004), who found no such relationship in an analysis of MP3 players – but was supported by a study of text-entry interfaces (Ben-Bassat, Meyer, & Tractinsky, 2006), and with mixed results for web pages (Lavie & Tractinsky, 2004; Thielsch, Engel, & Hirschfeld, 2015). In a meta-analysis of 25 reported correlations between aesthetics and usability across the literature, Hassenzahl and Monk (2010) found a general lack of consistency in both the methodology and results of these studies.

In response to this medley of results, researchers have discussed the complexity in accurately measuring subjective attributes and highlighted methodological differences between studies (Hassenzahl & Monk, 2010; Tractinsky, 2004). In particular, the manipulation and measurement of aesthetic and usability factors has varied substantially: Tractinsky et al. (2000) manipulated aesthetics exclusively through the interface's layout, whereas Hassenzahl (2004) used the layout and visual design, and Ben-Bassat et al. (2006) used only the visual design. Tractinsky (2004) also raised concerns that within-subjects experimental designs common in human–computer interaction research allow subjects to make comparative judgements between all possible design alternatives (rather than an independent value judgement), which does not reflect the environment that these decisions are typically made in.

To improve measurement validity, techniques from economics have been applied in recent work. Ben-Bassat et al. (2006) paid subjects according to their task performance and had them make auction bids on different interfaces they could use to complete tasks – hypothesising that the price subjects were willing to pay is indicative of the interface's value to them. They found

subjects bid significantly higher for conditions/interfaces that had easier task requirements, and that an interface's aesthetic quality had no effect on bid values. Similarly, Toomim, Kriplean, Pörtner, and Landay (2011) used labour supply curves and survival analysis to determine how much subjects would need to be paid to complete pointing tasks of varied difficulties, or data entry tasks with distracting aesthetic elements.

However, the trade-off subjects made in these experiments was to decrease their financial reward in return for easier experimental tasks (e.g. fewer actions), or perseverance with a difficult task. That is, the quality of the interaction itself was not manipulated[47] – rather, it was the difficulty of the tasks that subjects were asked to complete that elicited the effect. While this may be useful for plotting an indifference curve for a particular interface design or interaction technique,[48] it does not necessarily reflect preference.

47. Typing in Ben-Bassat et al. (2006) and pointing in Toomim et al. (2011).

48. The equity between task difficulty and price.

Furthermore, although auctions and labour supply models are well-established for economic problems, their methodological validity for interaction problems is unclear. In particular, it has not been established that willingness-to-pay or willingness-to-accept measures are indicative of user choice when the financial aspect is removed (nor the influence of the endowment effect, §2.6.2). Interactive outcomes rarely involve direct monetary incentives or costs, and although reducing interactive decisions to a monetary value facilitates analysis of those decisions, it does not reflect the environment such decisions are actually made in.

2.7.3 *Subjective and Objective Measures*

Objective measures of user performance – for example, the time to complete a task or the incidence of errors – are easy to obtain and are the most common device for inferring interface calibre. However, as with the paradoxes observed between the objective value of prospects and the choices people make, there are gaps between objective and subjective[49] measures of experience.

49. Sometimes referred to as *usability*.

Hornbæk (2006) reviewed current practices in assessing the subjective quality of interfaces in human–computer interaction research, including measurements of an interface's effectiveness, its efficiency, user satisfaction, and user attitude. Brave and Nass (2008) reviewed empirical methods for measuring subjective experience, including neurological measures (e.g. EEG and fMRI), autonomic activity (e.g. heart rate, pupil dilation, and skin conductivity), and self-report measures.

While it is widely recognised that an interactive experience's subjective qualities are important to its overall value, there is little agreement on the methods for measuring them (Frøkjær, Hertzum, & Hornbæk, 2000; Hornbæk, 2006; Hornbæk & Law, 2007; Kissel, 1995; Nielsen & Levy, 1994). In

a meta-analysis of 57 studies, Nielsen and Levy (1994) found that objective performance measures predicted user preferences in 75%. However, a more recent meta-analysis of 73 studies by Hornbæk and Law (2007) found a much weaker correlation between efficiency and satisfaction ($r = .20$), commenting that 'perceptions of phenomena are generally not correlated with objective measures' (p. 617).

Some of this variance may be due to different methodological approaches for collecting subjective data, as fewer than 10% of the papers reviewed by Hornbæk (2006) reported any efforts to confirm the validity or reliability of their satisfaction metrics. The inconsistency in measurement makes it difficult to assess and meaningfully discuss user preferences, or to identify trends and biases as they vary with interface manipulations.

2.7.4 *Interaction Biases*

User Performance. A cognitive bias specific to active interaction (i.e. where subject behaviour influences outcomes) is the *paradox of the active user*: expert users often settle on suboptimal strategies for completing tasks, and are reluctant to learn methods that could improve their overall productivity (Bhavnani & John, 2000; Carroll & Rosson, 1987). Fu and Gray (2004) found users preferred methods that were well-practiced and could be applied generally because they required less cognitive effort to maintain. For example, Lane, Napier, Peres, and Sándor (2005) found that even highly experienced users of Microsoft Word preferred to use menus and toolbars over keyboard shortcuts – despite the significant potential performance savings that would amortise the learning costs. Subsequent work has refined this paradox into models of cognitive satisficing (Charman & Howes, 2003; Fu & Gray, 2006; Gray, Sims, Fu, & Schoelles, 2006, cf. §2.1.2).

Examining how users perceive their own performance during interaction, Nicosia, Oulasvirta, and Kristensson (2014) had subjects complete pointing tasks in two interfaces and retrospectively judge their performance. In the first experiment, the two interfaces had identical pointing difficulties but with different width and size configurations. The results showed that subjects could reliably judge time performance, but that the distance and width also had an impact on their judgement (despite there being no theoretical difference in difficulty). A second experiment developed a model for the probability that a user could make such a judgement.

Interface Performance. Time perception effects (§2.5.4) have been observed in two studies of progress bars (Harrison, Amento, Kuznetsov, & Bell, 2007; Harrison & Hudson, 2010): subjects observed pairs of progress bars (both five seconds long) and chose which appeared to be faster. The first study found that users perceived accelerating rates of progress (i.e. slow start → fast finish)

to be faster than several alternatives, including decelerating progress. The second study found similar effects for progress bar animations: increasing rates of pulsing animation were perceived as faster than decreasing rates. They partially attributed these findings to peak-end effects (§2.5.2).

Users are also known to quickly form impressions of an interface's performance from very little exposure, and that these impressions result in stable judgements. In a series of experiments, Tractinsky, Cokhavi, Kirschenbaum, and Sharfi (2006) found that users formed aesthetic impressions of web pages after only 500 milliseconds of exposure, which were correlated with their ratings after a further 10 seconds of exposure (see also Lindgaard, Fernandes, Dudek, & Brown, 2006; Schenkman & Jönsson, 2000). Further work has connected these impressions to evaluations of trustworthiness and perceived usability (Lindgaard, Dudek, Sen, Sumegi, & Noonan, 2011; van Schaik & Ling, 2008), and identified mitigating factors (van Schaik & Ling, 2009).

Information Processing Biases. Several studies have examined behavioural economics tasks in computing contexts – such as leveraging status quo biases to influence dietary choices (Lee, Kiesler, & Forlizzi, 2011), the effects of social information on investment decisions (Zhao, Fu, Zhang, Zhao, & Duh, 2015), and the effects of website information framing on quality judgements (Hartmann, De Angeli, & Sutcliffe, 2008). However, these studies primarily manipulate the information that is presented to users, and not the quality of the interaction – that is, testing economic and psychological experimental tasks and methodologies, but on a computer system.

Studies of choice behaviour have identified applications of cognitive work on decision-making processes (§2.4) for interactive tasks. For example, studies of *choice overload* (Schwartz, 2004) when users interact with search engine results have found that their behaviour is influenced by the ordering of results, and that the number of results effects their satisfaction when under time pressure (Chiravirakul & Payne, 2014; Oulasvirta, Hukkinen, & Schwartz, 2009). Such work underlies choice frameworks for understanding how users develop interaction strategies (e.g. Bhavnani & John, 2000; Payne & Howes, 2013).

2.7.5 *Utility & Prospect Theory*

Models from behavioural economics (such as those reviewed in §2.2) have been postulated to apply in interaction, but with limited experimental support. Jameson et al. (2014) developed a framework for understanding the issues created by choices during computer use – for example, which application to use, whether to customise an interface or learn shortcuts, or how to configure a tool. Their framework focussed on six *choice patterns* that each represents an approach to evaluating alternatives and developing an interaction strategy. Within each pattern they reviewed issues for the goals and values

that users have, how users assess their context and options, how users learn from feedback and experience, and so forth.

Their analysis included prospect theory as a potential method for understanding how users evaluate the consequences of interaction outcomes (e.g. the time taken to complete a task relative to another method) – however, they cite no work that supports this hypothesis. In more targeted studies, Bergman, Tucker, Beyth-Marom, Cutrell, and Whittaker (2009) suggested that prospect theory may explain why users resist deleting files (the status quo bias establishes keeping a file as a reference point, and deleting it is a risky alternative; §2.3.1); and Banovic, Grossman, and Fitzmaurice (2013) briefly discussed the use of utility theories as a consideration when users are developing pointing strategies. However, the connection remains speculative.

Utility has also featured in interaction research as a potential assessment and user modelling tool. While the uses of the term are more general than the formulation of expected utility reviewed above, some of the models based around its premises have been successful in analysing interaction behaviour:

- Payne and Howes (2013) explored measures of utility and strategies of utility maximisation in a framework based on the influences of cognitive resources, context, and utility. They used the framework to understand interaction behaviour in button selection and online purchasing tasks.
- Norman (1983) modelled user satisfaction for an interface attribute (e.g. response time, display size) with a power function. Comparing the functions for pairs of attributes produces indifference curves for the trade-off between them, and can be collectively maximised to find an optimal interface configuration.
- Lelis and Howes (2011) developed and tested a utility model (although the measure of utility is unspecified) to describe how people use online ratings to inform their decision making.
- Gajos and Weld (2005) used utility functions learned from user feedback about user interface customisations to improve automatically generated interfaces.

However, these applications of utility are somewhat ad hoc as there is little direct connection to the broader economic or psychological models of utility reviewed earlier (§2.2).

2.7.6 *Persuasive Technology & Decision Support Systems*

In addition to observing behaviour, the *interactive* aspect of interfaces provides opportunities for helping users to make better decisions through *persuasive technology* and *decision support systems*.

Persuasive technology research studies how the ubiquity of technology can be used to influence people's behaviour and choices (reviewed by Consolvo, McDonald, & Landay, 2009; Fogg, 2003). Much of this work draws on cognitive models of behaviour from the social psychology literature (reviewed by Bandura, 2001; Baumeister, Vohs, DeWall, & Zhang, 2007), and studies their application and integration with interactive systems. For example, encouraging users to actively record and reflect on their experiences or negative behaviours (e.g. smoking or over-eating) can progressively improve subjective well-being and reduce bad habits (Hollis, Konrad, & Whittaker, 2015; Isaacs et al., 2013). Consolvo et al. (2009) reviewed major social theories and design strategies for effecting lifestyle behaviour changes through persuasive interactions; and Klasnja, Consolvo, & Pratt (2011) critiqued and reviewed the methodologies for evaluating interfaces that implement them.

In a similar vein, decision support systems aim to improve the choices people make by helping them analyse and process the available information. For example, Edwards and Fasolo (2001) reviewed a number of rules and devices for helping people make better decisions as part of a 19-step decision-making model. Many of their steps involved the routine application of the utility and probability models reviewed earlier, which Edwards and Fasolo suggested could be automated.

Zhang, Bellamy, and Kellogg (2015) explored two methods for reducing loss aversion and conservatism[50] biases when making investment decisions through a user interface. Their interfaces displayed investment options to the user, and showed the calculated expected returns of each alternative – and were successful in reducing the impact of the biases (see also Gunaratne & Nov, 2015). A similar suggestion system was tested by Solomon (2014) for a fantasy sports game, and found that users came to over-rely on the suggestions – suggesting that such systems need to be carefully designed to avoid creating their own biases.

50. People insufficiently adjust subjective probabilities when new information is presented (Erev, Wallsten, & Budescu, 1994).

The choices that interface designers make for even the static components of a user interface can also have substantial consequences for user behaviour. E. J. Johnson and Goldstein (2003), and Choi, Laibson, Madrian, & Metrick (2004) discussed the impact that default options have on behaviour: in particular, the default mode of an opt-in or opt-out choice has a substantial impact on resultant behaviour.[51] This is because default options can reduce the required decision-making effort, imply some endorsement, or establish a reference point for evaluating the alternatives (Dinner, Johnson, Goldstein, & Liu, 2011). Although this work has not studied user interfaces, users are known to only rarely actively customise their user interfaces and are likely to be similarly impacted by the ease and persuasiveness of defaults (Furnell, 2005; Gross & Acquisti, 2005; Mackay, 1990, 1991; Page, Johnsgard, Albert, & Allen, 1996).

51. Organ donation and retirement savings, respectively.

A MODEL OF INTERACTIVE PREFERENCES 3

This chapter develops a utility model of preferences for interactive prospects and describes the methodological issues for testing it (demonstrated in the following two chapters).

DURING INTERACTION with computer systems, users choose methods for completing tasks and make judgements about the value of the system's response. A collection of methods and corresponding system responses can be analysed as an *interactive prospect* in a manner analogous to how psychologists and economists analyse monetary prospects. The *utility* of an interactive prospect is a measure of its value: how much it is worth to the user in consideration of what they invested in obtaining it. As with economic prospects, users like interactions with positive utility, dislike those with negative utility, and prefer interactions with higher utility to those with lower.

Interactive tasks are oriented around a *goal* that a user wants to achieve, and therefore the utility of a system's response can be measured in terms of the *progress* made towards completing that goal. Perceptions of this progress are moderated by the user's *expectations* regarding the amount of progress that should be achieved for the effort invested. High expectations reduce the utility that a system's response can provide and exaggerate its failings (and vice versa for low expectations and its successes). The relationship between goals, progress, expectations, and utility can be expressed in an expected utility model for interactive prospects.

Empirically testing these relationships requires careful manipulation of interactive outcomes. Unlike economic experiments, where outcomes can be plainly stated, interactive outcomes are complex and are mixed with a user's *ability* to operate the interface. For example, a user's typing ability determines a substantial portion of the performance they stand to gain from a text-entry suggestion system. This complicates the application of experimental and analytical techniques from behavioural economics research in human–computer interaction, and necessitates careful adaptation of their methods to the characteristics of interactive prospects.

This chapter develops a model of reference-dependent preferences for interaction based on the model of Kőszegi and Rabin (2006, §2.3.3). Existing cognitive models of interaction (GOMS and KLM) are first reviewed to serve as the basis for analysing the structure of interactive tasks (§3.1), followed by the definition of an interactive prospect using that structure (§3.2.1). The utility-based reference-dependent preference model for interactive prospects is then detailed (§3.2.2), followed by the methodological issues for testing and validating it (§3.3) – applied empirically in the following two chapters.

3.1 COGNITIVE MODELS OF INTERACTION

The structure of an interactive task can be understood using *cognitive models* that decompose it into a series of actions that a user performs and system responses that they receive (reviewed by Dix, Finlay, Abowd, & Beale, 2004; Sutcliffe, 2000). Cognitive models mechanically express the steps that a user must undertake to achieve a particular outcome and the choices that they will encounter as they navigate through the system. This makes them useful tools for identifying the invested effort and received outcomes of an interactive prospect.

The GOMS family of cognitive models (Card, Moran, & Newell, 1980a) has extensive research support (e.g. Gray, John, & Atwood, 1993; John & Kieras, 1996a, 1996b), and decompose an interactive task into a set of *Goals* that a user wants to achieve, which are completed using elemental *Operators* (perceptual/motor/cognitive actions) assembled into *Methods*, with *Selection rules* for choosing between competing methods. For example, (Card et al., 1980a) decomposed text editing tasks into their constituent goals (e.g. *locate-line* and *modify-text*) and methods for achieving them (e.g. *use-qs-method* and *use-lf-method*). Through a series of experiments, they discovered the operators that completed those methods, and the selection rules users employed to choose between them.[1]

Given a task structure, performance models predict objective qualities about how an expert user will execute the task. For example, the Keystroke-Level Model (KLM; Card, Moran, & Newell, 1980b) predicts the time for a user to execute a GOMS method by summing together time predictions for its constituent operators. The KLM has components for actions that include typing, pointing, mental consideration, and system response – with a set of rules for assembling them into a time prediction.[2]

These cognitive models allow complex interactive tasks to be analysed as a series of elemental actions that a user performs, and responses from the system that they receive (Card, Moran, & Newell, 1983). However, these cognitive models are strictly concerned with the system's operation and do not assess methods or responses in terms of their utility to the user.

The actions required from a user to complete a task can be specified using a GOMS method, and a user's performance in executing those actions can be predicted by a KLM model – but these are not necessarily indicative of the *cost*

1 Alternative models, such as CCT (Kieras & Polson, 1985) and TAG (Payne & Green, 1986), produce functionally identical results, but with different grammars and decomposition rules. Any such model is appropriate for the purposes of the present research.

2 More granular analyses can be achieved using *cognitive architectures*, such as ACT (Anderson, 1996) and EPIC (Meyer & Kieras, 1997a, 1997b), on human–computer interaction tasks (e.g. Ehret, 2002; Hornof, 2004; Kieras & Meyer, 1997; Kieras, Wood, & Meyer, 1997; Salvucci & Lee, 2003).

to the user of performing those actions. For example, a 500 ms KLM pointing operator may be comparable in time to a 500 ms mental operator but differ in the amount of overall exertion. That is, even if two operators yield the same outcome, a user may not be indifferent to a choice between them (e.g. they may prefer pointing to thinking). Although time is a convenient metric that is easy to measure, interpret, and use for comparative judgements, it is not a comprehensive metric for the work performed by a user, nor does it predict or explain preferences.

Similarly, the outcome of a system's response is an objective state that can be measured against the task goal,[3] but this is not necessarily indicative of its *benefit* to the user. That is, the delta between the system's current state and the user's goal state can be measured, and can be analysed with respect to whether it increases or decreases as a result of an action – however, this delta does not indicate whether or not the user's effort invested to produce it is recompensed. For example, a text entry system that offers to complete partial words may be objectively excellent if it always suggests the user's intended word, but effectively worthless if it only does so as they are about to type the last character.

3. Such as the remaining GOMS operators or KLM time.

The perception of an interaction's value is also dependent on a user's expectations of how they *believe* the system should behave. These beliefs come from the user's experience of similar systems, reasoning through a model of interaction they possess, or their desires about how they want the system to perform. The *reference point* this creates changes over time as a user gains experience with the system and becomes accustomed to its particular behaviours and level of performance. For example, a text entry system that automatically corrects errors may initially have positive value when it corrects true errors, negative value when it incorrectly replaces entered text, and no value (neutral) when it offers no correction at all. However, as a user gains experience with the system they may come to expect corrections for certain habitual errors or shortcuts (e.g. automatically replacing *im* with *I'm*). These corrections will have less value than before because the baseline for assessing outcomes has increased with the user's expectations, and the system now has a negative value when it fails to offer them.

Interaction therefore involves a collection of objective and subjective factors: the objective user actions and system responses that can be described using cognitive models of interaction, and the subjective cost of performing those actions and the perceived benefit of the responses. The specification of an interactive prospect can be developed from the objective factors, with a utility model describing their relationship to the subjective factors – detailed in the following section.

3.2 A UTILITY MODEL OF INTERACTION

This model is based on Kőszegi and Rabin's (2006) model of reference-dependent preferences (§2.3.3). Given two alternative interactions I_1 and I_2 for completing a particular task (e.g. two methods from a GOMS model), users determine their preference based on the overall utility U:

$$I_1 \succ I_2 \iff U(I_1) > U(I_2). \tag{3.1}$$

The overall utility of the interaction U is determined by the sum of utilities u for the constituent sub-interactions i: $U(I) = \sum u(i)$, described below. Each sub-interaction is a *chunk* of an action and system response that are psychologically perceived by the user as a single unit (e.g. each step in a GOMS model). Chapter 7 discusses how this summation is knowingly naïve due to psychological effects (such as peak-end; §2.5.2), but it is sufficient for now.

3.2.1 *Interactions*

Each sub-interaction i consists of some user action a that yields an actual system outcome $\mathbf{c}$, as well as some expected system outcome $r(a)$:

$$i : a \to \langle \mathbf{c}, r(a) \rangle. \tag{3.2}$$

Actions. Each action a may be an elemental operation (such as a GOMS operator), or a small collection of them that form a *unit task* that has a routine solution (such as pointing and clicking on a button) and does not rely on extensive problem solving.

Outcomes. The actual outcome $\mathbf{c}$ is a K-dimensional *consumption bundle* that represents updates to the system's state arising from the user's actions. These updates include internal state changes, feedback elements to the user, and the temporal components of those changes (e.g. the response time). The bundle is deterministic and invariant to the user: given some initial state, a particular action a^* always produces the same outcome $\mathbf{c}^*$.

A consumption bundle encapsulates the multi-dimensional properties of an outcome. For example, typing a character on a mobile device may yield outcomes on several dimensions, including:

1. The character is added to the internal text buffer.
2. The character is displayed in a text field.
3. A *click* sound is played.
4. A haptic vibration is produced.

Some of these dimensions are observable by the user, but others are internal and can only be inferred/confirmed by feedback or subsequent actions. For example, the user infers the state of the internal text buffer from the contents

of the text field, but they are not necessarily identical (such as when entering a password).

Reference Expectations. For each action a, the user has some *reference expectations* for the outcome derived from that action, $r(a)$. The function r is specific to a particular user and is allowed to vary with their mental state (such as their past experiences and affective disposition) – no particular constraints are placed on the factors that influence r. It is desirable that r derives expectations rationally from recent experience with the system, but it is likely to be influenced by the user's beliefs and desires (Kőszegi, 2010) – even if they are known to not be realistic (e.g. reference expectations for a system that is consistently slow may reflect a desire for faster performance).

As reference expectations represent outcomes, they are also represented using K-dimensional consumption bundles. Actual outcomes $\mathbf{c}$ and expected outcomes $r(a)$ are pairwise comparable: for instance, if c_k contains the time for the system to respond, $r_k(a)$ is the expectation for that time. To maintain this comparability, either bundle may contain dimensions that are irrelevant to the other. For example, if a user expects haptic feedback for each keypress but none is produced by the device, then the dimension for haptic feedback in their expectations $r(a)$ is comparable with a null dimension for haptic feedback in the actual outcome $\mathbf{c}$ (and vice versa for an interface that produces such feedback that the user is not expecting).

Reference expectations are determined in consideration of the action a invested, which includes the cognitive, perceptual, and mental processing effort of producing that action. Users are likely to expect greater outcomes as the effort of achieving them increases (potentially in excess of what the system can actually provide). Expectations also change as users become familiar with new forms of interaction. In general, expectations increase as devices and interfaces improve: faster processing, brighter screens, higher resolution, improved design, and so forth.

3.2.2 *Utility*

The utility u of an interaction i (see Equation 3.2) can be expressed with Equations 3.3–3.5, which are functionally identical to those from Kőszegi and Rabin (Equations 2.14–2.16 in §2.3.3):

$$u(\mathbf{c}|a) = m(\mathbf{c}) + n(\mathbf{c}|a), \tag{3.3}$$

$$m(\mathbf{c}) = \sum_{k=1}^{K} m_k(c_k), \tag{3.4}$$

$$n(\mathbf{c}|a) = \sum_{k=1}^{K} \mu\big(\underbrace{m_k(c_k) - m_k(r_k(a))}_{\text{marginal utility}}\big). \tag{3.5}$$

Equation 3.3 states that the utility of an outcome $\mathbf{c}$ given action a is the sum of two functions: m and n, which respectively determine utility values for objective progress and the user's sense of gain-loss for the progress attained.

Consumption Utility (Equation 3.4). The consumption utility function m determines the *objective* utility of an outcome – either an actual outcome $\mathbf{c}$ or an expected outcome $r(a)$. Consumption utility is analogous to other measures of progress towards a goal, such as those evaluated by GOMS/KLM. Importantly, it is an objective measure and does not encapsulate subjective elements of the user's experience. The function yields a quantity that is positive when the outcome advances towards the goal (or assists in advancing towards the goal), negative when it moves away from the goal, and zero when there is no change.

As shown in Equation 3.4, consumption utility $m(\mathbf{c})$ is the sum of consumption utilities across the K dimensions of its argument (a consumption bundle). Separately applying the function to each dimension of a consumption bundle allows the importance of each dimension to be weighted (that is, each dimension k has a separately shaped consumption utility function m_k). Dimensions closely tied to attainment of the goal are likely to have more utility than superficial changes (Bordalo et al., 2013; Heath et al., 1999). For example, showing the typed character in the text field (dimension two in the above text entry example) is likely to have a higher utility weighting than the dimensions for audio or haptic feedback. Similarly, the visual feedback is likely to have higher utility than changes to the internal text buffer (dimension one) as it is the only directly observable representation of it.[4]

4. A lack of agreement between the internal text buffer and visual feedback is likely to lead to a *large* disparity between actual outcomes and reference expectations at a later point in the task.

These utility functions m_k share the properties of classic economic utility functions (e.g. Figure 2.1 and reviewed in §2.2), and in particular must be 'differentiable and strictly increasing' (Kőszegi & Rabin, 2006, p. 1138). However, empirical work in economics (or in human–computer interaction) does not, and is unlikely to, directly seek to expose the exact functions – instead, studies typically demonstrate that the predicted preference between two outcomes occurs within the manipulated dimensions and constraints of the model (reviewed in §3.3).

Gain-Loss Utility (Equation 3.5). The gain-loss utility function n evaluates the subjective elements of a user's sense of value derived from any disparity between actual and expected consumption utilities. The function μ has the properties of prospect theory's value function, which incorporates features of diminishing sensitivity and loss aversion (shown in Figure 2.3 and reviewed in §2.2.4). The argument to μ is the *marginal utility* between actual and expected consumption utility: $m_k(c_k) - m_k(r_k(a))$. Marginal utility considers the equity between expectations for the invested effort on the right, and the

received reward on the left. If actual consumption utility exceeds expectations ($m_k(c_k) > m_k(r_k(a))$), the user experiences a gain; conversely, if actual consumption utility fails to meet expectations, the user experiences a loss. The function μ transforms the marginal utility gains and losses into subjective value assessments according to prospect theory's value function, with losses having a stronger effect than gains.

For example, if a user types a character and expects to see it appear immediately, but its appearance is delayed by a second, then the user experiences a loss with respect to their expectation (amplified by the value function μ). That is, the consumption utility along the outcome's time dimension for the expected immediate output exceeds that for the actual, delayed output: $m_k(1 \text{ second delay}) < m_k(0 \text{ second delay})$. As another example, if a user expects audio feedback for each key-press but none is produced, the user experiences a loss along the relevant dimension – however, this may be recovered by gains along another dimension that has a greater m_k weight (such as offering a correct suggestion that completes the word).

3.3 MEASURING INTERACTIVE UTILITY

This utility model, as with utility theories in general, assumes that its components have *quantitative* values (i.e. m_k and μ are measured on ratio scales). However, actual values for these components are rarely measured directly in experiments that test utility theories due to the challenges in measuring them – particularly the psychological contributions. Instead, theories of utility are typically tested using the predicted relationships between components, and the ordering of preferences they imply – which can be used to construct a utility scale. For instance, many of the experiments reviewed in Chapter 2 presented subjects with two prospects (A and B) with controlled inputs and outputs (investments and returns), and asked them to indicate their preference between them (a *forced-choice* methodology).[5] From the pattern of these $A > B$ or $B > A$ responses, utility relationships were constructed and axiom conformance was tested. That is, while the inputs and outputs of an economic prospect are easily measured (such as the monetary value of investments and probabilities of monetary returns), the psychological contributions (such as endowment, framing, and mental accounting) and utility scores are only measurable indirectly through comparisons with alternatives.

5. Indifference was rarely permitted.

Although the forced-choice methodology is readily amenable to testing new types of prospects, interactive prospects have features that make them hard to clearly express in the style of economic experiments:

- *Multi-dimensionality*. Interactive prospects are difficult to reduce to a single feature, and comprise multiple disparate components (such as

task progress, time required, invested effort, feedback awareness, etc.). Some of these components are split between the inputs to the prospect and its outputs (such as the time required to perform an action vs. the time waiting for the system to respond), and their relationship with each other may be orthogonal or covariant.

- *Ineffability*. Many of the components of an interactive prospect are difficult to reliably quantify (such as invested effort and task progress). Even when components are measured in the same units, they may be difficult to compare and may not be fungible (e.g. pointing vs. thinking time).
- *Skilful*. A user's ability to process and understand the feedback from a system, and their skill in operating its input devices and interfaces has an effect on the outcomes they experience. That is, different users performing identical tasks may invest slightly different actions and receive/perceive different outcomes depending on their skill and experience with the system.
- *Opacity*. Users depend on feedback from the system to inform them of outcomes, which may be indirect and incomplete. For example, selecting a *copy* command (e.g. a file or some text) does not typically produce any feedback to indicate a successful copy operation – success or failure is only discovered upon a subsequent *paste* command.
- *Uncertainty*. The risk in probabilistic outcomes is not known a priori, and therefore interactive prospects usually feature uncertainty (if not ignorance; §2.2.5). For example, the efficacy of a text entry suggestion system is never told to users (and varies with their typing skill) – they can only infer its efficacy from repeated use and observation.

These features of interactive prospects raise significant challenges for experimenters that seek to test their utility, including: how to design experimental interfaces (§3.3.1), which methods to use for measuring utility (§3.3.2), how to elicit preferences (§3.3.3), and how to analyse responses (§3.3.4).

3.3.1 *Materials*

Interactive experiences are a combination of objective and subjective components that contribute to a user's overall satisfaction with a system. While objective metrics are readily quantifiable for some aspects of a system's outcomes (such as the time to complete a task or the number of errors in doing so) it is less clear how to address the subjective side. Additionally, unlike economic prospects where the inputs are readily manipulated (e.g. an investment of $100), the inputs to interactions are less amenable to precise manipulation and require active participation from the user (such as mental or physical

exertion). For instance, any two actions may differ in the time for execution (measurable) – or in the perceptual, cognitive, or motor costs of completing the action (largely unmeasurable). Testing utility theories for interactive prospects therefore requires designing experimental interfaces that can control and/or measure gains and losses in these components around a known reference point.

Reference Interfaces. The reference point for an interactive prospect is a baseline interface that the experimental interfaces will be judged against (i.e. the intended $\mathbf{r}$ or $r(a)$ for each experimental interface outcome $\mathbf{c}$). The experimental interfaces are those under test, and feature some trade-off in their outcomes relative to the reference interface – for instance, an assistive feature that works for some tasks but not others, or an interface that reduces one dimension of effort in exchange for increasing another (e.g. an interface that replaces the physical exertion of navigating menus with the mental demands of memorising keyboard shortcuts). In many cases the reference interface will be the status quo or a bland baseline that isolates the trade-off in the experimental interfaces. To validate that exposure to the reference interface sets the intended reference point, it should be compared with experimental control tasks that provide subjects with only the gains (or losses) of the experimental interface – that is, control tasks that feature no trade-off and should be uniformly accepted (or rejected) when compared to the reference.

Experimental Interfaces. The experimental interfaces should be controlled to manipulate a limited number of components relative to the reference interface (and to each other). As described in the model above, each component has its own utility function (m_k) and unnecessary manipulation of components can confound the interpretation of subject preferences. If the intention is to reveal a particular utility function m_k, then only that component should differ between the reference and experimental interfaces.

Constructing trade-offs is difficult for interactive prospects because the gains and losses are often on different scales, and it is not always clear how to equate or separate them to measure their relative contributions to overall preferences. For example, a system may decrease the time to perform an action but increase the physical or mental effort required to do so – an exchange across different scales that are unlikely to be independently manipulable. Such problems are not unique to interaction, but are also present in psychological research on negativity bias (§2.6.1) – for example, is it difficult to construct positive equivalents of traumatic experiences (Baumeister et al., 2001; Rozin & Royzman, 2001). Ensuring perfect scale and manipulation compatibility between interactive prospects is not essential for testing utility models, but it does limit the interpretation of results and demands further experimental work to tease apart the details. Rozin and Royzman (2001)

discussed several methods for constructing alternatives when it is difficult to equate their components – such as comparing series of combined stimuli to show that choices are more positive/negative than their algebraic sum implies, or introducing positive/negative events into a series of neutral events to show an asymmetric response (revisited in Chapter 6).

3.3.2 *Methods*

Although it appears advantageous to measure preferences on a scale or have subjects rank alternatives, there is no natural unit for an interface preference scale (cf. valuing monetary prospects in monetary units), and it is difficult to ensure internal validity and comparability of responses (reviewed in §2.7.3). However, an overall measure of preference can be obtained from a binary *choice* between two alternatives. That is, if a user consistently expresses a preference for interface A over interface B, it can be inferred that the combined objective (known) and subjective (unknown) experience of A is greater than that of B.[6] Although this does not indicate the strength of the preference, by tightly controlling the objective factors that distinguish A and B, repeated experimentation can uncover the ordering of preferences for its components and their respective weightings (e.g. §2.6.4).

6. And by Equation 3.1, A has a greater utility.

However, a forced-choice methodology implies that the completeness axiom and asymmetry of '$>$' holds[7] (§2.1.1). This may be permissible when the difference between outcomes is small and comparable along a limited number of similar dimensions (such as two text-entry suggestion algorithms), but becomes harder to justify when there are complex trade-offs between the outcomes (such as text entry using a keyboard vs. voice recognition). As it becomes harder for subjects to resolutely prefer one outcome over another, it likewise becomes harder for experimenters to understand or attribute meaning to their preferences patterns. In particular, critiques of preference patterns (e.g. tests of transitivity; §2.1.1) rely on these assumptions being defensible (Regenwetter & Davis-Stober, 2012).

7. Any two outcomes have either a '$>$' or '$\sim$' relationship.

Another commonly used methodology in behavioural economics has subjects matching outcomes to make them equivalent (e.g. the preference reversal experiments reviewed in §2.6.4). These methods find the point where two trade-offs balance each other and subjects would respond to a choice between them with indifference. Once this balance is found, a weighting for the different components of the prospect can be established. Matching methods include directly asking subjects for a balancing value, or repeatedly testing small manipulations with binary choices to home in on the point of indifference (e.g. Fischer et al., 1999). Patterns in these points of indifference can indicate the utility weightings of salient factors and other biases.

3.3.3 *Design*

There are many possible experimental designs for exposing subjects to alternatives and eliciting choices from them. The major trade-offs in alternative designs can be seen in the differences between how psychologists and economists have approached their experiments for studying behavioural biases (reviewed by Camerer, 1995; Camerer & Loewenstein, 2004; Hogarth & Reder, 1986b):

- Psychologists typically pose hypothetical questions to subjects, framed by descriptions of ordinary situations that subjects can empathise with; whereas economists typically lay bare a prospect's outcomes.
- Psychologists avoid giving subjects financial incentives to prevent contaminating the trade-off under test and confounding the primary dependent measure; whereas economics experiments are almost exclusively concerned with financial outcomes.
- Psychologists focus on subjects' first responses to a stimulus and avoid repeating choices; whereas economists are interested in how responses regress towards an equilibrium.

These differences are not matters of methodological superiority, but of differences in the underlying questions. Human–computer interaction researchers have interests that are closer to those of psychologists than economists when they seek to understand behaviour (rather than the prescriptive approach that is typical of economics), but they also have independent questions that will lead to distinct methodologies.

For instance, interactive choices are rarely made only once – they are usually repeated many times over the system's useful life. A user may be prompted each time the need for a decision arises, or their choice may be encoded as the default behaviour for the system. The repeated evaluation of an interactive prospect provides opportunities for potential costs of the choice to be recovered, but also may expose the user to an intertemporal decision about future interactions – which have their own psychological considerations (Loewenstein & Prelec, 1992, revisited in §7.6.2). The environment for these choices is difficult to replicate in a laboratory experiment, and subject responses for an experimental task in the short-term may differ from those when faced with long-term use. For example, Wedell and Böckenholt (1990) found that preferences for gambles became more consistent with their expected value when they were played multiple times.

Hershey, Kunreuther, and Schoemaker (1982) reviewed five sources of bias in procedures that may influence the results of utility theory experiments:

- *Response mode.* What is the method used by subjects to indicate their preferences?

- *Risk dimensions*. How many factors need to be distinguished between outcomes, and to what depth?
- *Domain*. Are the outcomes pure losses, pure gains, or mixed?
- *Transfer versus assumption of risk*. Are subjects being asked to give up the risk of an outcome, or obtain it?
- *Context*. Under what context is the decision being made under?

Even if the stated outcomes are identical, procedural differences along these dimensions can impact subjective responses. These may be intentional (such as tests of context effects), but requires cognisance when interpreting results or comparing them across experiments.

Experimenters also need to decide upon the experimental conditions that will be administered, and the assignment of subjects to conditions. Of course, if subjects are never exposed to a real trade-off (e.g. the experimental conditions objectively dominate reference conditions) or the set of experimental conditions do not sample a balanced range of trade-offs, then the results will be facile. A few experimental structures have support in the literature as standards for testing paradoxes of choice. For example, the HILO structure tests common consequence and common ratio effects (Chew & Waller, 1986), and preference reversal bets (Lichtenstein & Slovic, 1971). However, such methodologies often require that probabilities and outcomes can be arbitrarily combined (which is difficult to do for many interactive tasks).

Keren and Raaijmakers (1988) discussed the issue of within-subjects versus between-subjects experimental designs in tests of utility theories. Most economic work administers experimental choices using a within-subjects design due to the high experimental power and subject economy,[8] but this raises concerns about cross-condition contamination. In particular, if stimuli are highly similar then subjects may become aware of the experimental manipulation and adjust their responses to be proportionate (e.g. their reference point shifts to the average of the other conditions they have been exposed to, rather than the intended reference interface set by the experimenters). Between-subjects designs avoid these problems by only exposing subjects to a single choice, which preserves their first-impression responses and better reflects the environment that such choices are typically made in. However, between-subjects designs lose the ability to test for preference patterns within subjects (Hershey & Schoemaker, 1980).

8. Some exceptions were noted in §2.6.

3.3.4 *Analysis*

If all subjects have performed the same actions and received the same outcomes, then preferences across subjects can be tested with binomial sign tests (e.g. Birnbaum & McIntosh, 1996; Gonzalez & Wu, 1999; Tversky & Fox, 1995;

Wu & Gonzalez, 1998). However, as the number of choices being compared increases, so too does the chance of a Type-I error, and applying statistical corrections for those comparisons similarly increases the chance of a Type-II error (reviewed by Regenwetter et al., 2011). Statistical techniques for analysing the pattern of choices within subjects (e.g. the transitivity of their preferences; §2.1.1) remains an area of active research (reviewed by Regenwetter et al., 2011; Regenwetter & Davis-Stober, 2012).

In human–computer interaction experiments, the skilful nature of outcomes means that subjects are unlikely to all perform the same actions and receive the same outcomes. Variability in individual skill and ability across subjects will cause identical prospects to have different realised outcomes. For example, studies of pointing techniques typically manipulate the Fitts' law *Index of Difficulty* (ID) model (Fitts, 1954; Wright & Lee, 2013), and although the model is robust, its parameters vary with the individual (reviewed by Schmidt & Lee, 2005). A pointing technique that objectively lowers difficulty by two bits will not realise identical time savings across all users or compel an identical influence on their preferences (e.g. Quinn, Cockburn, & Delamarche, 2013; Quinn, Cockburn, Räihä, & Delamarche, 2011). In such cases, the realised outcome is itself a measurable variable that may influence choice.

If the outcomes that were actually received can be measured (such as time gained or lost), then choices can be analysed using a logistic regression – a regression model for a binomial dependent variable (i.e. the choice for either *A* or *B*). Logistic regression models predict the probability of the choice being one way or the other, given the outcome that was received. Logistic regression modelling is common in econometrics (reviewed by Cramer, 2003), and has been used to isolate variables in models of choice (e.g. Kusev, van Schaik, Ayton, Dent, & Chater, 2009; Larrick & Boles, 1995).[9]

9. Multinomial versions are used for multiple or ranked choices (reviewed by Louviere, 1991).

3.4 SUMMARY

The utility model of interaction presented in this chapter (§3.2) describes the relationship between a user's expectations and their evaluation of system outcomes: whether an interaction is perceived as a gain or a loss, and the influence of those gains and losses on overall preference. The model considers objective changes to a system's state, the utility weighting of each change, and the subjective sensation of gain and loss of utility. Although the model is based on established findings and models from behavioural economics, it is unfamiliar to human–computer interaction research and requires experimental validation in human–computer interaction contexts. This is challenging because the model describes interactive preferences in general terms (e.g. there are few constraints on how reference expectations are developed), and

it is not obvious how to adapt experimental methodologies from behavioural economics to human–computer interaction. These methodological issues are not specific to the model developed here but are general issues with experiments that aim to test preferences for interactive outcomes (§3.3). As reviewed in the previous chapter, small changes to experimental designs can have substantial effects on their results – but utility models that encapsulate these design choices aid in controlling and understanding these effects.

Given the broad scope of the model, three key aspects of it will be the focus of the experimental work in the remainder of this thesis:

- The salience of perceived progress towards a task goal in the weighting of utility functions m_k.
- Reference-dependence in the evaluation of interactive outcomes (the marginal utility in n).
- The psychological aversion to losses of progress in μ (loss aversion).

The following three chapters describe experiments that undertook this work. Chapter 7 reviews their results in the context of the model, discusses future directions for establishing support (including the use of risk), and potential applications of the model.

LOSS AVERSION IN INTERACTION 4

This chapter presents two experiments demonstrating reference-dependent loss aversion for interactive tasks – as described by the model in the previous chapter.

THE MODEL presented in the previous chapter makes claims and predictions about how users evaluate interactive prospects. A thorough examination of the model's claims requires extensive experimental work and methodological development (beyond that which is possible in a single thesis). However, its fundamental hypotheses are readily testable, and therefore the experimental work in this thesis focusses on the broad predictions it makes about task progress, the reference-dependence of preferences, and loss aversion. These aspects are fundamental to the model and establish parallels with the literature reviewed in Chapter 2.

In the two experiments that follow, a choice-based methodology was used to examine subjective preferences for interactive tasks engineered to contain controlled elements of progress gain and loss. In both experiments, subjects were given a choice between a neutral *reference* interface that worked equally well for all tasks, and an *experimental* interface that sometimes assisted task completion (yielding outcomes above expectations) and sometimes impeded it (yielding outcomes below expectations). By forcing a choice between these two interfaces across varied conditions of assistance and impediment, the magnitude of gain necessary to compensate for losses was revealed. The asymmetry of this compensation indicates biases for or against the differences between the two interfaces.

With respect to the model, the key difference between the two experiments concerns the utility of the actual outcome $m(\mathbf{c})$. Actions a and expectations $r(a)$ were largely held constant, and a dimension of the outcome $\mathbf{c}$ was manipulated to produce a controlled gain or loss (while still enabling subjects to complete the task successfully). In Experiment 4.1, certain sub-interactions required subjects to backtrack in order to complete the task – perceptually removing progress already attained. In Experiment 4.2, analogous sub-interactions were engineered to require actions and time costs comparable to those of Experiment 4.1, but without the appearance of backtracking and negative progress. These manipulations examine the consumption utility function m_k and its effect on overall utility u.

Both experiments were built around a text selection task, with the reference and experimental interfaces differing in their method of selection. Text selection provides a convenient method for examining the predicted effects of

the model because it uses simple mouse interactions that are well understood, allows for precise control over the objective gains and losses in time, is a familiar task for subjects, and exemplifies a design trade-off that has ecological validity (i.e. real interfaces incorporate the experimental behaviour).

The basic experimental method involved subjects completing a set of tasks with a conventional *letter-by-letter* selection technique (where text was selected one character at a time) to establish a reference point for *neutral* performance, and a set of identical tasks with a *word-by-word* technique that snapped the selection to word boundaries. After completing tasks with both letter-by-letter and word-by-word techniques, subjects chose which they preferred to use for a third set of identical tasks.

4.1 EXPERIMENT 4.1: OBSERVING LOSS AVERSION

Experimental tasks involved dragging a selection across an underlined portion of a sentence. The letter-by-letter technique selected all characters between the start of the dragging action and the current position of the cursor. The word-by-word technique began similarly, but snapped the selection to word boundaries after the selection crossed a space. This snapping behaviour could be disabled by backtracking the selection to anywhere within the initial word (reducing the selection back into the word it began in), after which selection followed letter-by-letter – illustrated in Figure 4.1. This behaviour has been implemented in some versions of Microsoft Word (U.S. Patent No. 5,832,528, 1998, claim 7).

Individual tasks were designed to be either *congruent* or *incongruent* with the word-by-word snapping behaviour. When congruent, the target selection was aligned with word boundaries, which allowed the snapping behaviour to assist selection (effectively enlarging the target size of the selection's start and end points). When incongruent, the target selection was misaligned with word boundaries, which impeded task completion due to the requirement to disable snapping by backtracking to the initial word.

The experimental method manipulated the overall marginal utility (Equation 3.5) of the word-by-word technique by controlling the number of congruent and incongruent tasks within the set of tasks completed by each subject. When all tasks were congruent, the marginal utility of word-by-word was positive with respect to the letter-by-letter reference condition, and subjects were hypothesised to prefer it. When all tasks were incongruent, there was a corresponding negative marginal utility, and subjects were hypothesised to reject it. However, when sets included a mixture of congruent and incongruent tasks, subjects were hypothesised to exhibit a bias against word-by-word due to the shape of μ – even when there were overall objective gains.

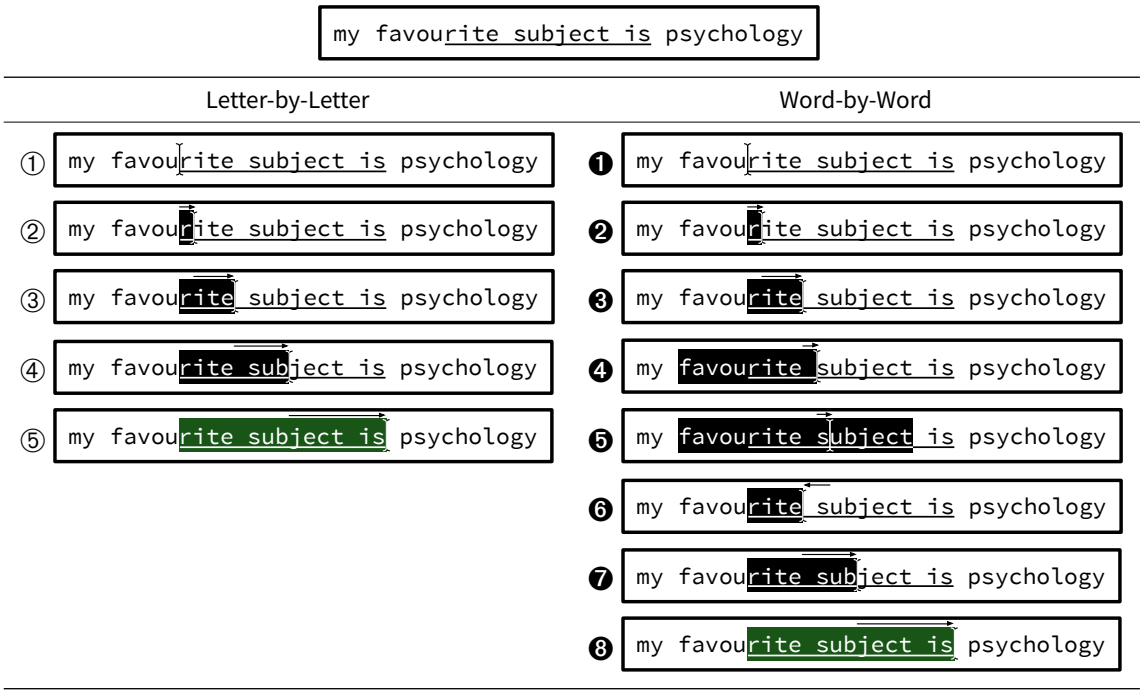


FIGURE 4.1 Examples of the letter-by-letter and word-by-word techniques in Experiment 4.1: tasks involved selecting the underlined portion of text, shown at the top. The letter-by-letter technique (left) allowed text to be selected between any two characters, one letter at a time (①–⑤). The word-by-word technique (right) began identically to letter-by-letter (❶–❸), but the selection snapped to word boundaries when a space character was crossed (❹ and ❺). In this example the snapping is not desired and is disabled by backtracking the selection to any point within the first selected word (❻); subsequent selection follows letter-by-letter (❼–❽).

4.1.1 Modelling Preferences

With letter-by-letter selection, all tasks were completed in the same manner and were equally difficult (given a fixed number of characters to select and a monospace font). During ideal letter-by-letter operation, dragging towards the end of the selection region continually advanced towards the goal, progressively increasing the number of correctly selected characters. That is, the subject's actions a moving the cursor rightwards (Figure 4.1) produced commensurate progress towards selecting the required characters; $\forall a \in I : m_k(c_k) = m_k(r_k(a))$, for the dimension k of character selection. The reference expectations in this case came from subjects' prior experience (as it emulates existing text selection practice) and what could be reasonably expected for such direct-manipulation actions.

However, word-by-word task difficulty hinged on whether the target selection was aligned with word boundaries (its *congruency* – illustrated in Figure 4.2). As the letter-by-letter and word-by-word techniques were being compared against each other, congruent tasks were intended to have a higher utility with word-by-word selection than letter-by-letter selection, and incongruent tasks were intended to have a lower utility with word-by-word selection than letter-by-letter selection.

The benefit of congruent tasks for word-by-word selection came from the positive marginal utility returned for a subject's pointing actions. A subject's reference expectation[1] was that exact pointing was required, but snapping to word boundaries allowed pointing actions to be less precise, and task completion was therefore faster. Such progress therefore exceeded letter-by-letter reference expectations, had positive marginal utility, and a higher overall utility than that for letter-by-letter: $\forall a \in I : m_k(c_k) \geq m_k(r_k(a))$.

1. i.e. letter-by-letter.

For incongruent tasks, the word-by-word selection interface yielded a lower utility than for letter-by-letter selection. When the snapping behaviour was misaligned with the target selection, additional actions were required to disable the behaviour (i.e. ❺–❻ in Figure 4.1). These actions were not present in the letter-by-letter method and required extra time and effort to perform. In addition, their outcome appeared to *remove* progress previously acquired. That is, the letters selected in steps ❹ and ❺ were deselected and reselected in steps ❻ and ❼. These actions to manipulate the snapping behaviour required negative progress away from the goal state.[2] These elements gave a sensation of progress that was negative (moving backwards before moving forwards): $\exists a \in I : m_k(c_k) < 0$. This element had an amplified effect due to the saliency of the dimension k and the loss aversion manifest in μ.

2. In actuality, progress was always positive overall as these actions were necessary as part of the method to reach the goal; the losses were perceptual.

However, the loss of progress for incongruent word-by-word tasks was not isolated from the additional actions that these tasks required (compared to letter-by-letter). Therefore, any preference for the letter-by-letter (or, bias against the word-by-word technique) technique might be explained by either (a) strongly felt progress losses (the hypothesis), or (b) the presence of additional actions. These issues are revisited in Section 4.1.7, and Experiment 4.2 isolated them using a technique that contained similar additional actions but without the perceived progress losses.

4.1.2 *Measuring Preferences and Hypothesis*

The primary dependent measure was subjects' response to whether or not they chose to repeat an experimental set of mixed congruent and incongruent tasks using the letter-by-letter selection technique or the word-by-word selection technique. An affirmative response for word-by-word indicated that

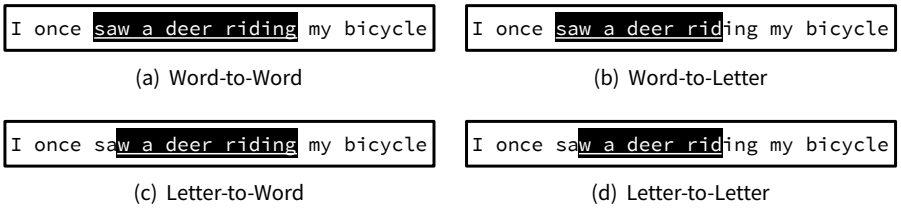


FIGURE 4.2
Examples of the different types of selection task. Only (a) is congruent for word-by-word selection.

they felt its behaviour was of higher utility than letter-by-letter, while a negative response indicated it was of lower utility.

The primary hypothesis was that the losses associated with incongruent tasks have subjectively more impact (i.e. fewer subjects choose to repeat them) than their objective performance measures would imply. That is, subjects would have a tendency to *reject* word-by-word selection when the set of tasks contained one or more incongruent tasks, even when combined with congruent tasks that more than compensated for the objective performance losses incurred. If confirmed, this supports the presence of aversion to the perceived loss of progress (and additional actions).

4.1.3 *Apparatus & Participants*

The experiment ran on Intel Core i7 computers running Linux Mint 17, rendering to a 22″ LCD monitor running at a resolution of 1680×1080 px, with input received through a wired Logitech optical mouse. The X server was configured to use a polynomial pointer acceleration profile, with a constant deceleration factor of 4. The software was written in Python and logged all user actions and responses.

Ninety volunteer undergraduate computer-science students took part in the experiment (21 female). Participation lasted approximately 10 minutes.

4.1.4 *Procedure*

To ensure familiarity with the two selection techniques, subjects first completed 5 letter-by-letter practice tasks and 10 word-by-word practice tasks. The 10 word-by-word tasks exposed subjects to all selection possibilities (5 congruent and 5 incongruent).

Subjects then completed three sets of 10 experimental selection tasks. In the first two sets, subjects used letter-by-letter and word-by-word techniques in a counterbalanced order (even numbered subjects used letter-by-letter first, odd numbered used word-by-word first). Subjects then selected which of the two techniques they wanted to use for the third set of 10 selections, which was subsequently completed using that choice (illustrated in Figure 4.3).

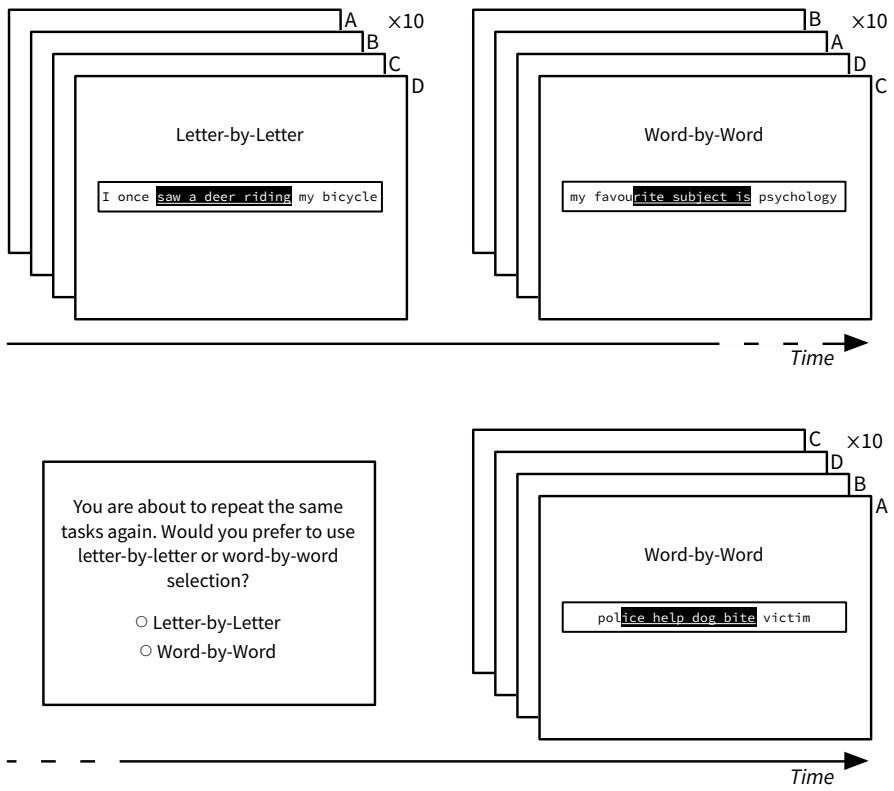


FIGURE 4.3 A schematic of the Experiment 4.1 procedure (following practise tasks). The first two sets of tasks (top row) were counterbalanced, and the final set (bottom-right) used the mode chosen by the subject on the preceeding screen (bottom-left). The order of tasks within each set (A, B, C, D) was randomised.

For each subject, the experimental tasks were identical in all three sets (the same sentences and selection ranges), but were administered each time in a random order. After completing the first two sets, they were asked:

> You are about to repeat the same tasks again. Would you prefer to use letter-by-letter or word-by-word selection?
> (Forced-choice radio button selection: 'Letter-by-Letter' or 'Word-by-Word'.)

For simplicity, this choice was coded as either accepting or rejecting word-by-word. Subjects were informed of this procedure (the two sets of identical tasks, the decision, and the final set of tasks using their choice).

Each task presented the sentence to be selected in a 880 × 66 px box, positioned in the centre of the screen. Subjects pressed a *Start* button directly beneath it to reveal the portion to be selected and begin the task. The text was rendered in a black monospace font (white when selected), which gave

each character a bounding box of 22 × 44 px. The selection highlight was normally black, but turned dark green when the selected region exactly matched the target (⑤ and ❽ of Figure 4.1). Releasing the mouse button completed the selection. If the selection matched the target, the next sentence was automatically cued (with the *Start* button reset), otherwise the sentence remained and subjects had to repeat the selection until it was correctly selected.

Selection could be performed in either direction. A message above the text box and before each set of tasks informed subjects of the initial selection mode ('Letter-by-Letter' or 'Word-by-Word').

4.1.5 Design

The sets of ten tasks were constructed to control the number of tasks that were congruent with word-by-word selection (i.e. aligned with word boundaries); the remaining tasks in each set were incongruent (i.e. misaligned with word boundaries) – see Figure 4.2. Six different selection set configurations were used, with {0, 2, 4, 6, 8, or 10} congruent tasks within each set (coded as the proportion of congruent tasks).

These six conditions evenly sampled the complete range of possible set constructions and therefore controlled the amount of intended gain and loss experienced. The conditions with 0 and 10 congruent tasks served as control conditions as they did not require subjects to respond to a trade-off between the two types of task – subjects were expected to reject the word-by-word interface when there were 0 congruent tasks (as there was never any benefit to using it), and they were expected to accept word-by-word when there were 10 congruent tasks (as there was never any cost).

Each subject completed only one of the conditions (between-subjects; in a counter-balanced order of 15 subjects per condition). This reduced the effects of practice and learning (which were not under investigation), and preserved subject sensitivity that may be confounded if they were exposed to repeated sets that were highly similar (see §3.3.3).

The tasks were randomly constructed from the MacKenzie and Soukoreff (2003) phrase set: each task was 15 characters long (±2 characters for those that needed alignment with word boundaries), with start and end points in words that were 6±2 characters in length. This ensured that the pointing difficulty was roughly equal across tasks.[3] No subject was exposed to the same sentence more than once per set.

3 These tasks can also be considered under a Fitts' law paradigm (Fitts, 1954): letter-by-letter tasks had an ID of approximately 4.70–5.09 bits, and congruent word-by-word tasks had an ID of approximately 1.70–3.09 bits. However, the additional cognitive and physical actions required to recognise and disable word-by-word snapping perverts such an analysis for incongruent tasks.

Figure 4.2(b–d) shows three types of incongruent tasks that can be constructed by placing the misaligned boundary at the start, end, or both positions of the selection. As the placement of the misaligned boundary may affect performance,[4] and pilot subjects were observed to consistently select from left-to-right, incongruent tasks were always misaligned at their leftward edge (Figure 4.2(c) and (d)).

4. e.g. with an incongruent task, a subject may not notice the problem until late in the task.

Each task was timed from when the *Start* button was pressed until successful completion. To avoid the influence of outliers, the geometric mean of a set's constituent task times was used as the measure of performance.[5] The time difference between letter-by-letter and word-by-word sets (prior to a subject choosing between them) was the objective measure of performance gain or loss that a subject received when using word-by-word.

5. Using the arithmetic mean does not significantly affect the results, but does increase the between-subjects variance.

4.1.6 Results

Time Performance. Analysis of selection time data showed that the manipulation of congruent and incongruent tasks had the intended effect on objective performance gains and losses – shown in Figure 4.4. Mean selection time[6] for each letter-by-letter task was 2.23 seconds (95% CI [2.13, 2.33]), with word-by-word ~30% faster for congruent tasks (1.61 s, 95% CI [1.53, 1.69], $t(162.94) = 10.59$, $p < .001$, $d = 1.63$)[7] and ~60% slower for incongruent tasks (3.51 s, 95% CI [3.31, 3.71], $t(107.04) = -11.80$, $p < .001$, $d = 1.94$). Although unimportant for the primary hypothesis, there was a small effect of practice. Comparing task times before and after subjects made their choice, letter-by-letter and congruent word-by-word tasks were significantly faster when repeated (paired $t(149.87) = 3.47$, $p < .001$, $d = 0.54$; and $t(57.57) = 3.17$, $p = .002$, $d = 0.64$, respectively) – but not incongruent word-by-word tasks ($p = .93$).

6. Before subjects chose between snapping behaviours.

7. All t-tests herein are for unequal variances with Welch-Satterthwaite degrees-of-freedom approximations.

Overall performance with letter-by-letter tasks was consistent across conditions (mean 2.26 s, 95% CI [2.03, 2.49], one-way ANOVA $p = .27$). In contrast, word-by-word time decreased linearly as the number of congruent tasks increased (Figure 4.5(a); $R^2 = .99$, $F(5, 84) = 29.59$, $p < .001$, $\eta_g^2 = .63$). These results confirm that the word-by-word selection technique had the intended effects of losing or gaining time (relative to letter-by-letter), and outperformed letter-by-letter selection when at least 60% of the tasks were congruent.

Errors. Errors were classified in terms of the number of excess selection attempts for each task. That is, if a subject was unable to complete a selection with their first attempt and had to try again (e.g. their start location was incorrect), each additional attempt was counted as *excess*. A Kruskal-Wallis test between letter-by-letter (mean 0.16 excess attempts, 95% CI [0.13, 0.18]), word-by-word for congruent tasks (0.13, 95% CI [0.09, 0.16]), and word-by-word for incongruent tasks (0.22, 95% CI [0.15, 0.26]), revealed a significant

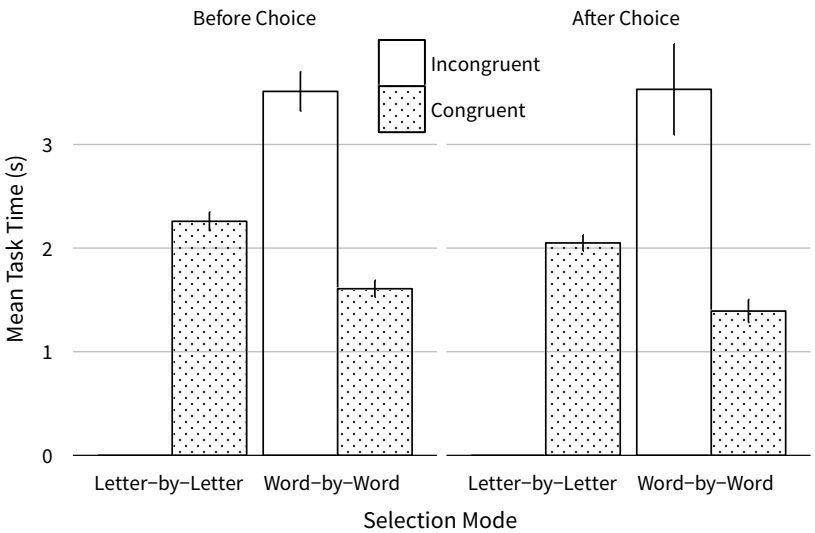
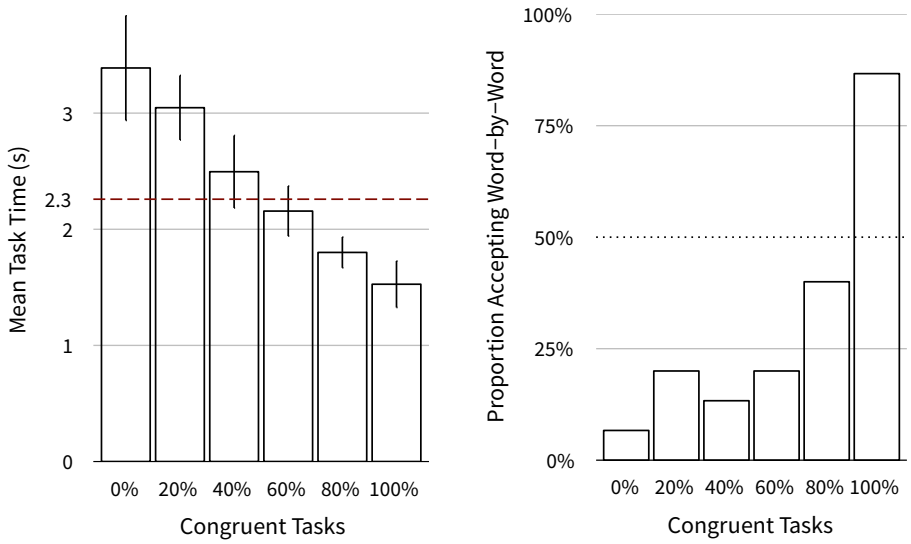


FIGURE 4.4 The mean time subjects took to complete each type of task in Experiment 4.1, before (left) and after (right) they made their choice (error bars show 95% confidence intervals).



(a) The mean time to complete each word-by-word task by condition. The dashed line shows the mean letter-by-letter task time across all conditions (error bars show 95% confidence intervals).

(b) The proportion of subjects that accepted the word-by-word snapping technique in each condition.

FIGURE 4.5 Experiment 4.1 time and choice results by condition.

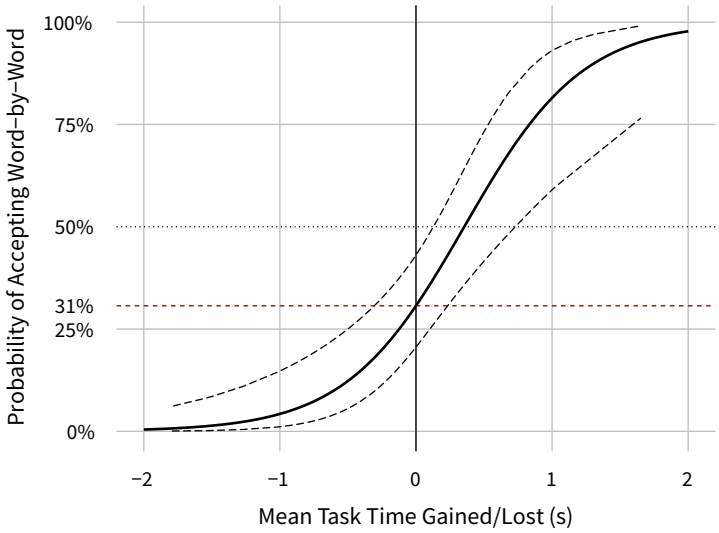


FIGURE 4.6 A binomial logistic regression model (dashed ±95% confidence interval band) for subject choice by the actual amount of time lost or gained in Experiment 4.1. The intersection with the vertical axis – when no time was lost or gained – is highlighted (dashed, maroon).

main effect ($\chi^2(2) = 6.68$, $p = .04$). Bonferroni-corrected pairwise comparisons using Dunn's test (Dunn, 1964) only found a significant difference between the two word-by-word task types ($z = -2.49$, $p = .02$).

Choice Response. In the control conditions with either 0 or 10 congruent tasks, all subjects received the expected loss or gain, respectively. When no tasks were congruent, all but one subject (14 of 15) rejected the word-by-word technique; conversely, all but two (13 of 15) accepted the word-by-word technique when all tasks were congruent – indicating that subjects were successful in identifying the impediment or benefit of the word-by-word selection technique. Figure 4.5(b) shows the proportion of subjects accepting word-by-word for each condition.

To examine the primary hypothesis – that losses have subjectively more impact on utility – subjects' decisions were analysed with respect to the objective gain or loss received. Figure 4.6 shows a binomial logistic model (§3.3.4) of the probability that a subject would accept the word-by-word technique given the mean time per task gained or lost using word-by-word ($\chi^2(1) = 25.31$, $p < .001$; Odds-Ratio = 9.94, 95% CI $[3.60, 34.79]$). The intercept of this model indicates the probability that a subject would have accepted the word-by-word technique over the letter-by-letter technique when the time difference between the two techniques was zero (that is, no objective gain or loss). This probability is 30.68% (95% CI $[19.30\%, 42.07\%]$), which is significantly less

than a null model of indifference (50%; binomial $p < .001$) – indicating a bias against the word-by-word technique.

This bias is also indicated in the contrast between Figures 4.5(a) and (b). The linear time reduction across the proportion of congruent tasks shown in Figure 4.5(a) is *not* reflected by a linear increase in the proportion of subjects accepting word-by-word in 4.5(b). Rather, Figure 4.5(b) shows that the majority of subjects rejected word-by-word across conditions (except for the 100% congruency condition in which subjects did not encounter any trade-off).

4.1.7 Discussion

The results of the logistic regression model (Figure 4.6) indicate that if objective performance outcomes were balanced (the zero point along the abscissa), a significant majority of subjects (69%) would reject the word-by-word technique. The word-by-word technique would need to exhibit a benefit of at least several seconds before subjects would be indifferent to its behaviour (the model in Figure 4.6 intercepts the 50% indifference point at ≈300 ms per task, representing a total of 3 s for the 10 tasks in a set). The primary hypothesis attributes this bias against the word-by-word technique to an overweighted aversion against the negative progress of incongruent tasks. However, there are several possible explanations for this bias, including the following:

- *Aversion to progress losses*. The word-by-word technique requires backtracking that perceptually removes previously completed progress towards the goal state. This removal of progress may have been subjectively overweighted as negative (the primary hypothesis).
- *Aversion to additional actions/workload*. The actions required to recognise and disable the snapping behaviour may contain extra cognitive costs that increased the workload of the technique beyond that captured in the measure of time. In other words, subjects may have disliked the actions and work required by the word-by-word technique more than they liked its overall time savings.
- *Aversion to time losses*. Incongruent word-by-word selection tasks were slower than neutral letter-by-letter selection, and subjects may have responded negatively to these time losses (rather than the progress, workload, or cognitive losses).
- *Speculative assessment*. Rather than responding to their experienced task sets, subjects may have chosen between techniques based on their speculation of what the word-by-word technique would be like for real use (rather than the task at hand).
- *Familiarity preference*. Subjects may have chosen letter-by-letter over word-by-word based on their familiarity with the technique.

Distinguishing between these alternatives requires manipulating one candidate explanation while holding all others constant. The next experiment addresses this by using identical tasks, but altering the mechanism used to disable snapping with the word-by-word technique. The key difference is that the modified technique preserves progress by using a key-press and time-consuming animation to disable snapping, rather than by requiring the user to backtrack the cursor and remove previously selected characters.

4.2 EXPERIMENT 4.2: NEUTRALISING LOSS AVERSION

To examine whether the preference choices observed in Experiment 4.1 stem predominantly from an aversion to progress losses or to one of the other explanations in Section 4.1.7, a second experiment was undertaken with a small change to the design of the word-by-word technique. The revised word-by-word technique was designed to require similar additional actions and time costs when compared with letter-by-letter, but removed the need to backtrack to turn off the snapping behaviour.

The behaviour of the letter-by-letter selection technique and the word-by-word selection technique for congruent tasks was identical to Experiment 4.1, but the method for disabling the snapping behaviour for incongruent tasks was modified – illustrated in Figure 4.7.

To disable snapping, subjects tapped on the *Control* key, which began an animation that gradually shrank the selection (at 80 ms per character) until it matched the letter-by-letter selection region (i.e. as if the task had been started with letter-by-letter selection). A similar snapping behaviour (without the animation) has also been implemented in some versions of Microsoft Word (U.S. Patent No. 5,832,528, 1998, claim 5). This mechanism preserved all characters that had already been selected, and did not require backtracking of the cursor: $\nexists a \in I : m_k(c_k) < 0$, for the dimension k of character selection.[8] The animation during the mode switch (steps ⑦–⑨ in Figure 4.7) was controlled to ensure that the objective time cost of the technique was at least as large as that of backtracking in Experiment 4.1.

8. Depending somewhat on where in the task the behaviour was disabled – see Figure 4.7.

While the key-press action was not *identical* to the backtracking action, the other elements of a subject's actions were largely comparable between both experiments:

- Both techniques involved a cognitive component in recognising a need to disable snapping.
- The time costs of performing the key-press and animation were engineered to match the time costs of backward pointing in Experiment 4.1.
- Both techniques required forward pointing to complete the selection once snapping was disabled.

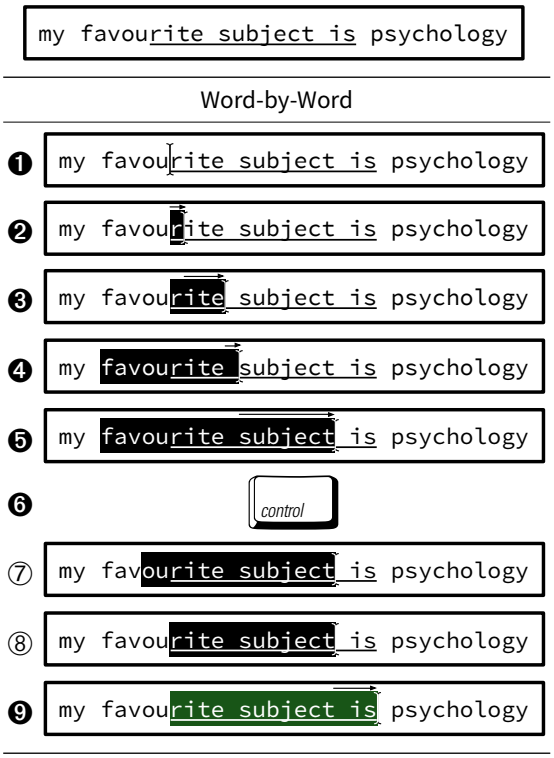


FIGURE 4.7 An example of the revised word-by-word selection technique in Experiment 4.2. The letter-by-letter technique behaved as in Figure 4.1 – as do steps ❶–❺ of the word-by-word technique. However, to disable the snapping behaviour, the *Control* key is tapped (❻), which initiates an animation that shrinks the selection to its letter-by-letter equivalent (⑦–⑧); subsequent selection follows letter-by-letter (❾).

- The key-press action required coordinating input across two hands, but freezing the selection region during the animation prohibited subjects from fully exploiting parallel actions.

In summary, there is relatively little reason for hypothesising that an additional small pointing action would be substantially less preferable than a key-press and frozen wait. These issues are further discussed in Section 4.3.1.

4.2.1 *Apparatus & Participants*

The experimental apparatus was identical to Experiment 4.1 (§4.1.3). Sixty-four volunteer undergraduate computer-science students participated in the experiment (14 female),[9] which lasted approximately 10 minutes.

9. None had participated in Experiment 4.1.

4.2.2 *Procedure & Design*

The procedure was identical to Experiment 4.1 (§4.1.4), but the experimental design was altered slightly. To increase experimental power and subject

economy, four selection sets were used (instead of six), and to evenly sample the range of possible set constructions, the number of tasks in each set was increased to 12 (instead of 10). The four selection sets varied in the number of congruent tasks, and contained either {0, 4, 8, or 12} congruent tasks. Sixteen subjects were assigned to each of the four selection sets (between-subjects, as with Experiment 4.1).

The selection tasks were constructed as with Experiment 4.1 (§4.1.5), but with the additional constraint that incongruent tasks either started or ended exactly five characters from the opposing word boundary. As the animation that shrank the selection progressed at one character every 80 ms, this ensured that disabling snapping took at least 400 ms, which pilot studies indicated would maintain the overall level of objective time loss observed in Experiment 4.1's incongruent tasks.

To prevent subjects from prematurely disabling the word-snapping behaviour (e.g. pressing the *Control* key at the start of the selection), the feature only worked after at least one word boundary had been crossed (i.e. word-snapping had begun). During the animation, the selection region was not manipulatable until the animation completed, after which the technique continued with letter-by-letter selection.

Practice tasks (5 with letter-by-letter, 10 with word-by-word) were identical to Experiment 4.1, as was the presentation of the tasks and time recording.

4.2.3 *Results*

Time Performance. Similar to the results of Experiment 4.1 (§4.1.6), the mean selection time for letter-by-letter tasks was 2.08 s (95% CI [1.98, 2.18]), with congruent word-by-word tasks ~20% faster (1.63 s, 95% CI [1.50, 1.76], $t(76.96) = 6.77$, $p < .001$, $d = 1.37$), and incongruent word-by-word tasks ~60% slower (3.33 s, 95% CI [3.12, 3.54], $t(58.00) = -10.96$, $p < .001$, $d = 2.32$) – shown in Figure 4.8. There was a small effect of practice for all three—

- Letter-by-letter: $t(83.90) = 2.71$, $p < .01$, $d = 0.53$.
- Congruent word-by-word: $t(71.90) = 3.14$, $p < .01$, $d = 0.68$.
- Incongruent word-by-word: $t(41.38) = 2.22$, $p = .03$, $d = 0.54$.

There was no effect of congruency level on letter-by-letter performance (mean 2.11 s, 95% CI [1.96, 2.26], one-way ANOVA $p = .81$), but word-by-word performance increased linearly as congruency increased ($R^2 = .99$; $F(3, 60) = 57.76$, $p < .001$, $\eta_g^2 = .74$) – shown in Figure 4.9(a).

Errors. For the number of excess selection attempts, a Kruskal-Wallis test between letter-by-letter (mean 0.14 excess attempts, 95% CI [0.10, 0.18]), word-by-word for congruent tasks (0.10, 95% CI [0.07, 0.13]), and word-by-word for incongruent tasks (0.26, 95% CI [0.18, 0.33]), revealed a significant

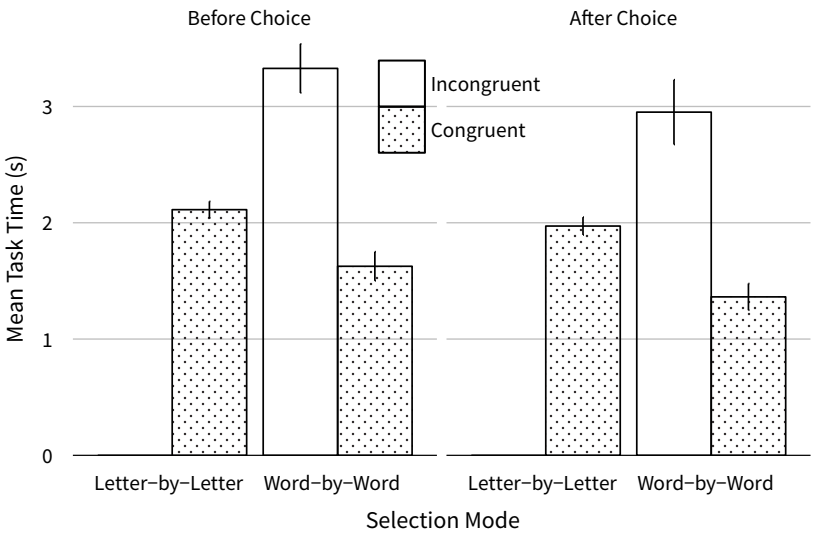
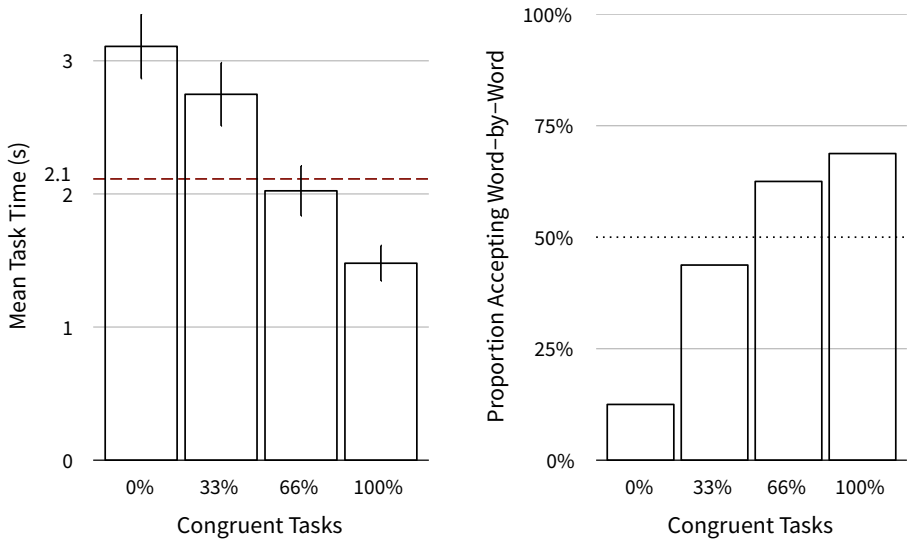


FIGURE 4.8 The mean time subjects took to complete each type of task in Experiment 4.2, before (left) and after (right) subjects made their choice (error bars show 95% confidence intervals).



(a) The mean time to complete each word-by-word task by condition. The dashed line shows the mean letter-by-letter task time across all conditions (error bars show 95% confidence intervals).

(b) The proportion of subjects that accepted the word-by-word snapping technique in each condition.

FIGURE 4.9 Experiment 4.2 time and choice results by condition.

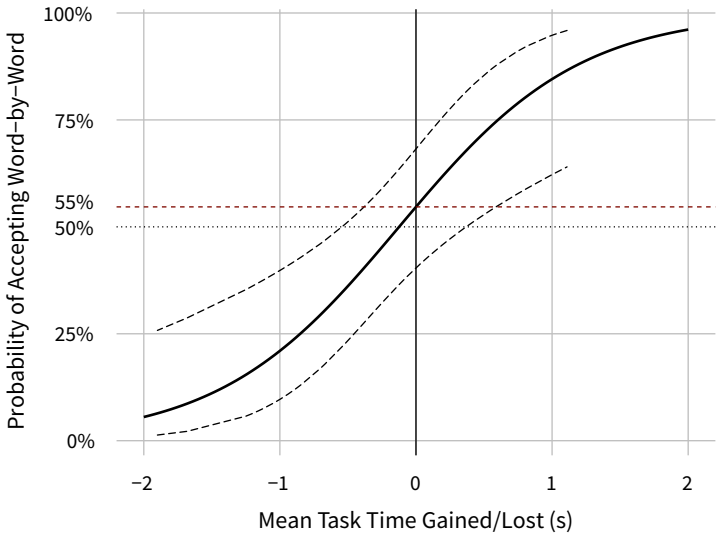


FIGURE 4.10 A binomial logistic regression model (dashed ±95% confidence interval band) for subject choice by the actual amount of time lost or gained in Experiment 4.2. The intersection with the vertical axis – when no time was lost or gained – is highlighted (dashed, maroon).

main effect ($\chi^2(2) = 13.34$, $p < .001$). With significant pairwise differences between letter-by-letter and incongruent word-by-word ($z = 3.08$, $p < .01$), and the two word-by-word task types ($z = 3.45$, $p < .001$).

Choice Response. In the control conditions, all but two subjects (14 of 16) rejected word-by-word when all tasks were incongruent, and all but five (11 of 16) accepted word-by-word at the opposite end. Figure 4.9(b) shows the proportion of subjects accepting word-by-word for each condition.

A binomial logistic regression for subject choice across the time gained or lost is shown in Figure 4.10 ($\chi^2(1) = 14.12$, $p < .001$; Odds-Ratio = 4.54, 95% CI $[1.98, 12.15]$). The probability that a subject would choose word-by-word when the time difference between the two techniques was zero is 54.67% (95% CI $[40.38\%, 68.95\%]$) – which is not significantly different from indifference (binomial $p = .73$). This indicates that subjective preferences were predominantly aligned with the objective outcomes of the two interfaces. This conclusion is also reflected in the comparison between Figures 4.9(a) and (b): the proportion accepting word-by-word roughly reflects the time gained or lost in the associated condition. In other words, there is no indication of a bias for or against the word-by-word technique.

These results indicate that if objective outcomes were balanced – that is, letter-by-letter and word-by-word had no difference in time performance – subjects would not express a preference for one technique over the other.

4.3 DISCUSSION

The main finding is in the contrast between results of Experiments 4.1 and 4.2. In Experiment 4.1, a significant majority of subjects rejected word-by-word snapping, even when it improved their task performance – that is, subjects exhibited a bias against the technique, which was hypothesised to be an aversion to the progress losses required to disable the snapping behaviour (backtracking). Experiment 4.2 showed that this bias was neutralised when subjects used a variant technique that replaced backtracking progress losses with a similarly time-consuming action that preserved progress.

However, the difference between a significant effect (Experiment 4.1) and a non-significant effect (Experiment 4.2) is itself not necessarily statistically significant (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). As Experiments 4.1 and 4.2 use the same dependent variable to predict choice (mean time saved per task), the two logistic models (Figures 4.6 and 4.10) can be compared by pooling the data and using a dummy variable to encode experiment membership. The logistic regression slope for this variable shows the effect of the experimental manipulation (the difference between Experiments 4.1 and 4.2), and is significantly different from 0 (1.00, 95% CI $[0.23, 1.81]$, $p = .01$). A likelihood ratio test also shows a significant difference between the two experimental models ($\chi^2(1) = 8.01$, $p = .02$).

The only substantial difference between the interactive tasks in Experiments 4.1 and 4.2 was the mechanism used to disable word-by-word snapping. Rather than backtracking the selection, in Experiment 4.2 the mode switch was accomplished by pressing a key that gradually shrank the selection to its letter-by-letter equivalent. Other elements of the two experiments were similar: the actions required to operate the word-by-word techniques were consistent in cognitive scope (requiring subjects to identify the need to disable snapping, then execute a motor action to achieve it); and both the time required to perform the actions and the incidence of selection errors were equitable (shown in Table 4.1).

With respect to the model of reference-dependent preferences (§3.2), the difference between the experiments is attributed to the utility of the outcomes while disabling word-by-word during incongruent tasks: $m_k(c_k)$ in Equation 3.5, where k is the dimension of character selection progress. In Experiment 4.1, the gain-loss utility while backtracking was negative because users were required to remove previously attained progress ($m_k(c_k) < 0$, while $m_k(r_k(a)) > 0$). However, in Experiment 4.2 the same component was at least neutral: $m_k(c_k) \approx 0$. This dimension was heavily weighted because of its saliency and relationship to the goal (more-so than the time required to disable the snapping technique), and therefore had a significant impact on

TABLE 4.1 Comparison of the task time and error rate metrics from Experiments 4.1 and 4.2.

Task Time	Letter-by-Letter	2.23 s [2.13, 2.33] 2.08 s [1.98, 2.18]	Error Rate	Letter-by-Letter	0.16 [0.13, 0.18] 0.14 [0.10, 0.18]
	Word-by-Word (Congruent)	1.61 s [1.53, 1.69] 1.63 s [1.50, 1.76]		Word-by-Word (Congruent)	0.13 [0.09, 0.16] 0.10 [0.07, 0.13]
	Word-by-Word (Incongruent)	3.51 s [3.31, 3.71] 3.33 s [3.12, 3.54]		Word-by-Word (Incongruent)	0.22 [0.15, 0.26] 0.26 [0.18, 0.33]

Note: Experiment 4.1 means are reported above those for Experiment 4.2 (both with 95% confidence intervals in brackets). Task times are the mean for a single selection before subjects chose between snapping behaviours, and *error rate* measures the mean number of excess selection attempts per task.

overall utility U. In both experiments, other elements of the interaction were substantially constant: the subjects' actions a, and therefore the utility of their expectations $m_k(r_k(a))$ from letter-by-letter selection, were consistent; as were errors, time costs, and aesthetics.

4.3.1 *Experimental and Modelling Issues*

Experimental work in human–computer interaction typically focuses on objective performance measures, such as time and errors. These measures are readily obtained, amenable to inferential statistics, easily replicated, and essential in contexts where there are performance requirements. However, the user's experience of interactive systems is also a critical design consideration. When this is the case, the forced-choice methodology provides a pragmatic approach for examining subjects' overall evaluation of the utility of user interface manipulations – avoiding the issues found with more granular psychometric measures of experience (reviewed in §2.7.3). Prior research has also demonstrated that subjective outcomes need not reflect objective measures (e.g. Hornbæk, 2006), and the present experiments demonstrate conditions under which this deviation between performance and subjective outcome occurs. The reference-dependent utility model establishes a framework for separating the objective and subjective components of interaction and understanding how manipulations of objective outcomes can influence subjective preferences. The following paragraphs highlight some experimental and modelling limitations. More general conclusions and opportunities for future work are described in Chapter 7.

Consumption versus Marginal Utility. Experiments 4.1 and 4.2 provide evidence of conditions under which loss aversion occurs during interaction: progress losses via elements of negative consumption utility. However, these

experiments have not been explicitly isolated the model's separate components for *consumption utility* and *marginal utility*.

Experiment 4.1 ensured a negative input to the value function μ by requiring subjects to encounter a perceptual *progress loss* – $m_k(c_k) < 0$, for the dimension k representing the number of characters selected – as they had to backtrack. Experiment 4.2 replaced these progress losses with elements of substantially neutral progress – $m_k(c_k) \approx 0$ (no loss in characters selected). However, the model of reference-dependent preferences explains that the input to the value function μ is not consumption utility, but *marginal utility* (Equation 3.5), which is the difference between actual and expected outcomes.

In both experiments, the overall marginal utility was negative during incongruent tasks because the additional actions to disable the snapping created expectations for outcomes that exceeded those attained (i.e. the actions are not required in the letter-by-letter reference condition for the same selection task). This marginal utility was very negative for Experiment 4.1, and slightly negative for Experiment 4.2 – making it easier to recover from when a set contained other, congruent tasks.

There are opportunities for further research that separately analyses the gains and losses of actual and expected consumption utility, as well as positive and negative marginal utilities. For example, a text entry interface might assist a user by correcting erroneous typing when the user expects to manually re-enter it (positive consumption utility, low expectations, positive marginal utility); it might fail to correct the word when the user expects it to be automatically corrected (neutral consumption utility, high expectations, negative marginal utility); or it might incorrectly replace a correctly typed word (negative consumption utility, moderate expectations, negative marginal utility).

Actions. The actions required to disable snapping were slightly different between the two experiments. In both experiments there was an initial cognitive component that required subjects to identify the need to disable snapping. Then, in Experiment 4.1 subjects had to drag to re-enter the initial word, and then re-drag out to reach the terminating selection point; in Experiment 4.2 they had to press the *Control* key, wait, and then complete the dragging movement (the time taken for these actions was controlled to be similar).

Therefore, the core difference in actions between the experiments comprises a leftwards pointing action in Experiment 4.1 versus a key-press and wait in Experiment 4.2 (other differences were summarised in §4.2). It is unlikely that this small difference in actions could cause the substantial difference in response, but the fact that there were any differences in actions remains an experimental risk that can be mitigated through further experimental work. For example, a future study might replace the key-press and wait of Experiment 4.2 with a vertical pointing action that maintains progress

(this technique was not used here because it lacks ecological validity and no viable text selection technique could use this method because vertical cursor movement is reserved for multi-line selections).

Consumption Bundle Dimensions. In the reference-dependent preferences model, utility values are summed across the dimensions of outcome consumption bundles ($\mathbf{c}$ and $r(a)$). Both experiments held the dimensions of these bundles constant (with the caveats about the actions above), except for the dimension of the number of characters selected. Character selection progress was chosen as the dimension for analysis because it is among the most salient for the task, and therefore most heavily weighted (Bordalo et al., 2013) – giving the greatest experimental sensitivity. However, the model and method are also applicable to other dimensions of interface outcomes that are not directly related to visible progress – for example, the perceived value of other objective qualities relating to feedback latencies, and so on.

Number of Losses versus Magnitude of Loss. Both experiments varied the magnitude of loss by altering the *number* of congruent tasks within the task set, where each incongruent task involved the same *magnitude* of loss. However, the model predicts that interactions with different magnitudes of objective loss will generate different levels of utility, creating opportunities for further investigation. That is, eight incongruent tasks that lose one second each will, according to the model, have a different utility to a single task with an eight second loss. However, it is difficult to arbitrarily manipulate interactive losses – which often depend on human factors – in this way.

Analysing the data of Experiments 4.1 and 4.2 in terms of the proportion of congruent or incongruent tasks does not change the results as this factor is correlated with the amount of time lost (and the manipulation is equitable between experiments). However, the collinearity of these variables also means that the effect of introducing incongruent tasks cannot be isolated.

4.3.2 *Summary*

The two experiments presented in this chapter have demonstrated the presence of an aversion to perceived losses of interactive task progress (in excess of the objective performance losses), and its subsequent neutralisation by altering the perceptual output of that progress loss. These results can be understood in the context of the model presented in Chapter 3: in Experiment 4.1, a deficit was created between a subject's expectations and actual interactive outcomes, resulting in a substantially negative gain-loss utility; in Experiment 4.2 this deficit was narrowed, and preferences become more moderate. This is not a conclusive demonstration of the model's validity, but connects its key components with the foundational behavioural economics work in Chapter 2.

A POSITIVITY BIAS FOR PROGRESS 5

THE EXPERIMENTS in the previous chapter developed support for the model of reference-dependent preferences from Chapter 3: they found that perceived progress losses overwhelmed subjective evaluation, even when subjects experienced objective performance gains. The experimental tasks were either congruent or incongruent with a potentially-assistive feature[1] to stimulate certain gains or losses. However, a single incongruent task always inflicted a loss upon subjects, and the outcome of an incongruent task was always objectively negative. That is, although the experiments demonstrated an aversion to perceived progress losses under a constant objective loss, they did not examine a manipulation of the objective loss or the perception of progress gains under such losses.

This chapter describes an experiment that uses a new task to develop further support for the model by demonstrating a positivity bias for progress over objective losses.

1. Word snapping text selection.

This chapter describes a new experimental task that allowed the marginal utility of incongruent tasks to be manipulated (§5.1). That is, the gain or loss of the technique could be tested independently of the task's congruency. This removed the task's congruency as a prima facie indication of gain or loss, and allowed experimental conditions to be constructed where incongruent tasks performed objectively better than the reference condition – or objectively worse, but with perceptual gains in progress. Table 5.1 compares the task manipulations for the two experiments in the previous chapter with the experiment presented below using this new task. The task is based around simple two-dimensional drag-and-drop actions with a potentially-assistive snap-to-grid behaviour. This task is substantially similar to the text selection tasks in Experiments 4.1 and 4.2: both involved simple pointing actions that emulated existing interface behaviour.

Experiment 5.1 examined the effects of these new manipulations on subjective preferences using a methodology similar to that of Experiments 4.1 and 4.2. However, unlike those experiments, subjects were exposed to incongruent tasks that contained gains in progress with a variable amount of objective performance (with respect to the neutral reference condition). Subjects were hypothesised to exhibit a bias in favour of the experimental technique due to the salience of the progress gains.

5.1 EXPERIMENTAL MANIPULATION

The text selection task used in Chapter 4 was successful in creating conditions where subjects either gained or lost time. However, the magnitude of

TABLE 5.1 Summary of the Experiment 4.1–5.1 task manipulations for perceived progress and objective time.

	Congruent Tasks		Incongruent Tasks	
Experiment	Progress	Time	Progress	Time
4.1	+	+	−	−
4.2	+	+	::	−
5.1	+	+	+/::	+/::/−

Note: The +/::/− symbols denote intended gain (i.e. faster progress or less time), proportionate, and loss outcomes (compared to the neutral reference condition), respectively.

these objective gains and losses could not be easily manipulated: for example, incongruent word-by-word tasks forced subjects to revert back to letter-by-letter selection, which set an upper-limit for task performance (i.e. it was not possible to create incongruent word-by-word tasks with objective performance above what could be achieved using letter-by-letter selection). Rather, congruent and incongruent tasks were mixed in a set to create a particular level of overall objective gain or loss. Similarly, incongruent tasks always appeared to be negative (4.1) or neutral (4.2) with respect to progress as subjects reverted to letter-by-letter behaviour. That is, incongruent tasks never offered a progress advantage and always had a performance disadvantage.

To remove these limitations, and explore the effects of manipulated gains and losses for either congruent or incongruent tasks, a *grid snapping* task was developed wherein subjects performed drag-and-drop tasks under controlled levels of grid snapping assistance or impediment. As with text selection, these tasks used simple mouse and keyboard interactions, and leveraged an ecologically valid interaction that was already familiar to subjects. However, the technique's design allowed the both objective time and perceived progress gained or lost to be experimentally controlled.

The basic experimental method had subjects complete a series of tasks with no grid snapping assistance to establish a reference point for *neutral* performance, followed by a series of visually identical tasks with some manipulated *snapping* behaviour.

5.1.1 *Task Design*

Each task involved dragging a black square into a hollow square of exactly the same size. The square turned blue as it was dragged and green once its boundaries exactly matched the target. During neutral tasks the square moved directly according to mouse control (without any snapping), and during snapping tasks the object initially moved between locations aligned with an invisible grid of some resolution R_1 (i.e. to the screen pixel location closest

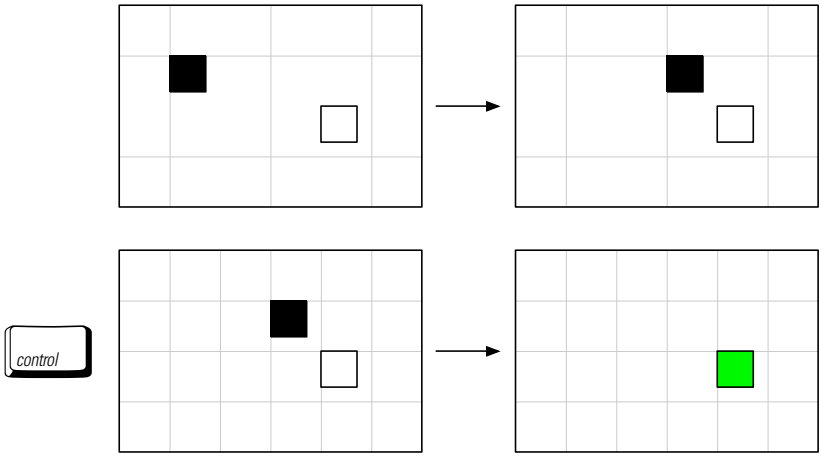


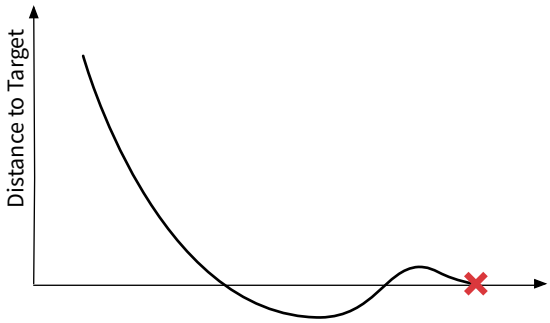
FIGURE 5.1 An example of a grid snapping task with illustrative grid lines (not shown to subjects). The object (solid, black) cannot be aligned with the target (hollow, white) in the initial grid resolution (top row), and so the object must be released and reacquired with the *Control* key held to increase the resolution such that it can be aligned (bottom row). The top row corresponds to segment ❶ in Figure 5.2(b), and the bottom row corresponds to segments ❷ and ❸.

to the cursor that was a multiple of R_1). In *congruent* conditions the target location was aligned with this grid and the task could be completed quickly;[2] however, in *incongruent* conditions the final target location was purposely misaligned with the grid – subjects had to release the object and reacquire it with the *Control* key held. This action increased the resolution of the grid to R_2 – chosen such that the target would be aligned with the grid, allowing task completion (illustrated in Figure 5.1).[3]

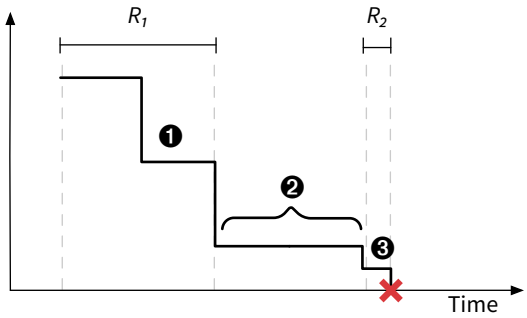
2. Akin to congruent word-by-word snapping.

These actions and behaviours are both conceptually and physically similar to those of the selection technique in Experiment 4.2. Both tasks are fundamentally a dragging action towards a target location, with a physical key-press to adjust the resolution of the object–target alignment: in Experiment 4.2 a switch from word-by-word to letter-by-letter, and here a switch from R_1 to R_2. The key differences are: (a) Experiment 4.2 used textual stimuli, whereas this task uses geometric shapes; (b) Experiment 4.2 involved a one-dimensional pointing task, whereas this task is two-dimensional; and (c) Experiment 4.2 maintained the dragging action throughout (with a forced wait period), whereas this task uses a release-and-reacquire action.

3 Under a Fitts' law paradigm (Fitts, 1954; MacKenzie, 1992), congruent tasks have an ID of $\log_2(D/R_1 + 1)$. Incongruent tasks can cover a distance of $D^* = \text{nint}(D/R_1) \cdot R_1$ using the R_1 grid, with the balance under R_2: $\log_2(D^*/R_1+1)+\log_2([D-D^*]/R_2+1)$; however, as with Experiments 4.1 and 4.2, the release-and-reacquire action jeopardises such an analysis.



(a) *Neutral*: Smooth and continuous drag-and-drop.



(b) *Snapping*: Grid snapping along ❶ at a resolution of R_1; if the resolution is not aligned with the target, the subject releases and reacquires during ❷, and completes dragging along ❸ at a resolution of R_2.

FIGURE 5.2
Characterisation of grid snapping task progress.

5.1.2 *Modelling Preferences*

Progress towards the target in each type of task is characterised in Figure 5.2. With neutral tasks (5.2(a)), subjects could rapidly reduce the distance to the target (potentially overshooting), and then precisely align the object with the target (a typical aimed movement: Meyer, Abrams, Kornblum, Wright, & Smith, 1988; Woodworth, 1899). With snapping tasks (5.2(b)), the task contained three elements (with respect to neutral behaviour):

❶ Target approach assistance through coarse snapping.

❷ An element of impediment in requiring the object to be released and reacquired to change the grid resolution.

❸ An element of assistance when the terminating resolution is larger than that of the reference neutral task.

By controlling the resolution of the grid during approach (R_1; ❶) and final positioning (R_2; ❸), the amount of overall assistance and impediment from the grid could be controlled. The release-and-reacquire component (❷) was a constant impediment.

In all cases, progress towards the goal was preserved: the dragged object always moved towards the target, and switching from the R_1 to R_2 grid maintained its location (the object never moved away from the target unless the user dragged it so – as with Experiment 4.2). However, the *rate* of progress was manipulated through the R_1 and R_2 resolutions. For instance, if the R_2 resolution was greater than that of a neutral task (i.e. there was some element of snapping during ❸), then the achievable progress and performance after the release-and-reacquire action could exceed that for a neutral task. This allowed experimental conditions to be created that had objective performance benefits using incongruent tasks – without mixing them in a series with congruent tasks (as was necessary in Experiments 4.1 and 4.2).

In the context of the model presented in Chapter 3, there was always neutral or positive progress during any snapping task: $\nexists a \in I : m_k(c_k) < 0$, for the dimension k of the distance to the target (the ordinate in Figure 5.2). Utility losses were created with the release-and-reacquire action, which was not required during neutral tasks ($m_k(c_k) < m_k(r_k(a))$), and could now be balanced with gains in the R_1 or R_2 resolution.

5.1.3 *Grid Resolution Selection*

To achieve the above utility relationships, R_1 must be a multiple of R_2 (i.e. $R_1 = \gamma R_2$, where $\gamma, R_2 \in \mathbb{Z}^+$), and R_2 must be a multiple of the target distance.[4] The grid resolution creates an effective width for the target: even if the visual target is smaller (or larger) than the grid spacing,[5] the cursor only needs to be within an R_2-sized square centred around the grid intersection for the object to become aligned with the target. If the dragged object and target location are not aligned along a cardinal axis, the grid resolution needs to be adjusted to maintain the same effective width (for a given fixed target distance). For example, along an ordinal axis, R_1 and R_2 are scaled by a factor of $1/\sqrt{2}$ (rounded to an integral number of pixels).

4. If $\gamma = 1$, or R_1 is a multiple of the task distance, then the task is congruent.

5. e.g. the bottom row of Figure 5.1

These requirements restrict the number of possible R_1 and R_2 combinations that create completable tasks (particularly if the target distance is held constant). As such, the following experiment used a single task distance (252 px) that was large enough to create a meaningful task and had many divisors to create viable grids.

5.2 EXPERIMENT 5.1: POSITIVITY BIAS

An experiment was conducted to examine the effects of variable gains and losses under a constant task type (congruent or incongruent) on subjective choices. The experiment used the task described above to create conditions

of gain or loss, but unlike the previous experiments, all tasks within a condition were either congruent or incongruent and the grid resolution was manipulated to create the overall gain/loss in time or progress. As with Experiment 4.2, progress towards the target was never negative in the experimental conditions and therefore subjects were not expected to display a bias against the snapping behaviour. Furthermore, subjects could not use the congruency of a task per se to determine whether they gained or lost time from the snapping behaviour.

The method involved a set of 10 tasks in a *neutral* condition that offered no assistance (Figure 5.2(a)), followed by a set of 10 tasks using a *snapping* interface that was engineered to include elements of impediment while also delivering variable levels of assistance (Figure 5.2(b)). After the two sets of tasks, subjects were asked to choose whether they would enable or disable the snapping behaviour if they were to repeat the set of tasks. That is, whether or not the trade-off between the assisting and impeding elements of the snapping behaviour was acceptable to them.

The primary dependent measure was each subjects' choice of whether they would use the neutral or snapping behaviour if repeating the tasks. An affirmative response to use the snapping behaviour indicated that it was subjectively valued higher than neutral; a negative response indicated the opposite.

5.2.1 *Apparatus & Participants*

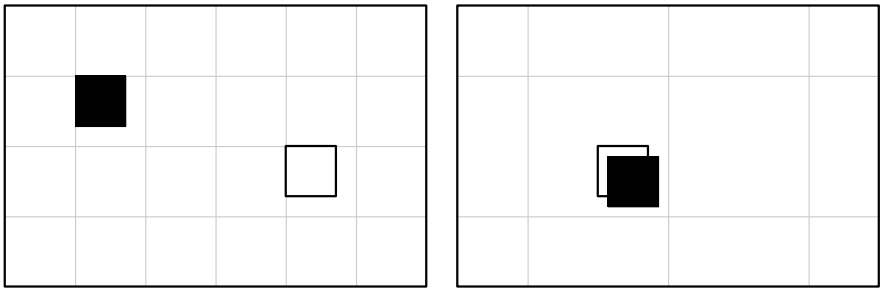
The experiment ran on Intel Core i7 computers running Linux Mint 13, rendering to a 22″ LCD monitor running at 1680×1080 px. Input was received through a wired HP Optical mouse, configured with the default polynomial acceleration profile of the X server. The software was written in Python and logged all user actions and subjective responses.

Twenty-seven volunteer undergraduate computer-science students took part in the experiment (six female). None participated in the prior experiments. Participation lasted approximately 20 minutes.

5.2.2 *Procedure*

Subjects completed 12 practice tasks of selections in representative neutral, congruent snapping, and incongruent snapping configurations. They then completed 17 conditions administered in a random order, with each involving 10 neutral tasks followed by the same number of tasks with a snapping configuration.

As the differences between experimental conditions were sometimes subtle (see below) and tasks within each condition were uniformly congruent



(a) *Pure-assisting*: The object can be aligned with the target without an R_2 component.

(b) *Pure-impeding*: When the object is selected, it will snap to the closest grid location – further away from the target.

FIGURE 5.3 Illustrations of the control tasks in Experiment 5.1.

or incongruent, there was less risk of subjects becoming aware of the manipulation and artificially constraining their responses to be consistent across conditions. Therefore, a within-subjects design was used to increase power and economy, and allow preferences across conditions for a single subject to be analysed.[6] At the end of each condition, subjects were asked:

> If you were to repeat the last set of tasks, would you prefer the snapping behaviour to be on or off?
> (Forced-choice selection: 'On' or 'Off'.)

This choice was coded as whether or not they preferred snapping. Subjects did not repeat the tasks.

6. In Experiments 4.1 and 4.2 subjects only needed to count the number of congruent and incongruent tasks to reliably distinguish conditions.

5.2.3 *Design*

All tasks involved selecting a 50×50 px solid black square, and dragging it a distance of 252 px onto a hollow target (a 50×50 px unfilled square with a 1 px black outline). While selected, the black square turned blue, and when positioned exactly over the target it turned green (and completed the task when released, as illustrated in Figure 5.1).

The start locations were randomly placed anywhere on the screen within the grid alignment constraints, but not within 100 px of the display edge. The target location was computed to be along one of the randomly selected cardinal/ordinal directions (within the same constraints as the start location). Grid lines were never shown to subjects to avoid foresight or anticipation about each task's congruency.

All neutral tasks were completed with a 4 px movement resolution, without the need to release and reacquire the object.[7] This 4 px resolution eased the difficulty of a task that would have otherwise required a high degree of precision,[8] and was not noticeable to subjects – the object appeared to move smoothly and continuously.

7. A Fitts' law ID of 6 bits.

8. A Fitts' law ID of 8 bits.

Four control conditions were used to confirm that snapping tasks with only assistive components were preferred to neutral tasks, and neutral tasks were preferred to snapping tasks with only impeding components:

- In two *pure-assisting* conditions (Figure 5.3(a)), the target was aligned with the R_1 resolution grid of either 18 or 84 px – which negated the need for the release-and-reacquire and R_2 components.
- In two *pure-impeding* conditions (Figure 5.3(b)), the object was initially placed 12 or 61 px from the target with an R_1 resolution of 126 or 510 px, respectively. This caused the object to snap further away from the target when first moved (i.e. negative progress). The *pure-impeding* R_2 resolution was set to 4 px, offering no advantage over neutral tasks.

Tasks in the remaining 13 conditions included elements of both assistance and impediment. These conditions are described with tuples of the form (R_1, R_2), and were chosen to sample a range of gains and losses. Larger values for R_1 provide greater assistance during target approach, and larger values for R_2 provide greater assistance during final positioning. All of these conditions require the user to release-and-reacquire the object, and tasks could always be completed using the R_2 resolution at any time. The conditions (in pixels, ordered by R_2) were:

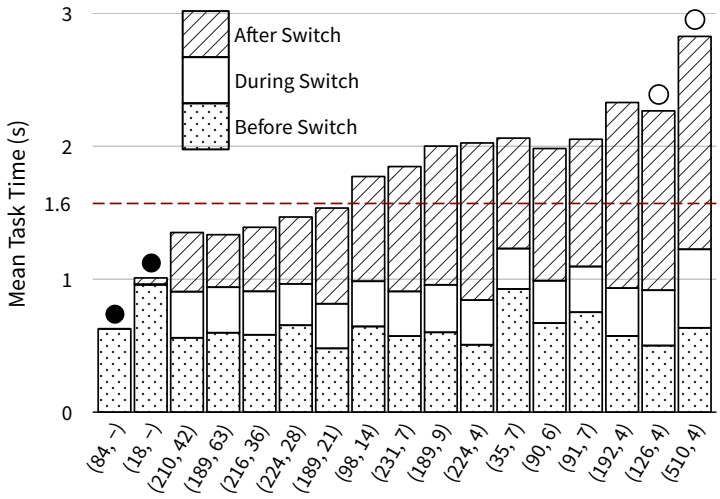
(192, 4),	(224, 4),	(90, 6),	(35, 7),	(91, 7),
(231, 7),	(189, 9),	(98, 14),	(189, 21),	(224, 28),
(210, 42),	(216, 36), and	(189, 63).		

5.2.4 *Results*

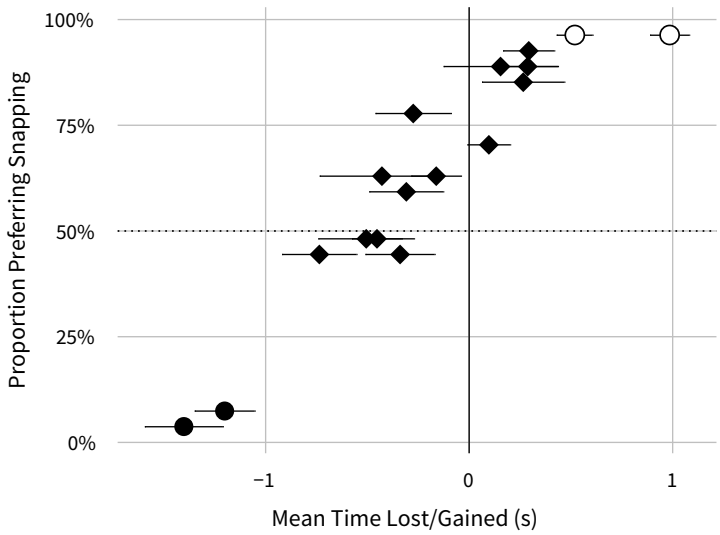
Control Conditions. Across all conditions (control and experimental), the mean time to complete a neutral task was 1.57 s (95% CI [1.54, 1.60]).[9] As intended, pure-assisting snapping tasks were significantly faster than neutral tasks: pooled mean 0.82 s, 95% CI [0.75, 0.94], paired $t(53) = 16.37$, $p < .001$, $d = 2.23$. Similarly, pure-impeding snapping tasks were significantly slower than neutral tasks: pooled mean 2.55 s, 95% CI [2.53, 2.57], paired $t(53) = -20.78$, $p < .001$, $d = 2.81$. Subjects' choices confirmed the expected subjective preferences for the control conditions: 95% chose the pure-assisting snapping over neutral, whereas only ~5% chose the pure-impeding snapping. The time and choice data for these control conditions are shown at the extreme left and right of the plots in Figure 5.4.

9. Given the 10 tasks in each set, total time is ~10 times higher.

Experimental Conditions. The mean time to complete the remaining (non-control) snapping tasks varied between 1.34 and 2.33 s – between 0.74 s slower and 0.29 s faster than for neutral tasks. Figure 5.4(a) summarises the task time results by condition, and highlights the division of time around the R_1/R_2 switch: *before* (❶ in Figure 5.2(b)), *during* (❷), and *after* (❸). In general, the



(a) The mean division of task time in each condition. The dashed line shows mean neutral task time, and the ●/○ dots indicate the pure-assisting and pure-impeding control conditions, respectively.



(b) The proportion of subjects who preferred snapping by each condition's mean time lost/gained per task (±95% confidence interval). The ●/◆/○ marks indicate the pure-assisting/experimental/pure-impeding conditions, respectively.

FIGURE 5.4
Summary of the time and choice results for Experiment 5.1.

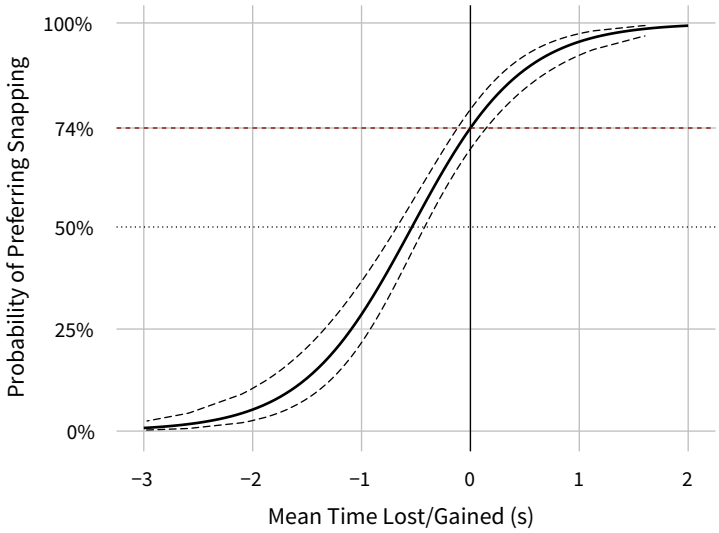


FIGURE 5.5 A binomial logistic regression model (dashed ±95% confidence interval band) for subject choice by the actual amount of time lost or gained in Experiment 5.1. The intersection with the vertical axis – when no time was lost or gained – is highlighted (dashed, maroon).

TABLE 5.2 The division of time around the R_1/R_2 switch. Plotted per condition in Figure 5.4.

	Mean (s)	95% CI
Before Switch (❶)	0.63	[0.57, 0.71]
During Switch (❷)	0.32	[0.25, 0.39]
After Switch (❸)	0.80	[0.57, 1.04]

Note: CI = Confidence Interval.

time spent before and during the switch were relatively constant, with the overall gain or loss largely determined by the final R_2 component (Table 5.2).

Choice Response. The proportion of subjects choosing the snapping behaviour in each condition is shown in Figure 5.4(b). As expected, this proportion correlates with the mean time lost or gained from the snapping behaviour (R^2 = .89).

When snapping offered no performance benefit, the majority of subjects (~75%) chose to turn snapping on (top-left quadrant of Figure 5.4(b)). Even when the snapping behaviour cost subjects half a second per task (around five seconds in total), ~48% chose to turn it on. Only the pure-impeding conditions showed a strong majority of subjects rejecting the snapping behaviour.

Subject choices were analysed using a binomial logistic model (shown in Figure 5.5). The model was fit to the data for all conditions completed by all subjects and describes the overall probability of a subject choosing the snapping behaviour when given a particular mean task gain or loss ($\chi^2(1)$ =

137.93, $p < .001$; Odds Ratio = 7.20, 95% CI $[4.85, 11.11]$). Of particular interest is the intercept: the probability of a subject preferring the snapping behaviour when there was no time gained or lost by using it. Objectively, subjects should be indifferent to such an outcome (a probability of 50%), but the intercept here at 74% (95% CI $[69\%, 79\%]$) confirms a significant bias in favour of the snapping behaviour when it offered no performance benefit (binomial $p = .02$).

The within-subjects design allows analysis of the choices made by each subject using a logistic model across only their samples (the model shown in Figure 5.5 is the average of these individual models). This analysis provides insight into the variability of individual decision-making choices, rather than only the global average behaviour. In particular, the consistency of subjects can be classified from the separation of the models: is there some task gain/loss time below which snapping is uniformly rejected, and above which snapping is uniformly accepted? A lack of separation would indicate that there was a range of objective time gains or losses that the subject was either insensitive or indifferent to. Across all 27 subjects, only 4 (15%) had models with complete separation – indicating that the conditions they rejected could be reliably distinguished from those they accepted based on the objective time alone. For the others, there was some overlap between the greatest gain where they rejected snapping and the greatest loss where they accepted snapping (mean 0.85 s, 95% CI $[0.58, 1.12]$).

5.2.5 *Discussion*

The results of this experiment indicate a positivity bias in favour of the snapping behaviour – even when it delivered objective performance losses. This can be viewed in contrast to the negativity bias in Experiment 4.1 and the balanced responses in Experiment 4.2 (although the experimental designs do not permit a statistical comparison), and can be understood within the context of the model presented in Chapter 3.

The release-and-reacquire action added mental and motor costs to the task, and added a constant time overhead. However, any possible negativity towards these actions did not overwhelm subjective responses: subjects chose to turn snapping on despite its additional actions and slower performance. Even when performance with the snapping behaviour was 45% slower than without it, 44% of subjects preferred it – suggesting that the additional task time and actions had a relatively weak effect on experience. This is not simply a preference for any sort of snapping behaviour as the pure-impeding conditions were almost uniformly rejected (indicating that subjects could at least identify when snapping was obviously not beneficial).

A key feature of the experimental snapping conditions was that progress towards the goal was consistently positive: The snapping prior to the grid resolution switch (during R_1; ❶ in Figure 5.2(b)) always provided greater progress towards the goal than that for neutral tasks, and even after the switch (during R_2, ❸) the progress was also often greater than that for neutral tasks (and never any worse). That is, the snapping behaviour was assistive (both perceptually and actually) in terms of bringing the object closer to the target – even if this was not the case for the objective time consumed in doing so. In terms of the model, the snapping behaviour featured consistently positive consumption utility: $m_k(c_k) \geq 0$ for the dimension k of distance to the target. And even though the release-and-reacquire action had a negative marginal utility as progress was not being made ($m_k(c_k) < m_k(r(a))$), this was ameliorated by the positive marginal utility before and possibly after it. The pattern of responses from subjects indicates that this dimension has substantially more weight that any alternative dimension (such as objective task time – which generally had negative marginal utility during and after the release-and-reacquire action), and therefore supported a higher overall utility for the snapping behaviour.

This is not to say that time is of no concern to users – only that the consumption utility weighting of progress is greater than that for time. The effects of duration neglect, violations of temporal dominance, and other studies of time perception (§§2.5.3–2.5.4) also support a model of preferences that underweights the importance of time on subjective experience. However, these two components (time and progress) are not perfectly separable: progress correlates with, and is bounded by, the objective time taken to complete the task. That is, tasks that take more time to complete when using the snapping behaviour necessarily have some element of negative marginal progress.

Both R_1 and R_2 were positively correlated with the proportion of choices to turn snapping on (Pearson's $r = 0.43$ and $r = 0.82$, respectively), suggesting that both forms of assistance were positively valued. The fact that subjects chose to enable the snapping behaviour even when it was slower overall suggests that subjects positively valued these elements of assistance more than they negatively valued the release-and-reacquire impediment.

This highlights the multi-dimensional nature of progress, and calls for further experimentation. Positive marginal progress in one dimension (in this case, the visual distance to the target) cannot overwhelm negative marginal progress in another indefinitely (e.g. losses in time) – at some point the losses will become too great to ignore. Although these results show a preference for the positve visual progress of the snapping functionality despite the loss of time, the balance between these two methods of measuring progress has not been revealed (i.e. the point where the extra effort to disable snapping or the loss of time doing so overwhelms gain-loss utility).

The lack of separation in subjects' responses – that they could not perfectly discriminate the amount of time they gained or lost – is not particularly surprising as a certain amount of variance and error is expected in any human response. However, the time overlap between conditions that subjects rejected and those that they accepted provides insight into the range of objective gains and losses that users either cannot reliably distinguish between, or are indifferent to. The results of this experiment indicate this is approximately one second per task, but as this was not the primary hypothesis, further work is required to draw a robust conclusion.

5.3 SUMMARY

This chapter developed a new experimental task that can be used to further test the model and explore the psychological biases described in this thesis. An experiment using the task found a positivity bias for its drag-and-drop snapping behaviour (Experiment 5.1). These results can be compared with those of Experiments 4.1 and 4.2, and complement them when understood in the context of the model presented in Chapter 3: the positive perceived progress towards the task goal that the snapping behaviour provided weighed heavily on subjective preferences – in excess of the negative objective performance losses it inflicted.

PEAK-END EFFECTS & EXPERIENCED UTILITY 6

THE TASK developed in the previous chapter is general enough to create experiences with positive, negative, and proportionate objective and subjective experiences (Table 5.1). This flexibility allows for some of the other psychological biases (reviewed in Chapter 2) that are not covered by the model to be examined by constructing interactive experiences that have specific experiential properties. In particular, it can be used to test the effects of experienced utility (§2.5), wherein the order of moments in an experience influences subjective experience (specifically, remembered utility). Experiments on experienced utility effects have demonstrated that the remembered utility of an event can diverge from the arithmetic sum of its components (total utility).

This chapter uses the task developed in the previous chapter to examine two experienced utility biases that are beyond the scope of the model: the peak-end rule and duration neglect.

The model of reference-dependent preferences does not capture these effects of experienced utility. Consumption utility m_k is independent of the actions or outcomes that preceded it, or will succeed it. That is, the model describes the total utility of an experience,[1] and not its remembered utility. As noted in Section 3.2, this is knowingly naïve of the effects studied in the experienced utility literature.

1. The time integral of moment-to-moment instant utilities.

This chapter presents an experiment that examined two well-established experienced utility effects: the peak-end rule and duration neglect. Experiment 6.1 used the drag-and-drop task detailed in Section 5.1 with the forced-choice methodology detailed in Section 3.3 to create series of tasks that would have the hypothesised peak-end and duration neglect effects.

6.1 EXPERIMENT 6.1: PEAK-END AND DURATION NEGLECT

The peak-end rule suggests that people evaluate an experience primarily from its most intense moment and its terminating moment. That is, people prefer experiences that terminate with a positive moment, or spread painful moments over a long period of time (to avoid creating a peak). The latter also relates to the effects of duration neglect, wherein people do not discriminate between experiences based on their duration. Duration neglect (and other time perception effects) are alluded to in the experiments presented so far: the gain or loss of objective time has not been a reliable predictor of preference (although they are correlated). However, these experiments did not isolate these effects in their designs.

The drag-and-drop task with grid snapping can be used to analyse these effects by constructing series of tasks that mix various snapping configurations (grid resolutions) to create controlled peaks, ends, and series durations.

For example, adding a few congruent snapping tasks to the end of a series of incongruent snapping tasks can create a positive end to the experience of completing a substantially negative series. Such a series can be compared to one with the congruent tasks at the start of the series (or spread throughout). Both series will have identical objective completion times, but the peak-end rule predicts an asymmetric response to a choice between them. In particular, that more subjects will prefer the series with a positive end – indicating that the order of the tasks has an effect on subjective experience.

The method of this experiment followed that of the other experiments in this thesis, and the peak-end experiment for interactive tasks of Cockburn, Quinn, and Gutwin (2015): subjects completed two series of drag-and-drop tasks, and then selected which of the two they would prefer to repeat a third time. However, unlike the other experiments in this thesis, there was no reference condition because a utility/preference scale was not being constructed.

6.1.1 *Materials*

Task series were constructed from mixtures of three snapping behaviours:

- *Control.* The object moved directly according to mouse control (as with the neutral baseline in Experiment 5.1). Targets were at a distance of 252 px, with an inconspicuous movement resolution of 4 px.
- *Positive.* The target was aligned with an 84×84 px R_1 grid, which provided snapping assistance (as with the pure-assisting control conditions in Experiment 5.1).
- *Negative.* The object snapped to a 224×224 px R_1 grid and the target was misaligned by 24 px – with an R_2 resolution of 4 px (as with the (224, 4) experimental condition in Experiment 5.1).

The control behaviour is a neutral experience, with the intention that subjects would prefer positive snapping behaviour over the control, and prefer the control over the negative snapping behaviour.

Six conditions were constructed that involved subjects choosing between a pair of task series (illustrated in Table 6.1):

- Control Conditions:
  1. *Positive versus Control.* Ten control tasks were compared against 10 positive tasks.
  2. *Control versus Negative.* Ten control tasks were compared against 10 negative tasks.
- Peak-End Conditions:
  3. *Positive End versus Positive Start.* A random mix of 10 control and 10 negative tasks, followed by 4 positive tasks (a positive end experience), were compared with the inverse. Subjects were hypothesised to prefer the positive end over the positive start.

4. *Negative Start versus Negative End*. A random mix of 10 control and 10 positive tasks, followed by 4 negative tasks (a negative end experience), were compared with the inverse. Subjects were hypothesised to prefer the negative start over the negative end.

- Duration Neglect Conditions:
  5. *Positive End versus Short*. A random mix of 10 control and 10 negative tasks, were compared with same followed by 4 positive tasks (a longer series with a positive end). Subjects were hypothesised to prefer the longer series extended by positive tasks over the unextended series.
  6. *Long versus Negative End*. A random mix of 15 control and 15 positive tasks, were compared with a similar mix of 10 control and 10 positive tasks, followed by 4 negative tasks (a shorter series with a negative end). Subjects were hypothesised to prefer the longer series over the shorter series that had a negative end.

6.1.2 *Apparatus & Participants*

The experiment ran on Intel Core i7 computers running Linux Mint 17, rendering to a 22″ LCD monitor running at a resolution of 1680×1080 px, with input received through a wired Logitech optical mouse. The X server was configured to use a polynomial pointer acceleration profile, with a constant deceleration factor of 4. The software was written in Python and logged all user actions and subjective responses.

Forty-six volunteer undergraduate computer-science students took part in the experiment (nine female). None participated in the other experiments. Each received a $5 café voucher for their participation, which lasted approximately 10 minutes.

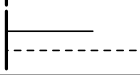
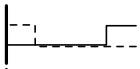
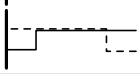
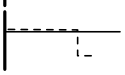
6.1.3 *Procedure & Design*

Subjects were instructed that they would make several choices between two series of drag-and-drop tasks, choosing which they would prefer to repeat if asked to do so. They first completed a series of 18 practice tasks to familiarise themselves with the control, positive, and negative task behaviours.

They then completed the six experimental conditions: the two control conditions first, followed by the remaining four in a random order. Within each condition, the order of exposure to the two series was counterbalanced. After completing the second series in each condition, they were asked:

> If you were to repeat one of the last two sets of trials, which would you prefer to do?
> (Forced-choice selection: 'First' or 'Second'.)

TABLE 6.1 Summary of the condition construction and choice results of Experiment 6.1.

	Condition		Experience	Choice (%)		
	A	B		A	B	*p*
1	Positive (10P)	Control (10C)		**87**	13	<.001
2	Control (10C)	Negative (10N)		**72**	28	.004
3	+End (10[CN]+4P)	+Start (4P+10[CN])		63	37	.10
4	-Start (4N+10[PC])	-End (10[PC]+4N)		63	37	.10
5	+End (10[CN]+4P)	Short (10[CN])		61	39	.18
6	Long (15[PC])	-End (10[PC]+4N)		**78**	22	<.001

Note: The *A* conditions were hypothesised to be preferred and the *B* conditions were hypothesised to be not preferred. The task makeup in parentheses describes their construction – for example, '10[CN]+4P' indicates 10 control and 10 negative tasks in a random order, followed by 4 positive tasks. The *Experience* plot shows the hypothesised subjective experience across time for the *A* (solid) and *B* (dashed) conditions. The *Choice* results show the proportion of subjects that preferred each condition, with statistically significant preferences (binomial sign test) in boldface.

TABLE 6.2 Summary of the time results and series differences for each condition in Experiment 6.1.

	Condition		Mean Time and 95% CI (s)		Difference			
	A	B	A	B	B−A	95% CI	*t*(45)	*p*
1	Postive	Control	7.62 [7.06, 8.18]	17.15 [16.27, 18.02]	**9.53**	[8.61, 10.45]	20.78	<.001
2	Control	Negative	17.08 [16.40, 17.75]	27.48 [24.77, 30.19]	**10.41**	[7.72, 13.09]	7.80	<.001
3	+End	+Start	44.28 [41.80, 46.76]	44.51 [42.72, 46.31]	0.23	[-1.47, 1.94]	0.28	.78
4	-Start	-End	35.28 [33.76, 36.79]	34.62 [33.24, 35.99]	-0.66	[-2.02, 0.70]	-0.98	.33
5	+End	Short	46.26 [44.10, 48.43]	40.34 [38.76, 41.92]	**-5.93**	[-7.67, -4.18]	-6.93	<.001
6	Long	-End	36.26 [34.93, 37.59]	34.46 [32.95, 35.98]	**-1.80**	[-2.97, -0.62]	-3.08	=.003

Note: Refer to Table 6.1 for the construction of each condition. Unlike previous experiments, times are reported for the entire series (rather than per task). CI = Confidence Interval. Significant differences (paired *t*-tests) are in boldface.

The primary dependent measure was subjects' response to this question. Subjects did not repeat the tasks.

Tasks were generated as they were for Experiment 5.1 (§5.2.3).

6.1.4 Results

The results are analysed separately for each of the conditions, and are summarised in Tables 6.1 and 6.2.

Control Conditions. Nearly all participants preferred (40 of 46, 87%) the positive snapping behaviour in favour of the control (Condition 1; binomial $p < .001$) and rejected (33 of 46, 72%) the negative snapping behaviour in favour of the control (Condition 2; binomial $p = .004$). The snapping behaviours also had the intended gain or loss in time, with positive snapping performing significantly faster than the control ($d = 3.85$), and negative snapping performing significantly slower than the control ($d = 1.56$) – both detailed in Table 6.2.

These results confirm that the manipulated snapping behaviours in isolation had the desired subjective preference responses.

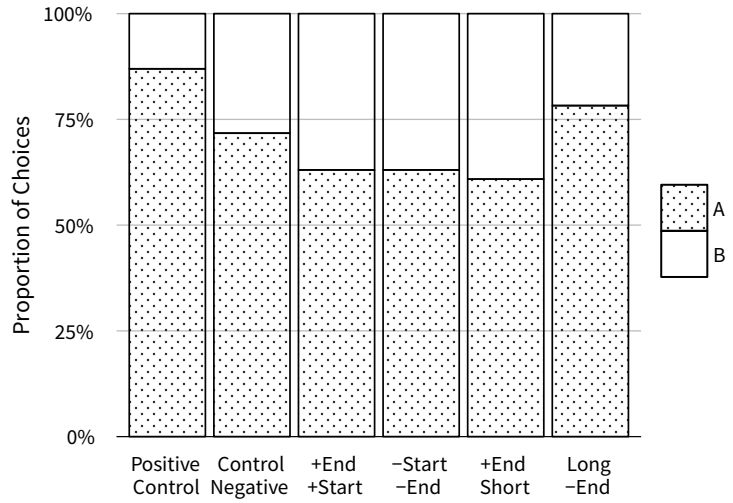
Peak-End Conditions. Neither peak-end condition (Conditions 3 and 4) had significantly different time performance between the series (as intended – the tasks in each condition were identical, but simply ordered differently). Although a majority of subjects favoured the positive end (Condition 3) and negative start (Condition 4) series (29 of 46, 63% for both), the proportions were not statistically significant (binomial $p = .10$, revisited in §6.1.5).

These results do not confirm the presence of a peak-end effect.

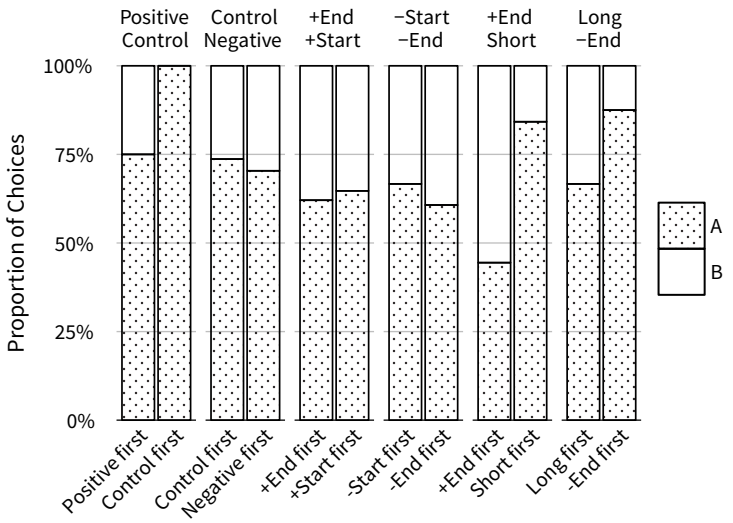
Duration Neglect Conditions. Both duration neglect conditions (5 and 6) had a significantly different time performance between the series in the intended direction: the short ($d = 0.93$; Condition 5) and long ($d = 0.38$; Condition 6) series performed as described against their counterpart series. Condition 6 found a significant proportion preferring the longer series over the series with a negative end (36 of 46, 78%, binomial $p < .001$), but although a majority preferred the positive end over the shorter series in Condition 5 (28 of 46, 61%), this was not significant (binomial $p = .18$).

These results provide some support for a duration neglect effect in the presence of a negative end effect.

Order Effects. Figure 6.1(a) shows the choice data, and 6.1(b) showing the same data divided by the series completed first by subjects – for example, the left-most pair of bars in Figure 6.1(b) shows that all subjects chose the hypothesised preferred condition (A – positive) when they completed the control series first, but only 75% chose it when they completed the positive snapping series first. This was significant for Conditions 1 (two-sample test for equality of proportions $\chi^2(1) = 4.3$, $p = .04$) and 5 ($\chi^2(1) = 5.8$, $p = .02$).



(a) Choice results for the hypothesised preferred (A) and not preferred (B) series in each condition (see Table 6.1).



(b) Order effects in the preference choices. The proportion of choices for the hypothesised preferred (A) and not preferred (B) series in each condition are divided by the series completed first (see Table 6.1).

FIGURE 6.1
The choice results for Experiment 6.1.

6.1.5 Discussion

These results provide support for the peak-end and duration neglect psychological biases in human–computer interaction tasks, but the results do not reach the level of statistical significance. This follows prior human–computer interaction work that has either sought (Cockburn et al., 2015) or appealed (Harrison et al., 2007; Harrison, Yeo, & Hudson, 2010) to the peak-end rule, but using the choice-based methodology described in Chapter 3. However, there are some significant limitations to these results, which can be resolved through further experimentation.

Peak-End. Although neither of the peak-end conditions found a significant effect ($p = .10$ at $\alpha = .05$ for both conditions), a one-tailed test in the hypothesised direction (Fisher, 1954, 1971; Kimmel, 1957) of the peak-end effect would find statistical significance ($p = .05$ for both conditions). Such an analysis was similarly performed for the seminal peak-end experiment of Kahneman et al. (1993) to reach the same threshold. However, as the use of one-tailed tests in psychological experiments is controversial (Burke, 1953; Eysenck, 1960; Hick, 1952; Marks, 1951), such an appeal will not be made here. Rather, these results support the presence of peak-end effects in an interactive context – but further work is required to verify this claim, establish the size of the effect, and the conditions under which it appears.

Unlike prior psychological work which has used strong hedonic stimuli such as physical pain (e.g. Ariely, 1998; Fredrickson & Kahneman, 1993; Kahneman et al., 1993), this experiment used comparatively weak stimuli for inducing the subjective peaks and end effects. In particular, the negative experience was not particularly aversive (as physical pain is), which is likely to contribute to a weakening of the effect. Therefore, more punishing interactions (such as task failure) may more readily expose it. For example, the pure-impeding control conditions in Experiment 5.1 offered tasks where the snapping was effectively useless (although there were concerns that it would overwhelm experience if used in this experiment).

Duration Neglect. Only one of the duration neglect conditions showed an effect, with significantly more subjects preferring a longer series of tasks to one that finished with a negative end (Condition 6). However, the lack of a corresponding effect for Condition 5 and the lack of any negative snapping tasks in the longer series raises an alternate explanation: subjects responded to the presence of negative tasks, rather than the extension of the series.

The lack of an effect for Condition 5 may also be due to the substantial time difference between the two series. Subjects were ~6 s faster with the shorter series in Condition 5, whereas they were only ~2 s slower with the longer series in Condition 6. Therefore, the large difference in Condition 5 may have

made time more salient to subjects, or the small difference in Condition 6 may not have been detectable. The duration neglect hypothesis describes an underweighting of the duration of an event in favour of the magnitude of its constituent experiences, and it should be emphasised that this is distinct from ignorance or a lack of sensitivity to the duration. For instance, in the cold-pressor experiment (Kahneman et al., 1993) subjects could reliably detect the difference in condition durations – but it was not subjectively felt as much as the water temperature. To be confident about the presence of a duration neglect effect it is important that subjects are aware of a time difference but openly choose to ignore it. As such, these results for duration neglect are inconclusive.

Other Experimental Issues. There is an interesting contrast between the results of Experiments 5.1 and 6.1. The negative snapping behaviour used in this experiment was also one of the conditions in Experiment 5.1 (224, 4). However, while 44% of subjects accepted snapping in that condition of Experiment 5.1, only 28% did so in Condition 2 of this experiment. Although an unplanned comparison does not find these proportions to be significantly different ($\chi^2(1) = 1.33$, $p = .25$),[2] there are some factors of the experimental design that may contribute to a difference in subjective responses. In particular, Experiment 5.1 featured conditions with behaviour that was much worse than that in this experiment – which may have established a preference scale for subjects.[3] That is, subjects' choices in that experiment may not only have been influenced by the reference condition without snapping, but also by the conditions subjects knew they could potentially experience. This influence is undesirable, but difficult to eliminate entirely: even with a between-subjects experimental design, subjects would still need to be exposed to the behaviour at some point to ensure a lack of familiarisation doesn't impair performance.

2. The small number of subjects creates wide confidence intervals.

3. Although conditions were counterbalanced, subjects were exposed to the different varieties of snapping behaviour during the practice session.

This contamination of reference expectations is indicative of a decision-by-sampling approach to decision making (Stewart et al., 2003; Vlaev et al., 2011, §2.4), wherein the set of options presented to a subject influences their choice. Even though subjects were not given a set of options to select between and were not choosing between conditions they had not experienced in these experiments, the knowledge of what the range of conditions was[4] may have induced a similar effect (e.g. Quiggin, 1982).

4. In particular, the most positive and most negative.

In contrast to many of the prior peak-end studies that used continuous stimuli which were gradually adjusted over the course of the experiment (e.g. Ariely, 1998; Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Redelmeier et al., 2003; Schreiber & Kahneman, 2000), the experiential changes in this experiment were discrete events. Each task had a distinct positive, negative, or neutral sensation. This is consistent with the experience-by-moments hypothesis (§2.5.1), but it presents an opportunity for subjects to

become more cognisant of each episode. Although most interaction tasks are discretised, there does exist continuous interaction tasks that may be manipulated and explored in a manner closer to prior peak-end studies (e.g. games: Gutwin, Rooke, Cockburn, Mandryk, & Lafreniere, 2016).

6.2 SUMMARY

This chapter described an experiment that used the choice-based methodology and drag-and-drop tasks from earlier chapters to test the effects of experienced utility (reviewed in §2.5)[5] – in particular, the peak-end rule and duration neglect (Experiment 6.1). The hypotheses of Experiment 6.1 were not captured by the model, but can be tested using the same experimental methodology and analysis practices. The experimental results were mixed, but provided some support for the presence of the peak-end rule for interaction tasks.

5. As opposed to the total utility studied previously.

GENERAL DISCUSSION 7

This chapter reviews the experimental results in the context of the judgement and decision-making literature, and discusses its applications, limitations, and directions for future work.

THE EXPERIMENTAL results of this thesis found that subjects responded negatively to an interface with objective benefits (Experiment 4.1), positively to an interface with objective costs (Experiment 5.1), and proportionately with the objective costs and benefits of an interface (Experiment 4.2). These discordant results can be understood coherently with a model of reference-dependent preferences (Chapter 3): the objective costs and benefits did not weigh on subjects' preferences as much as the perceived progress towards the task goal, which had a higher overall utility. The model describes the utility of an interaction as a psychologically-biased weighting of the difference between the outcomes that a user is expecting and those they actually receive. These outcomes are multi-dimensional bundles, with one dimension for each aspect that was either expected or received (such as the response time, system state changes, and feedback). Crucially, the received outcomes are compared with the user's reference expectations of the outcome. Failing to meet expectations is more detrimental to overall utility than exceeding expectations is beneficial. Not all dimensions are weighted equally, with the experimental results demonstrating an overweighting of the perceived progress towards a task goal, and an underweighting of the objective time performance.

The methodology used to test this model exposed subjects to two conditions: a baseline interface to establish reference expectations, and a manipulated experimental interface. After completing a set of tasks with each, subjects were asked which they wanted to use when repeating the tasks (either hypothetically or actually). These choices imply a utility relationship between the two conditions and construct a preference scale that tested the predictions of the model. The methodology can also be used to test psychological preference biases in general – demonstrated for the peak-end rule and duration neglect (Experiment 6.1).

The model has broad implications. These experiments were designed to parallel work on asymmetric experience in the behavioural economics literature (Chapter 2): negativity bias, prospect theory's value function, and the reference-dependence of subjective evaluation. This expands the applicability of behavioural economics methods and models to a new choice domain, and introduces them to the practice of human–computer interaction research. Given the breadth of the related literature, the work here is not comprehensive and offers many opportunities for future work to refine the model or investigate other behavioural biases.

This chapter reviews the applications for the findings presented here (§7.1), limitations of the presented experiments (§7.2), and opportunities for further development: including the effects of risk and uncertainty (§7.3), the presence of other psychological effects (§7.5), and helping improve user productivity through prescriptive models (§7.4). A brief discussion of the potential applicability of other economic fields to interaction is also given (§7.6).

7.1 APPLICATIONS

Interface designers often develop systems that seek to assist task completion by adapting to the user's input. Examples are abundant: suggestions in mobile text entry, URL type-ahead in web browsers, search query rewriting, autocorrection, recency lists, history, automatic formatting, spam filters, and so on. These potentially assistive features will periodically fail, requiring additional user actions beyond those that would have been necessary if the feature was absent or worked correctly. This raises questions about how to design interfaces that avoid provoking strong negative responses, and how to experimentally measure asymmetries between objective and subjective experience.

The experiments in this thesis used word selection and drag-and-drop tasks that are supported in real-world user interfaces – representing real design issues that may have been considered prior to their release. For example, the word snapping behaviour is implemented in current versions of Microsoft Word, and grid snapping is a common feature of graphics applications. The presence of these features indicates that the application's designers felt that the potential for assistance was at least balanced with the potential for impediment (although their actual design approach is not known).

7.1.1 *Implications for Design*

In general, the results presented in this thesis suggest that designers should exercise conservative design in the provision of assistive features if they aim to maximise user experience (rather than objective performance). They also suggest that assistive features, where possible, should avoid returning the user to a previously attained state or moving them further away from their task goal. For example, there are risks associated with an assistive text-entry system that automatically replaces a user's typed characters because doing so may trigger a loss aversion that creates a negative subjective experience (despite overall improved performance). Passively offering text corrections, in contrast, does not saliently remove progress, and may be preferred despite the additional cognitive requirements to attend and act upon them (demonstrated by Quinn & Zhai, 2016, Appendix B).

A design focus on the potential benefits of an assistive technique is risky if users will periodically encounter minor impediments that are likely to have a stronger influence on their subjective experience. The negativity bias associated with losses may overcome the positive sensations derived from many occurrences of correct assistance. This is not to suggest that assistive techniques should be avoided, but that the design of their failure cases should receive proportionate attention. This is further emphasised by *confirmation bias* (Wason, 1968): the psychological tendency to seek information that confirms one's hypotheses rather than rejects them. This suggests that designers tend to overemphasise the potential for their systems to assist, and underemphasise their potential impediments – leading to naïve and vacuous experimentation (Hornbæk, 2015). Negativity bias suggests that designers should closely scrutinise the impact of potential failures to assist the user, as they may be more formative of their overall assessment.

However, designers should not be cynical about assistive features, but take a considered approach to their design. This echoes approaches to interaction design which emphasise the same aspects of goals and progress that are suggested by this work. For instance, *activity-centred design* (Norman, 2005, 2006) and *goal-directed design* (Cooper, Reimann, & Cronin, 2007) advocate approaching interaction design in terms of the goals that a user wants to accomplish (in contrast to a focus on the discrete features or tasks of an application that users assemble to complete a goal). For example, a text-entry suggestion system may be designed with an awareness of the text-entry context (e.g. a text message to a friend vs. an email to a work colleague) to improve the content of its suggestions for achieving common types of messaging goals. The model of reference-dependent preferences embodies a similar sentiment: utility is maximised when assistive features help users complete their goals with monotonically increasing progress.

7.1.2 Implications for Research

The model and methodology (Chapter 3) provide tools for researchers to understand user preferences and conduct controlled experiments that examine how specific components of a system's behaviour influence those preferences. The model decomposes the components of an interactive task, and formalises the relationship between interactive actions and their utility. The model leverages existing research tools in human–computer interaction – such as GOMS and KLM – to find these components and their objective value, using judgement and decision-mkaing principles from economics and psychology.

The methodology can be used to test the model, but also offers a generic framework for testing user preference manipulations. Generic experimental

frameworks support research through a consistent experimental design and common dependent variables that are amenable to the comparison of results between experiments, and meta-analyses of research agendas (which are rare in the human–computer interaction literature [Hornbæk, 2015]). This is intended to advance the current practice of arbitrary experimental design for subjective measures of user experience – which have produced turbid results (e.g. Hassenzahl & Monk, 2010, and reviewed in §2.7.2). The choice-based dependent measure lessens the problems of validity and calibration associated with preference scales (reviewed in §3.3.2), and can be used to construct such a scale from a series of preference orderings (although this requires more experimental control and data collection).

The creation of a reference point for comparing interface preferences is not common in human–computer interaction experimental practice: typically, the subjective features of experimental interfaces are studied in isolation[1] or ranked across experimental conditions.[2] This makes it difficult to draw generalisable conclusions from the data or conduct meta-analyses of experiments as the ground truth for the scale is unknown.[3] In contrast, concerns about the baseline for such comparisons are common in economic and psychological experiments – and are a central feature of the model of reference-dependent preferences. The work presented in this thesis promotes analogous experimental controls in human–computer interaction research.

1. e.g. preference or workload scales, such as the NASA-TLX (Hart & Staveland, 1988).
2. In contrast, objective measures are usually compared against a reference interface.
3. e.g. rating scales may not have the same meaning between subjects or experiments.

7.2 LIMITATIONS

As mentioned throughout this thesis, the behavioural economics literature has significant breadth and depth. This thesis has served primarily to introduce the literature to human–computer interaction and provide a foundation for future exploration into it. As such, many features of the model and biases in the literature have not been established or tested. The following sections describe general areas for future work, with this section describing some particular limitations of the experimental work.

Additional Actions. The experimental conditions all involved some additional action to trigger the outcome trade-off (the switch between snapping modes): a mouse movement in Experiment 4.1 and a key-press in the other experiments. The design of Experiment 4.2 was intended to alleviate concerns that subjects were responding to the additional action, rather than the progress offered by the experimental behaviour. However, an ideal experimental condition would expose subjects to the trade-off without their actions differing from the reference condition – that is, subjects would perform the same actions in both conditions, but receive different outcomes. This is difficult to engineer for the low-level pointing tasks used in these experiments

as subjects expect a direct-manipulation style of interaction for their mouse movements. Higher-level tasks or more novel interaction techniques may be able to create such conditions without confusing subjects.

Consumption Utility Shape. Each condition used an experimental interface that offered subjects a controlled consumption utility m_k (Equation 3.4). However, the shape of this utility curve was not examined.[4] The economic literature assumes that consumption utility follows 'the outcome-based utility classically studied in economics' (Kőszegi & Rabin, 2006, p. 1134), which is well-understood for economic outcomes (§2.2.2) but has not been verified for interactive outcomes. However, doing so is difficult because existing test methodologies (§2.2.3 and §4.3.1) require conditions that have precisely manipulable outcomes (cf. §3.3).

4. The independent variable was marginal utility; §4.3.1.

Furthermore, interactive outcomes are often bounded and are likely to exhibit ceiling effects. For example, although a monetary gamble can always win or lose greater sums of money, the grid snapping behaviour has a limit to its assistance (when R_1 equals the target distance) and its impediment (when the user must drag the object across the entire length of the screen).

Reference Point Manipulation. The experiments used a neutral interface without any assistance or impediment for the reference point – either conventional text selection (Experiments 4.1 and 4.2) or conventional drag-and-drop behaviour (Experiment 5.1). This reference point was never manipulated as an experimental factor.[5] That is, the experiments did not verify that if the reference point was adjusted, then preferences for the experimental interfaces would adjust accordingly. The model predicts that if the utility of the reference point increases, then the utility of an experimental interface with positive marginal utility must increase by a larger amount in order to maintain the same overall utility and preference for the experimental interface. That is, if $m(\mathbf{r}^*) = m(\mathbf{r}) + \omega$, $\omega > 0$; then for $u(\mathbf{c}|\mathbf{r}) = u(\mathbf{c}^*|\mathbf{r}^*)$ it must be that $m(\mathbf{c}^*) = m(\mathbf{c}) + \psi$ and $\psi > \omega$.

5. The peak-end experiment (6.1) compared interfaces without a reference point, but used comparisons between different series constructions.

This requires a reference interface with outcomes that are manipulable without changing the actions of the interaction – which may not be always possible for interactive tasks (e.g. in the text selection experiments, it is hard to imagine a reference condition that is somehow better than letter-by-letter but does not introduce any new actions and works equally well for all tasks). However, the task in Experiment 5.1 (object drag-and-drop) may support testing such reference point manipulation by giving the neutral reference interface some degree of congruent snapping or pointing enhancement.

Expectation Development. The choices that subjects made were intended to be independent of the other experimental conditions. That is, subjects were expected to choose between the neutral reference interface and the experimental interface without consideration of the other conditions used in

the experiment or speculation about performance beyond the experimental tasks. To help ensure that subjects were aware of this and had both experiences clear in their mind, the instructions to subjects emphasised the purpose of the choice and the two interfaces were always completed in pairs. However, as discussed in Section 6.1.5, it is difficult to confirm that subjects were not influenced by the other conditions or other expectancies. For example, in Experiments 4.1 and 4.2 all subjects practised on incongruent tasks, even if they never completed any during the experimental tasks. Their choices to use letter-by-letter or word-by-word snapping may therefore have been influenced by their speculation about the general utility of word-by-word snapping if they were to have encountered incongruent tasks.

Confirming that exposure to the reference condition creates a reference point may be achievable through reference point manipulation (described above), or by using more abstract tasks. For example, if the experimental interface and the tasks are unlike any interaction subjects regularly encounter, they may be less likely to speculate about its performance – but at the cost of ecological validity.

Procedure Invariance. The forced-choice question in all of the experiments asked subjects whether they wanted the snapping behaviour 'on' or 'off'. This question was designed to be neutral: it did not presuppose any default opinion or state of affairs. This avoided potential procedure invariance violations (wherein choices to *reject* are not inverses of choices to *accept*; §2.6.4), but may not represent the environment that interactive choices are usually made within. That is, most interfaces do not demand that users configure their features, but use a default configuration that the designers believe will benefit most users. Users are known to only rarely customise these defaults (§2.7.6), and the default configuration may instil an endowment for that behaviour (§2.6.2). If the experimental question was framed to ask subjects to deviate from a default – for example, 'Snapping will be enabled for the next set of tasks, do you want to turn it off?' ('Yes'/'no' choices, with 'no' already selected) – their responses may exhibit a bias for the default behaviour.

7.3 RISK AND UNCERTAINTY

Although risk is a central component of judgement and decision-making research (§2.2.5), the experimental work in this thesis has focussed exclusively on *riskless* prospects. That is, subjects were always informed of, and were exposed to, the outcomes they were choosing between – there was no ambiguity about what they would receive. This significantly simplified the experimental design and interpretation of the results: their responses can be attributed to

the received progress, rather than an aversion to risk or uncertainty in the experimental behaviour (e.g. aversion to the possibility of incongruent trials).

However, some of the most interesting effects in the literature involve perceptions of risk, and interactive prospects that users encounter are frequently *risky*. As neither designers nor users know with certainty what tasks will be encountered, they must consider the risk of impediment before deploying or engaging a system. For example, choosing to enable (or disable) a predictive text entry feature requires speculating about future interactions: the frequency of words that will be unknown to the predictive system, and the efficacy of the system in predicting desired words.

As reviewed in Section 3.3, it is difficult to design and expose experimental subjects to risky interactive prospects in the same manner as behavioural economics experiments: the probability of an interactive outcome cannot simply be told to subjects, it must be learnt through experience. This means that experimental designs need to carefully isolate the elements of risk (the probability of obtaining an outcome) from those of uncertainty (the lack of knowledge about probabilities). For example, methodology of Experiments 4.1 and 4.2 could be modified so that the third set of tasks is random – that is, the tasks completed after a subject makes their choice are not a repetition of the ones before the choice. This would examine the aversion to uncertainty that subjects have for the experimental interface (in consideration of the construction of the first two sets), but it would not test their aversion to risk as they have no information to estimate probabilities with.

Mobile text entry is a convenient interactive task to examine manipulations of risk. Fluent text entry on mobile devices is hindered by the small size of the keyboard and there are significant opportunities for systems to offer suggestions that complete words or correct errors (e.g. Weir, Pohl, Rogers, Vertanen, & Kristensson, 2014). Such suggestions are necessarily probabilistic – based on dictionaries and language models (e.g. Fowler et al., 2015; Goodman, Venolia, Steury, & Parker, 2001) – and users can choose to attend to them, or ignore them. This presents an experimental opportunity for examining how users perceive and respond to the risk in these systems: manipulation of the suggestion system's efficacy and measurement of changes in subjects' choices to engage with the system reveals the non-linearity of their decision weights (i.e. π in Equation 2.7).

The reference-dependent preferences model provided by Kőszegi and Rabin (2006) captures risk as the integration of utility over a probability distribution (§2.3.3). An extension to the model refined its approach to risk attitudes (Kőszegi & Rabin, 2007), making it consistent with the major empirical work on risk aversion, risk seeking, disappointment aversion, and uncertainty aversion. In contrast to other theories of risk attitudes, its reference-dependent

nature allows for prior expectations of risk in a prospect to moderate a person's level of risk aversion (e.g. Irwin, 1953).

7.4 RATIONAL BEHAVIOUR

No claims are made about the rationality of the choices subjects made in these experiments. Even though their choices are described as *biased*, this does not imply that they were poor or deficient – as reviewed in Sections 2.1 and 4.1.7, subjects undoubtedly had reasons and justifications for their choices. However, the biases subjects revealed in their choices present an opportunity for *correction* in the sense that Savage (1972) found his normative violations of the Allais paradox (§2.2.2) in *error*: 'There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be' (p. 103).

7.4.1 *Interactive Rationality*

The economics literature has sought normative models that yield optimal choices when followed. Although these models are not strictly adhered, a person may be able to improve the choices they make by adopting such a model. A contemporary application of this is *nudge theory* (Sunstein, 2014; Thaler & Sunstein, 2008; Thaler, Sunstein, & Balz, 2013), wherein people are surreptitiously guided to subdue their biases and make better choices with respect to their finances, health care, education, relationships, and so on.

The motivating force behind much human–computer interaction research into assistive interfaces and new interaction techniques is the improvement of user *productivity*. New interfaces are often judged on objective criteria (such as the time to complete a task or the incidence of errors in doing so), and models of interaction stress these qualities in their predictions. For example, if a text entry system improves average typing efficiency, this is presented evidence for its merit and legitimacy. In this sense, *productivity* is a normative ideal for human–computer interaction – it is the quality that designers and researchers most frequently strive to improve, with the assumption that users are seeking to maximise it.

However, as demonstrated by the empirical results in this thesis (and supported by general work on psychological biases), even if users are seeking to maximise their productivity, there are forces that inhibit a clear perception and attainment of that ideal. All of the experimental work here has found a correlation between objective time and choice – subjects generally preferred the interfaces that were faster[6] – however, subjects were poor at identifying whether they gained or lost, and relied upon more salient cues (such as perceived progress). If objective productivity was more salient, subjects may be

6. Except for Experiment 6.1, where manipulating this correlation was intentional.

successfully *nudged* towards the interface that improved it. For example, if Experiments 4.1–5.1 notified subjects of the exact time they gained or lost before they made their choice, their responses may have been led by it.

7.4.2 Interactive Nudging

A unique quality of interactive systems is their ability to adapt and alter their behaviour in response to a user's behaviour. That is, a system can adjust the information or options presented to the user in response to their past behaviour. For example, if a text entry system observes that a user continually rejects a particular word correction suggestion, it may choose to stop offering that suggestion to them. Quinn and Zhai (2016, Appendix B) explored one metric for managing this – *interface assertiveness*: the tendency for an interface to interrupt the user with probabilistic suggestions.

Awareness of the biases in users' perceptions and subjective experiences of interactive systems can be used by designers to create interfaces that guide user choices to be both productive and subjectively satisfying. Persuasive technology and decision support research actively explores using interactive systems to effect changes in user behaviour (reviewed in §2.7.6), but only recently have the behavioural changes been effected for the interactive tasks themselves. For example, Kim et al. (2016) examined positive and negative information framing in a productivity-tracking application – either informing subjects with a measure of their 'productive time' (positive framing) or their 'distracted time' (negative framing). They found that the negative framing led to a significant improvement in subjects' overall productivity (compared with no improvement for the positive framing). This provides prescriptive insight into how a system can use judgement and decision-making biases to improve both user performance and satisfaction. Others have explored similar interactive and interface techniques for helping users improve their security (e.g. Turland, Coventry, Jeske, Briggs, & van Moorsel, 2015), shopping (e.g. Kalnikaite et al., 2011; Todd, Rogers, & Payne, 2011), and privacy choices (e.g. Choe, Jung, Lee, & Fisher, 2013; Wang et al., 2014).

7.5 OTHER PSYCHOLOGICAL BIASES

This thesis has reviewed many psychological biases, but only tested a small fraction of them for their applicability to interaction. Some of the untested biases are captured by the model presented in Chapter 3, but some are not modellable with its parameters or construction. For example, Section 3.2 stated that the utility of an interaction I is determined by the sum of the utilities of its sub-interactions: $U(I) = \sum u(i)$. This simplification provides a starting point for model-based analysis but is knowingly naïve with respect

to biases such as peak-end (Chapter 6). Other biases that have been documented in the psychology literature that could be experimentally examined in human–computer interaction contexts include the following.

Hedonic Dimensions of Consumption. Kőszegi and Rabin (2004) discussed an extension to their model with a vector of hedonic dimensions that were applicable to a consumption bundle. For example, even if two goods are objectively perfect substitutes for each-other, hedonic properties (such as brand loyalty) may factor into gain and loss assessments. For interfaces, such dimensions may include aesthetics or social influences.

Endowment Effects. People have an aversion to change: preferring to keep items they already own, rather than exchange them for items of equal value (§2.6.2). This suggests that a new interface (e.g. that removes features or substantially changes interactions) must return additional utility to the user to overcome endowment effects. This is captured in the formulation of our model (by reference from Kőszegi and Rabin [2006]), and is likely reflected in additional utility for interfaces that the user is already familiar with (e.g. Sedley & Müller, 2013).

Sunk Costs. Another interpretation of the word-by-word technique's snapping behaviour in Experiment 4.1 is that it contained a sunk cost (§2.6.5). The actions to select text prior to disabling its snapping were *sunk* with respect to the subsequent actions: some of the text that subjects selected prior to disabling the snapping behaviour must be deselected, only to then be reselected. This interpretation is consistent with the progress loss explanation – a loss of attained progress is sunk – but other consequences of the cost are only implied (e.g. an aversion to future interactions).

Framing Effects. As discussed with the experimental limitations (§7.2), the choice questions in the experiments were designed to be neutral. However, framing effects (§2.6.3) can have a powerful influence on how choices are perceived, and have been demonstrated for interactive tasks (Kim et al., 2016). Careful management of framing may be important for user adoption and acceptance of new interaction techniques, or managing users' perceptions of changes to existing interfaces.

7.6 OTHER ECONOMIC FACTORS

As with the above psychological biases, there are other areas of economics that treat problems which share many properties with user interaction. Given the fundamental presence of expected utility (i.e. §2.2.2) in economic study, much of the material discussed in this thesis forms the basis for more advanced economic principles that may be adaptable to interactive contexts.

7.6.1 Game Theory

The inspiration for modern economic work on decision making was the formulation of utility and the axiomatic system of decision making given in game theory (von Neumann & Morgenstern, 1947). However, game theory's primary contribution was its analysis of games: where one or more agents make a series of choices (either sequentially or simultaneously) to advance towards certain pay-offs. The decisions agents face may depend on the choices made by themselves or other agents at prior points in the game, and their knowledge of the state of the game can be complete or partial. Game theory provides the tools for representing these games and analysing the decisions each agent faces. Research questions in game theory often involve how agents, in collaboration or competition, make choices as a game progresses to maximise their individual pay-offs.

Analyses of how an agent should respond to a decision in a game follow many of the normative principles reviewed in Chapter 2 – that is, the normative choices for an agent depend upon all agents in the game behaving rationally (and believing that the other agents will behave rationally, too). As with simple economic decisions, games played experimentally by human agents quickly violate these assumptions (reviewed by Camerer, 1997; Kneeland, 2015; Rabin, 1993).

Such a structure may also be useful for analysing user interactions: as a game between the user and the system.[7] That is, an interaction could be analysed as a series of decisions and pay-offs: the system's decisions about how to assist the user, and the user's decisions about how to respond to the system. These games would usually not be competitive (the system's pay-offs are likely to be aligned with the user's), but feature incomplete information sets wherein the system lacks perfect knowledge about the user's intentions and the user lacks perfect knowledge about the system's state. Analyses, strategies and solutions for such games from the game theory literature may provide interaction designers with insight into how to manage assistive systems (e.g. their assertiveness [Quinn & Zhai, 2016]).

7. And with other users in the case of collaborative systems.

7.6.2 Intertemporal Choice

When choices are made about prospects which have outcomes that will not be received immediately or will be received over a period of time, two questions arise: an economic question of how to discount the outcomes to their present value, and a psychological question of how people perceive the outcomes. These questions are closely related to those of utility and psychological value, but evoke biases in time perception and prediction (i.e. how people

form beliefs about the benefit of a delayed outcome). The original economic model for these prospects is P. A. Samuelson's (1937) *discounted expected utility*, wherein the utility (i.e. Equation 2.2) for an outcome that will be received at time period t is discounted by some factor δ^t – where δ is the discount factor for one period (Loewenstein & Prelec, 1992).[8] However, a number of biases have been observed when people face prospects that involve a temporal aspect in their outcomes (reviewed by Frederick et al., 2002):

8. Samuelson was explicit that his model should not be interpreted normatively: 'The idea that the results of such a statistical investigation could have any influence upon ethical judgments of policy is one which deserves the impatience of modern economists' (p. 161).

- The discount rate δ that people apply increases with t (known as *hyperbolic discounting*).
- People discount future gains more than future losses.
- People discount small outcomes more than large outcomes.
- People discount more to avoid the delay of outcome than they do to expedite the receipt of an outcome (a framing effect).
- Normative axioms are violated.

A number of models – some built upon prospect theory (e.g. Loewenstein, 1988; Loewenstein & Prelec, 1992) – aim to capture these effects (also reviewed by Frederick et al., 2002).

Interactive choices share many features with intertemporal choices as the outcomes of a choice to use a particular system or feature are typically realised over a period of time. For example, the gains of a word-by-word snapping behaviour may not be realised for an individual selection task, but become realised over the useful life of the system if a large proportion of all tasks are congruent with the behaviour. The discount rate applied to the utility of an interactive prospect is also likely to increase with the initial investment cost. For example, a spam filter may require a period of manual training and periodic checks for false-positives. Similarly, users exhibit an *aversion to changes* in the features of established products (Sedley & Müller, 2013) that may be analysed with these models. Models of intertemporal choice may assist human–computer interaction design and research by explaining and predicting how users assess these types of interactive prospects.

7.6.3 *Neuroeconomics*

The success of the collaboration between economics and psychology has recently expanded to include principles and findings in neuroscience that may explain economic phenomena (and vice versa) – with the intersecting field known as *neuroeconomics* (reviewed by Camerer, 2013; Loewenstein, Rick, & Cohen, 2008). The fundamental insight from neuroscience that can be applied to economic problems is that decision making is not a unitary process, but an interaction of multiple processes that guide behaviour (reviewed by

Evans, 2008, and briefly in §2.4). For example, the dual system theory of Kahneman and Frederick (2002, 2005) argued that decision making is driven by two systems – an unconscious and emotive system that operates quickly on intuitions and heuristics, and a more contemplative and analytic system that operates slowly on rules and reasoning (Stanovich & West, 2000). It is the evocation or suppression of these systems that drives psychological biases and non-normative economic behaviour (e.g. Fudenberg & Levine, 2006; Loewenstein & O'Donoghue, 2004).

Support for dual-system models is provided by work that employs methods from neuroscience (typically fMRI scanning) to understand the neural systems that correlate with behavioural biases – for example, there is a correlation between loss aversion responses and dopaminergic activity in the midbrain (Tom, Fox, Trepel, & Poldrack, 2007), and between activity in the amygdala/orbitofrontal cortex and ambiguity aversion (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005). Although such a level of precision is unlikely to be of interest to human–computer interaction practice, the models they produce may be of use in picking apart the competing forces that guide user behaviour and inform user experience in complex interactive systems.

CONCLUSIONS 8

Behavioural economics and psychological research on judgement and decision making provide theoretical foundations for understanding and explaining important issues in user experience. In particular, the economics literature has developed models that suggest (a) non-linear utility scores connect objective returns with subjective preferences, (b) gains and losses are divided around a neutral reference point, and (c) there is an asymmetry between the value derived from gains and the value derived from losses. The psychological literature has developed empirical evidence that supports such models, but also substantial evidence that subtle biases can disrupt them.

This thesis has sought to bridge this literature into human–computer interaction by adapting a model of reference-dependent preferences from the economics literature to interactive contexts. It has also developed an experimental methodology that can be used to test the model and other issues of user experience. The model explains how gains and losses are defined by the difference between an actual interface outcome and a user's expectations for that outcome. These gains and losses are then transformed into subjective experiences through an asymmetric value function that amplifies the experience of losses. The methodology tests this using a series of binary choices made by subjects. The pattern of choices establishes a preference scale that can be used to infer utility relationships and psychological biases.

The model's predictions were tested in three experiments. The first two measured user preferences for an assistive text selection interface in comparison to a neutral reference interface:

- In the first experiment, the assistive interface snapped selections to include entire words and required subjects to backtrack in order to disable the snapping behaviour when it was an impediment to task completion. Subjects displayed a bias against the snapping interface, even when it improved their performance (consistent with a negativity bias against progress losses).
- In the second experiment, the assistive interface again snapped to select entire words, but the mechanism to disable selections was altered to maintain the user's progress: subjects disabled snapping by tapping the *Control* key and waiting for an animation to complete. The bias against the snapping interface was neutralised.

The third experiment used a new experimental task: object drag-and-drop with grid-snapping assistance. Subjects switched snapping resolutions by

holding the *Control* key with a release-and-reacquisition of the object. Subjects displayed a bias in favour of this snapping behaviour, even when it impaired their performance. All of these results are understood within the context of the model: the objective costs and benefits did not weigh on subjects' preferences as much as the perceived progress towards their task goal.

Two further psychological biases that are not captured by the model – the peak-end rule and duration neglect – were also tested using the methodology. The results showed some support for the effects but require further work.

These empirical findings establish a relationship between behavioural economics and human–computer interaction that can be applied to enhance research and design practice. There are extensive opportunities for further work examining the application of the model and extending it to account for and validate other psychological biases. The forced-choice methodology also has promise for examining nuanced factors influencing user experience.

以上 ———.

WORKS CITED

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, *101*(2), 470–492. doi: 10.1257/aer.101.2.470

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine [Behaviour of rational man in the face of risk: Critique of hypotheses and axioms of the American school]. *Econometrica*, *21*(4), 503–546. doi: 10.2307/1907921

Allais, M. (1979a). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school. In M. Allais & O. Hagen (Eds.), *Expected utility and the Allais paradox* (pp. 27–145). London, England: D. Reidel.

Allais, M. (1979b). The so-called Allais paradox and rational decisions under uncertainty. In M. Allais & O. Hagen (Eds.), *Expected utility and the Allais paradox* (pp. 27–145). London, England: D. Reidel.

Anand, P. (1987). Are the preference axioms really rational? *Theory and Decision*, *23*(2), 189–214. doi: 10.1007/BF00126305

Anand, P. (1993). *Foundations of rational choice under risk*. Oxford, England: Clarendon Press.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365. doi: 10.1037/0003-066X.51.4.355

Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making*, *11*(1), 19–45. doi: 10/cspgqg

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140. doi: 10/crq2tt

Arrow, K. J. (1958). Utilities, attitudes, choices: A review note. *Econometrica*, *26*(1), 1–23. doi: 10.2307/1907381

Arrow, K. J. (1970). *Essays in the theory of risk bearing*. Amsterdam, The Netherlands: North-Holland.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1–26. doi: 10.1146/annurev.psych.52.1.1

Banovic, N., Grossman, T., & Fitzmaurice, G. (2013). The effect of time-based cost of error in target-directed pointing tasks. In *Proceedings of the 2013 ACM annual conference on human factors in computing systems* (pp. 1373–1382). New York, NY: ACM. doi: 10.1145/2466110.2466181

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *The Journal of Economic Perspectives*, *27*(1), 173–195. doi:

10.1257/jep.27.1.173

Barberis, N. C., Huang, M., & Santos, T. (2001). Prospect theory and asset prices. *The Quarterly Journal of Economics*, *116*(1), 1–53. doi: 10/dt96fn

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. doi: 10.1016/0001-6918(80)90046-3

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*(8), 1561–1580. doi: 10/bqmhpq

Bateman, I., Munro, A., Rhodes, B., Starmer, C., & Sugden, R. (1997). A test of the theory of reference-dependent preferences. *The Quarterly Journal of Economics*, *112*(2), 479–505. doi: 10.1162/003355397555262

Battalio, R. C., Kagel, J. H., Rachlin, H., & Green, L. (1981). Commodity-choice behavior with pigeons as subjects. *Journal of Political Economy*, *89*(1), 67–91. doi: 10.1086/260950

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. doi: 10.1037/1089-2680.5.4.323

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, *11*(2), 167–203. doi: 10.1177/1088868307301033

Becker, G. M., & McClintock, C. G. (1967). Value: Behavioral decision theory. *Annual Review of Psychology*, *18*(1), 239–286. doi: 10/d8wwt9

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981. doi: 10.1287/opre.30.5.961

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, *33*(1), 1–27. doi: 10.1287/opre.33.1.1

Bell, D. E., Raiffa, H., & Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision making. In D. E. Bell, R. Howard, & A. Tversky (Eds.), *Decision making* (pp. 9–30). Cambridge, England: Cambridge University.

Ben-Bassat, T., Meyer, J., & Tractinsky, N. (2006). Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction*, *13*(2), 210–234. doi: 10/c767md

Bentham, J. (1789/1948). *An introduction to the principle of morals and legislations*. Oxford, England: Blackwell.

Bergman, O., Tucker, S., Beyth-Marom, R., Cutrell, E., & Whittaker, S. (2009). It's not that important: Demoting personal information of low subjective importance using grayarea. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 269–278). New York, NY: ACM. doi: 10.1145/1518701.1518745

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk (L. Sommer, Trans.). *Econometrica*, *22*(1), 23–36. (Reprinted from *Commentarii academiae scientiarum imperialis Petropolitanae*, 1738, *5*, 175–192) doi: 10.2307/1909829

Bessière, K., Newhagen, J. E., Robinson, J. P., & Shneiderman, B. (2006). A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Computers in Human Behavior*, *22*(6), 941–961. doi: 10/fxb5pv

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, *25*(3), 187–217. doi: 10.1086/209535

Bhavnani, S. K., & John, B. E. (2000). The strategic use of complex computer systems. *Human-Computer Interaction*, *15*(2–3), 107–137. doi: 10/d3j27j

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463–501. doi: 10.1037/0033-295X.115.2.463

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes*, *67*(1), 91–110. doi: 10.1006/obhd.1996.0067

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, *121*(5), 803–843. doi: 10.1086/673885

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409–432. doi: 10.1037/0033-295X.113.2.409

Brave, S., & Nass, C. (2008). Emotion in human-computer interaction. In A. Sears & J. A. Jacko (Eds.), *The human-computer interaction handbook* (pp. 77–92). New York, NY: Lawrence Erlbaum Associates.

Briggs, R. (2014). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2014 ed.). Stanford University. Retrieved from http://plato.stanford.edu/archives/fall2014/entries/rationality-normative-utility/

Bunt, A., Conati, C., & McGrenere, J. (2007). Supporting interface customization using a mixed-initiative approach. In *Proceedings of the 12th international conference on intelligent user interfaces* (pp. 92–101). New York, NY: ACM. doi: 10.1145/1216295.1216317

Burke, C. J. (1953). A brief note on one-tailed tests. *Psychological Bulletin*, *50*(5), 384–387. doi: 10.1037/h0059627

Camerer, C. F. (1995). Individual decision making. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 587–703). Princeton, NJ: Princeton University Press.

Camerer, C. F. (1997). Progress in behavioral game theory. *The Journal of Economic Perspectives*, *11*(4), 167–188. doi: 10.1257/jep.11.4.167

Camerer, C. F. (1998). Bounded rationality in individual decision making. *Experimental Economics*, *1*(2), 163–183. doi: 10.1023/A:1009944326196

Camerer, C. F. (2004). Prospect theory in the wild: Evidence from the field. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 148–161). Princeton, NJ: Princeton University Press.

Camerer, C. F. (2013). Goals, methods, and progress in neuroeconomics. *Annual Review of Economics*, *5*(1), 425–455. doi: 10/bhp6

Camerer, C. F., Babcock, L., Loewenstein, G., & Thaler, R. H. (1997). Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, *112*(2), 407–441. doi: 10.1162/003355397555244

Camerer, C. F., & Loewenstein, G. (2004). Behavioral economics: Past, present, future. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 3–51). Princeton, NJ: Princeton University Press.

Camerer, C. F., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370. doi: 10.1007/BF00122575

Caplin, A., & Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, *116*(1), 55–79. doi: 10.1162/003355301556347

Card, S. K., Moran, T. P., & Newell, A. (1980a). Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive Psychology*, *12*(1), 32–74. doi: 10.1016/0010-0285(80)90003-1

Card, S. K., Moran, T. P., & Newell, A. (1980b, July). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410. doi: 10.1145/358886.358895

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. London, England: Lawrence Erlbaum Associates.

Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In *Interfacing thought: Cognitive aspects of human-computer interaction* (pp. 80–111). Cambridge, MA: MIT Press.

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., & Shneiderman, B. (2004). Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction*, *17*(3), 333–356. doi: 10/bj2rqn

Chapman, G. B. (1998). Similarity and reluctance to trade. *Journal of Behavioral Decision Making*, *11*(1), 47–58. doi: 10/d49rc7

Charman, S. C., & Howes, A. (2003). The adaptive user: An investigation into the cognitive and task constraints on the generation of new methods. *Journal of Experimental Psychology: Applied*, *9*(4), 236–248. doi: 10.1037/1076-898X.9.4.236

Chew, S. H. (1983). A generalization of the quasilinear mean with applications

to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica*, *51*(4), 1065–1092. doi: 10/czjqjr

Chew, S. H., & Waller, W. S. (1986). Empirical tests of weighted utility theory. *Journal of Mathematical Psychology*, *30*(1), 55–72. doi: 10/bn25w2

Chiravirakul, P., & Payne, S. J. (2014). Choice overload in search engine use? In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1285–1294). New York, NY: ACM. doi: 10/bhqc

Choe, E. K., Jung, J., Lee, B., & Fisher, K. (2013). Nudging people away from privacy-invasive mobile apps through visual framing. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2013* (Vol. 8119, pp. 74–91). Berlin/Heidelberg, Germany: Springer. doi: 10.1007/978-3-642-40477-1_5

Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2004). For better or for worse: Default effects and 401(k) savings behavior. In D. A. Wise (Ed.), *Perspectives on the economics of aging* (pp. 81–125). University of Chicago Press. Retrieved from http://www.nber.org/books/wise04-1

Chou, P.-H., Chou, R. K., & Ko, K.-C. (2009). Prospect theory and the risk-return paradox: Some recent evidence. *Review of Quantitative Finance and Accounting*, *33*(3), 193–208. doi: 10.1007/s11156-009-0109-z

Cockburn, A., Quinn, P., & Gutwin, C. (2015). Examining the peak-end effects of subjective experience. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 357–366). New York, NY: ACM. doi: 10.1145/2702123.2702139

Consolvo, S., McDonald, D. W., & Landay, J. A. (2009). Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 405–414). New York, NY: ACM. doi: 10/b8dbk9

Coombs, C. H., Bezembinder, T. G., & Goode, F. M. (1967). Testing expectation theories of decision making without measuring utility or subjective probability. *Journal of Mathematical Psychology*, *4*(1), 72–103. doi: 10.1016/0022-2496(67)90042-9

Coombs, C. H., & Komorita, S. S. (1958). Measuring utility of money through decisions. *The American Journal of Psychology*, *71*(2), 383–389. doi: 10.2307/1420083

Cooper, A., Reimann, R., & Cronin, D. (2007). *About face 3: The essentials of interaction design*. Indianapolis, IN: Wiley.

Cramer, J. S. (2003). *Logit models from economics and other fields*. Cambridge, England: Cambridge University Press. doi: 10.1017/CBO9780511615412

Crawford, V. P., & Meng, J. (2011). New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *The American Economic Review*, *101*(5),

1912–1932. doi: 10.1257/aer.101.5.1912

Cubitt, R., Starmer, C., & Sugden, R. (1998). Dynamic choice and the common ratio effect: an experimental investigation. *The Economic Journal, 108*(450), 1362–1380. doi: 10.1111/1468-0297.00346

Daston, L. (1988). *Classical probability in the enlightenment*. Princeton, NJ: Princeton University Press.

Davidson, D., Suppes, P., & Siegel, S. (1957). *Decision making: An experimental approach*. Stanford, CA: Stanford University Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. doi: 10/dd5xsk

de Fermat, P. (1894). *Œuvres de Fermat* (Vol. 2). Paris: Gauthier-Villars. (Original work published 1654)

de Finetti, B. (1980). La prévision: ses lois logiques, ses sources subjectives (H. E. Kyburg Jr., Trans.). In H. E. Kyburg Jr. & H. E. Smokler (Eds.), *Studies in subjective probabilities* (2nd ed., pp. 53–118). Huntington, NY: Robert E. Krieger. (Reprinted from *Annales de l'institut Henri Poincaré*, 1937, 7[1], 1–68)

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), 315–372. doi: 10/cx2vsx

Diecidue, E., & Wakker, P. P. (2001). On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty*, *23*(3), 281–298. doi: 10/c7bnmj

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology: Applied*, *17*(4), 332–341. doi: 10.1037/a0024354

Dix, A., Finlay, J. E., Abowd, G. D., & Beale, R. (2004). *Human-computer interaction* (3rd ed.). Essex, England: Pearson Education.

Do, A. M., Rupert, A. V., & Wolford, G. (2008). Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review*, *15*(1), 96–98. doi: 10.3758/PBR.15.1.96

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241–253. doi: 10.2307/1266041

Edwards, W. (1953). Probability-preferences in gambling. *The American Journal of Psychology*, *66*(3), 349–364. doi: 10.2307/1418231

Edwards, W. (1954a). Probability-preferences among bets with differing expected values. *The American Journal of Psychology*, *67*(1), 56–67. doi: 10.2307/1418071

Edwards, W. (1954b). The reliability of probability-preferences. *The American Journal of Psychology*, *67*(1), 68–95. doi: 10.2307/1418072

Edwards, W. (1954c). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. doi: 10/crnw6p

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology: General*, *50*(3), 201–214. doi: 10.1037/h0041692

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, *12*(1), 473–498. doi: 10.1146/annurev.ps.12.020161.002353

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review*, *69*(2), 109–135. doi: 10.1037/h0038674

Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, *52*(1), 581–606. doi: 10.1146/annurev.psych.52.1.581

Ehret, B. D. (2002). Learning where to look: Location learning in graphical user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 211–218). New York, NY: ACM. doi: 10.1145/503376.503414

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. doi: 10.1037/0033-295X.85.5.395

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, *32*, 53–88. doi: 10.1146/annurev.ps.32.020181.000413

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, *92*(4), 433–461. doi: 10/c79wh7

Einhorn, H. J., & Hogarth, R. M. (1986). Decision making under ambiguity. *The Journal of Business*, *59*(4, Pt. 2), S225–S250. Retrieved from http://www.jstor.org/stable/2352758

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. doi: 10.2307/1884324

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527. doi: 10.1037/0033-295X.101.3.519

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review*, *67*(4), 269–271. doi: 10.1037/h0048412

Fehr-Duda, H., & Epper, T. (2012). Probability and risk: Foundations and economic implications of probability-dependent risk preferences. *Annual Review of Economics*, *4*(1), 567–593. doi: 10/bhp7

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology*, *91*(5), 797–813. doi: 10.1037/0022-3514.91.5.797

Festinger, L. (1957). *A theory of cognitive dissonance*. New York, NY: Harper & Row.

Findlater, L., & McGrenere, J. (2010). Beyond performance: Feature awareness in personalized interfaces. *International Journal of Human-Computer Studies*, *68*(3), 121–137. doi: 10.1016/j.ijhcs.2009.10.002

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1548–1553. doi: 10.1037/a0020605

Fischer, G. W., Carmon, Z., Ariely, D., & Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, *45*(8), 1057–1075. doi: 10.1287/mnsc.45.8.1057

Fishburn, P. C. (1978). On Handa's "new theory of cardinal utility" and the maximization of expected return. *Journal of Political Economy*, *86*(2), 321–324. doi: 10.1086/260670

Fishburn, P. C. (1979). On the nature of expected utility. In M. Allais & O. Hagen (Eds.), *Expected utility and the Allais paradox* (pp. 243–257). London, England: D. Reidel.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision*, *13*(2), 139–199. doi: 10.1007/BF00134215

Fishburn, P. C. (1986). The axioms of subjective probability. *Statistical Science*, *1*(3), 335–358. doi: 10.1214/ss/1177013611

Fishburn, P. C. (1988). Normative theories of decision making under risk and under uncertainty. In D. E. Bell, R. Howard, & A. Tversky (Eds.), *Decision making* (pp. 78–98). Cambridge, England: Cambridge University Press.

Fisher, R. A. (1954). *Statistical methods for research workers* (11th ed.). New York, NY: Hafner.

Fisher, R. A. (1971). *The design of experiments* (9th ed.). New York, NY: Hafner Press.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movements. *Journal of Experimental Psychology*, *47*(6), 381–391. doi: 10.1037/h0055392

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. San Francisco, CA: Morgan Kaufmann.

Fortin, C., & Rousseau, R. (1998). Interference from short-term memory processing on encoding and reproducing brief durations. *Psychological Research*, *61*(4), 269–276. doi: 10.1007/s004260050031

Fowler, A., Partridge, K., Chelba, C., Bi, X., Ouyang, T., & Zhai, S. (2015). Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 649–658). New York, NY: ACM.

doi: 10.1145/2702123.2702503

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603. doi: 10.2307/2946693

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401. doi: 10.1257/jel.40.2.351

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55. doi: 10.1037/0022-3514.65.1.45

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279–304. doi: 10.1086/256692

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 345–352). New York, NY: ACM. doi: 10.1145/332040.332455

Fu, W.-T., & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science*, *28*(6), 901–935. doi: 10.1016/j.cogsci.2004.03.005

Fu, W.-T., & Gray, W. D. (2006). Suboptimal tradeoffs in information seeking. *Cognitive Psychology*, *52*(3), 195–242. doi: 10/ff5zdz

Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *The American Economic Review*, *96*(5), 1449–1476. doi: 10/dbf4xf

Furnell, S. (2005). Why users cannot use security. *Computers & Security*, *24*(4), 274–279. doi: 10.1016/j.cose.2005.04.003

Gajos, K., Czerwinski, M., Tan, D. S., & Weld, D. S. (2006). Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the working conference on advanced visual interfaces* (pp. 201–208). New York, NY: ACM. doi: 10.1145/1133265.1133306

Gajos, K., & Weld, D. S. (2005). Preference elicitation for interface optimization. In *Proceedings of the 18th annual ACM symposium on user interface software and technology* (pp. 173–182). New York, NY: ACM. doi: 10.1145/1095034.1095063

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. doi: 10.1198/000313006X152649

Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *The Quarterly Journal of Economics*, *116*(4), 1233–1260. doi: 10.1162/003355301753265561

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. doi: 10/dkkzjd

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2006). *Measuring consumer risk-return tradeoffs* [Unpublished Manuscript]. doi: 10/fzbk4z

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166. doi: 10.1006/cogp.1998.0710

Gonzalez, R., & Wu, G. (2003). *Composition rules in old and new prospect theory.* Unpublished manuscript.

Goodman, J., Venolia, G., Steury, K., & Parker, C. (2001). *Language modeling for soft keyboards* (Tech. Rep. No. MSR-TR-2001-118). Microsoft Research. Retrieved from http://research.microsoft.com/pubs/69903/tr-2001-118.pdf

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction*, *8*(3), 237–309. doi: 10/b4n5hn

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*(3), 461–482. doi: 10.1037/0033-295X.113.3.461

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, *95*(3), 537–557. doi: 10.2307/1885092

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, *69*(4), 623–638. Retrieved from http://jstor.org/stable/1808708

Grondin, S. (Ed.). (2008). *Psychology of time.* Bingley, England: Emerald Group.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on privacy in the electronic society* (pp. 71–80). New York, NY: ACM. doi: 10/cqhxwg

Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, *59*(3), 667–686. doi: 10.2307/2938223

Gunaratne, J., & Nov, O. (2015). Informing and improving retirement saving performance using behavioral economics theory-driven user interfaces. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 917–920). New York, NY: ACM. doi: 10/bhqj

Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R. L., & Lafreniere, B. (2016). Peak-end effects on player experience in casual games. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5608–5619). New York, NY: ACM. doi: 10.1145/2858036.2858419

Hansson, S. O., & Grüne-Yanoff, T. (2012). Preferences. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2012 ed.). Stanford University. Retrieved from http://plato.stanford.edu/archives/win2012/

entries/preferences/

Hardie, B. G. S., Johnson, E. J., & Fader, P. S. (1993). Modeling loss aversion and reference dependence effects on brand choice. *Marketing Science, 12*(4), 378–394. doi: 10.1287/mksc.12.4.378

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

Harrison, C., & Hudson, S. E. (2010). Minput: Enabling interaction on small mobile devices with high-precision, low-cost, multipoint optical tracking. In *CHI '10: Proceedings of the 28th international conference on human factors in computing systems* (pp. 1661–1664). New York, NY: ACM. doi: 10.1145/1753326.1753574

Harrison, C., Yeo, Z., & Hudson, S. E. (2010). Faster progress bars: Manipulating perceived duration with visual augmentations. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1545–1548). New York, NY: ACM. doi: 10/c6d27k

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, pp. 139–183). North Holland, The Netherlands: Elsevier. doi: 10/fbms3r

Hartmann, J., De Angeli, A., & Sutcliffe, A. (2008). Framing the user experience: Information biases on website quality judgement. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 855–864). New York, NY: ACM. doi: 10/fktzv4

Hartmann, J., Sutcliffe, A., & De Angeli, A. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction, 15*(4), 15:1–15:30. doi: 10/b57cf9

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction, 19*(4), 319–349. doi: 10.1207/s15327051hci1904_2

Hassenzahl, M. (2005). The thing and I: Understanding the relationship between user and product. In M. A. Blythe, K. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology* (Vol. 3, pp. 31–42). The Netherlands: Springer. doi: 10.1007/1-4020-2967-5_4

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products – facets of user experience. *Interacting with Computers, 22*(5), 353–362. doi: 10/b29bhx

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction, 25*(3), 235–260. doi: 10/cxcs4n

Hastie, R. (2001). Problems for judgement and decision making. *Annual*

Review of Psychology, *52*(1), 653–683. doi: 10/ctfnw7

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, *38*(1), 79–109. doi: 10.1006/cogp.1998.0708

Heidhues, P., & Kőszegi, B. (2008). Competition and price variation when consumers are loss averse. *The American Economic Review*, *98*(4), 1245–1268. doi: 10.1257/aer.98.4.1245

Hershey, J. C., Kunreuther, H. C., & Schoemaker, P. J. H. (1982). Sources of bias in assessment procedures for utility functions. *Management Science*, *28*(8), 936–954. doi: 10.1287/mnsc.28.8.936

Hershey, J. C., & Schoemaker, P. J. H. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance*, *25*(3), 395–418. doi: 10.1016/0030-5073(80)90037-9

Herweg, F., Müller, D., & Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *The American Economic Review*, *100*(5), 2451–2477. doi: 10.1257/aer.100.5.2451

Hick, W. E. (1952). A note on one-tailed and two-tailed tests. *Psychological Review*, *59*(4), 316–318. doi: 10.1037/h0056061

Highhouse, S., & Johnson, M. A. (1996). Gain/loss asymmetry and riskless choice: Loss aversion in choices among job finalists. *Organizational Behavior and Human Decision Processes*, *68*(3), 225–233. doi: 10/c56zzq

Ho, T. H. (2006). Modeling the psychology of consumer and firm behavior with behavioral economics. *Journal of Marketing Research*, *43*(3), 307–331. doi: 10.1509/jmkr.43.3.307

Hogarth, R. M., & Reder, M. W. (Eds.). (1986a). The behavioral foundations of economic theory [Special Issue]. *The Journal of Business*, *59*(4). Retrieved from http://www.jstor.org/stable/i340501

Hogarth, R. M., & Reder, M. W. (1986b). Editors' comments: Perspectives from economics and psychology. *The Journal of Business*, *59*(4, Pt. 2), S185–S207. Retrieved from http://www.jstor.org/stable/2352756

Hollis, V., Konrad, A., & Whittaker, S. (2015). Change of heart: Emotion tracking to promote behavior change. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2643–2652). New York, NY: ACM. doi: 10.1145/2702123.2702196

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. doi: 10.1016/j.ijhcs.2005.06.002

Hornbæk, K. (2015, October). We must be more wrong in HCI research. *interactions*, *22*(6), 20–21. doi: 10.1145/2833093

Hornbæk, K., & Law, E. L.-C. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 617–626). New York, NY: ACM. doi:

10.1145/1240624.1240722

Hornof, A. J. (2004). Cognitive strategies for the visual search of hierarchical computer displays. *Human Computer Interaction*, *19*(3), 183–223. doi: 10.1207/s15327051hci1903_1

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, *310*(5754), 1680–1683. doi: 10.1126/science.1115327

Hurst, P. M., & Siegel, S. (1956). Prediction of decisions from a higher ordered metric scale of utility. *Journal of Experimental Psychology: General*, *52*(2), 138–144. doi: 10.1037/h0040860

Huygens, C. (1714). *De ratiociniis in ludo aleæ* [The value of all chances in games of fortune] (W. Browne, Trans.). London, England: S. Keimer. (Original work published 1657)

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality*, *21*(3), 329–335. doi: 10.1111/j.1467-6494.1953.tb01775.x

Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., & Whittaker, S. (2013). Echoes from the past: How technology mediated reflection improves well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1071–1080). New York, NY: ACM. doi: 10.1145/2470654.2466137

Jameson, A., Berendt, B., Gabrielli, S., Cena, F., Gena, C., Vernero, F., & Reinecke, K. (2014). Choice architecture for human-computer interaction. *Foundations and Trends in Human-Computer Interaction*, *7*(1–2), 1–235. doi: 10.1561/1100000028

John, B. E., & Kieras, D. E. (1996a). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, *3*(4), 320–351. doi: 10.1145/235833.236054

John, B. E., & Kieras, D. E. (1996b). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction*, *3*(4), 287–319. doi: 10.1145/235833.236050

Johnson, E. J., & Goldstein, D. G. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. doi: 10.1126/science.1091721

Johnson, S. G. B., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76. doi: 10.1016/j.cogpsych.2015.01.003

Jokinen, J. P. P. (2015). Emotional user experience: Traits, events, and states. *International Journal of Human-Computer Studies*, *76*, 67–77. doi: 10.1016/j.ijhcs.2014.12.006

Kahneman, D. (2000a). Evaluation by moments: Past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 693–708).

Cambridge, England: Cambridge University Press.

Kahneman, D. (2000b). Experienced utility and objective happiness: A moment-based approach. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 673–692). Cambridge, England: Cambridge University Press.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. doi: 10.1037/0003-066X.58.9.697

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). Cambridge, England: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–294). Cambridge, England: Cambridge University Press.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405. doi: 10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348. doi: 10.1086/261737

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206. doi: 10.1257/jep.5.1.193

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. doi: 10.1037/h0034747

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. doi: 10.2307/1914185

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157. doi: 10.1016/0010-0277(82)90023-3

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350. doi: 10.1037/0003-066X.39.4.341

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, England: Cambridge University Press.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, *112*(2), 375–405. doi: 10.1162/003355397555235

Kalnikaite, V., Rogers, Y., Bird, J., Villar, N., Bachour, K., Payne, S., ... Kreitmayer, S. (2011). How to nudge in situ: Designing lambent devices to deliver salient information in supermarkets. In *Proceedings of the 13th*

international conference on ubiquitous computing (pp. 11–20). New York, NY: ACM. doi: 10.1145/2030112.2030115

Kay, R. (2008). Exploring the relationship between emotions and the acquisition of computer knowledge. *Computers & Education*, *50*(4), 1269–1283. doi: 10.1016/j.compedu.2006.12.002

Keren, G. B., & Raaijmakers, J. G. W. (1988). On between-subjects versus within-subjects comparisons in testing utility theory. *Organizational Behavior and Human Decision Processes*, *41*(2), 233–247. doi: 10/b2nkrd

Keynes, J. M. (1921). *A treatise on probability*. London, England: Macmillan.

Kőszegi, B. (2010). Utility from anticipation and personal equilibrium. *Economic Theory*, *44*(3), 415–444. doi: 10.1007/s00199-009-0465-x

Kőszegi, B., & Rabin, M. (2004). *A model of reference-dependent preferences* (Working Paper No. E04-337). Department of Economics, University of California, Berkeley.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, *121*(4), 1133–1165. doi: 10/cqgpmq

Kőszegi, B., & Rabin, M. (2007). Reference-dependent risk attitudes. *The American Economic Review*, *97*(4), 1047–1073. doi: 10/djbb7t

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*(4), 391–438. doi: 10/dpb5j8

Kieras, D. E., & Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, *22*(4), 365–394. doi: 10.1016/S0020-7373(85)80045-6

Kieras, D. E., Wood, S. D., & Meyer, D. E. (1997). Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task. *ACM Transactions on Computer-Human Interaction*, *4*(3), 230–275. doi: 10.1145/264645.264658

Kim, Y.-H., Jeon, J. H., Choe, E. K., Lee, B., Kim, K., & Seo, J. (2016). TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the SIGCHI conference on human factors in computing systems*. New York, NY: ACM. doi: 10.1145/2858036.2858428

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*(4), 351–353. doi: 10.1037/h0046737

Kissel, G. V. (1995). The effect of computer experience on subjective and objective software usability measures. In *Conference companion on human factors in computing systems* (pp. 284–285). New York, NY: ACM. doi: 10.1145/223355.223676

Klasnja, P., Consolvo, S., & Pratt, W. (2011). How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3063–3072). New

York, NY: ACM. doi: 10.1145/1978942.1979396

Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, *83*(5), 2065–2079. doi: 10.3982/ecta11983

Knetsch, J. L. (1992). Preferences and nonreversibility of indifference curves. *Journal of Economic Behavior & Organization*, *17*(1), 131–139. doi: 10.1016/0167-2681(92)90082-M

Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston, MA: Houghton Mifflin.

Köbberling, V., & Wakker, P. P. (2005). An index of loss aversion. *Journal of Economic Theory*, *122*(1), 119–131. doi: 10.1016/j.jet.2004.03.009

Krantz, D. H., & Kunreuther, H. C. (2007). Goals and plans in decision making. *Judgement and Decision Making*, *2*(3), 137–168.

Kreps, D. (1988). *Notes on the theory of choice*. Boulder, CO: Westview.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23–55. doi: 10.1006/obhd.1998.2781

Kusev, P., van Schaik, P., Ayton, P., Dent, J., & Chater, N. (2009). Exaggerated risk: Prospect theory and probability weighting in risky choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1487–1505. doi: 10.1037/a0017039

Kwatinetz, A., Leblond, A., Peters, G. C., & Hirsch, S. M. (1998). *U.S. Patent No. 5,832,528*. Washington, DC: U.S. Patent and Trademark Office.

Lane, D. M., Napier, H. A., Peres, S. C., & Sándor, A. (2005). Hidden costs of graphical user interfaces: Failure to make the transition from menus and icon toolbars to keyboard shortcuts. *International Journal of Human-Computer Interaction*, *18*(2), 133–144. doi: 10/bqxz2b

Langer, T., Sarin, R., & Weber, M. (2005). The retrospective evaluation of payment sequences: Duration neglect and peak-and-end effects. *Journal of Economic Behavior & Organization*, *58*(1), 157–175. doi: 10/c77nqg

Larrick, R. P., & Boles, T. L. (1995). Avoiding regret in decisions with feedback: A negotiation example. *Organizational Behavior and Human Decision Processes*, *63*(1), 87–97. doi: 10.1006/obhd.1995.1064

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Interaction*, *60*(3), 269–298. doi: 10.1016/j.ijhcs.2003.09.002

Lazar, J., Jones, A., Hackley, M., & Shneiderman, B. (2006). Severity and impact of computer user frustration: A comparison of student and workplace users. *Interacting with Computers*, *18*(2), 187–207. doi: 10/ck4rqb

Lee, M. K., Kiesler, S., & Forlizzi, J. (2011). Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 325–334).

New York, NY: ACM. doi: 10.1145/1978942.1978989

Lejeune, H., & Wearden, J. H. (2009). Vierordt's 'The Experimental Study of the Time Sense' (1868) and its legacy. *European Journal of Cognitive Psychology*, *21*(6), 941–960. doi: 10.1080/09541440802453006

Lelis, S., & Howes, A. (2011). Informing decisions: How people use online rating information to make choices. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2285–2294). New York, NY: ACM. doi: 10.1145/1978942.1979278

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188. doi: 10/bzntst

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology: General*, *89*(1), 46–55. doi: 10.1037/h0031207

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology: General*, *101*(1), 16–20. doi: 10.1037/h0035472

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge, England: Cambridge University Press.

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction*, *18*(1), 1:1–1:30. doi: 10.1145/1959022.1959023

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, *25*(2), 115–126. doi: 10/dqm55n

List, J. A. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics*, *118*(1), 41–71. doi: 10/cgknrg

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, *72*(2), 615–625. doi: 10/bb7vdz

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. doi: 10.1037/0003-066X.57.9.705

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *The Economic Journal*, *97*(387), 666–684. doi: 10.2307/2232929

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science*, *34*(2), 200–214. doi: 10.1287/mnsc.34.2.200

Loewenstein, G., & O'Donoghue, T. (2004). *Animal spirits: Affective and deliberative processes in economic behavior*. doi: 10.2139/ssrn.539843

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, *107*(2),

573–597. doi: 10.2307/2118482

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annual Review of Psychology*, *59*(1), 647–672. doi: 10/bwbjmk

Loomes, G. (1988). Further evidence of the impact of regret and disappointment in choice under uncertainty. *Economica*, *55*(217), 47–62. doi: 10.2307/2554246

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824. doi: 10.2307/2232669

Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, *53*(2), 271–282. doi: 10.2307/2297651

Loomes, G., & Sugden, R. (1987). Testing for regret and disappointment in choice under uncertainty [Supplement: Conference Papers]. *The Economic Journal*, *97*, 118–129. doi: 10.2307/3038234

Lopes, L. L. (1994). Psychology and economics: Perspectives on risk, cooperation, and the marketplace. *Annual Review of Psychology*, *45*(1), 197-227. doi: 10.1146/annurev.ps.45.020194.001213

Louviere, J. J. (1991). Experimental choice analysis: Introduction and overview. *Journal of Business Research*, *23*(4), 291–297. doi: 10/c27xhw

Lovell, M. C. (1986). Tests of the rational expectations hypothesis. *The American Economic Review*, *76*(1), 110–124. Retrieved from http://jstor.org/stable/1804130

Luce, R. D. (1958). A probabilistic theory of utility. *Econometrica*, *26*(2), 193–224. doi: 10.2307/1907587

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. III). New York, NY: John Wiley & Sons.

MacCrimmon, K. R., & Larsson, S. (1979). Utility theory: Axioms versus 'paradoxes'. In M. Allais & O. Hagen (Eds.), *Expected utility and the Allais paradox* (pp. 333–409). London, England: D. Reidel.

Machina, M. J. (1982). "Expected utility" analysis without the independence axiom. *Econometrica*, *50*(2), 277–323. doi: 10.2307/1912631

Machina, M. J. (1987). Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, *1*(1), 121–154. doi: 10/bhqf

Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, *27*(4), 1622–1668. Retrieved from http://www.jstor.org/stable/2727025

Mackay, W. E. (1990). Patterns of sharing customizable software. In *Proceedings of the 1990 ACM conference on computer-supported cooperative work* (pp. 209–221). New York, NY: ACM. doi: 10.1145/99332.99356

Mackay, W. E. (1991). Triggers and barriers to customizing software. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 153–160). New York, NY: ACM. doi: 10.1145/108844.108867

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction*, *7*(1), 91–139. doi: 10.1207/s15327051hci0701_3

MacKenzie, I. S., & Soukoreff, R. W. (2003). Phrase sets for evaluating text entry techniques. In *CHI '03 extended abstracts on human factors in computing systems* (pp. 754–755). New York, NY: ACM. doi: 10/fttfp5

Manktelow, K. I., & Over, D. E. (Eds.). (1993). *Rationality: Psychological and philosophical perspectives*. London, England: Routledge.

March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics*, *9*(2), 587–608. doi: 10/dwb5xv

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*(2), 151–158. doi: 10.1086/257177

Marks, M. R. (1951). Two kinds of experiment distinguished in terms of statistical operations. *Psychological Review*, *58*(3), 179–184. doi: 10/fdqsxx

Matlin, M. W., & Stang, D. J. (1978). *The Pollyanna Principle: Selectivity in language, memory, and thought*. Cambridge, MA: Schenkman.

McAfee, R. P., Mialon, H. M., & Mialon, S. H. (2010). Do sunk costs matter? *Economic Inquiry*, *48*(2), 323–336. doi: 10/bmg3cm

McAuley, J. D. (1995). *Perception of time as phase: Toward and adaptive-oscillator model of rhythmic pattern processing* (Unpublished doctoral dissertation). Indiana University, Bloomington, IN.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. University of Minnesota Press.

Mellers, B. A., Schwartz, A., & Cooke, A. D. J. (1998). Judgement and decision making. *Annual Review of Psychology*, *49*(1), 447-477. doi: 10/dkztgm

Mercer, J. (2005). Prospect theory and political science. *Annual Review of Political Science*, *8*(1), 1–21. doi: 10/d5wfcq

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., & Smith, J. E. K. (1988). Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review*, *95*(3), 340–370. doi: 10/d5ms4n

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*(1), 3–65. doi: 10/fwrwd8

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*(4), 749–791. doi: 10/fqbd6z

Morgenstern, O. (1972). Descriptive, predictive and normative theory. *Kyklos*,

25(4), 699–714. doi: 10.1111/j.1467-6435.1972.tb01077.x

Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica*, *29*(3), 315–335.

Nasiry, J., & Popescu, I. (2011). Dynamic pricing with loss-averse consumers and peak-end anchoring. *Operations Research*, *59*(6), 1361–1368. doi: 10.1287/opre.1110.0952

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Interaction*, *45*(6), 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. doi: 10/cqzrs6

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, *29*(5), 1093–1109. doi: 10/dpdm9k

Nicosia, M., Oulasvirta, A., & Kristensson, P. O. (2014). Modeling the perception of user performance. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems* (pp. 1747–1756). New York, NY: ACM. doi: 10.1145/2556288.2557355

Nielsen, J., & Levy, J. (1994, April). Measuring usability: Preference vs. performance. *Communications of the ACM*, *37*(4), 66–75. doi: 10/bt8736

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. doi: 10.1038/nn.2886

Norman, D. A. (1983). Design principles for human-computer interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1–10). New York, NY: ACM. doi: 10.1145/800045.801571

Norman, D. A. (2005, July). Human-centered design considered harmful. *interactions*, *12*(4), 14–19. doi: 10.1145/1070960.1070976

Norman, D. A. (2006, November). Logic versus usage: The case for activity-centered design. *interactions*, *13*(6), 45–63. doi: 10/fdg6ph

Novemsky, N., & Kahneman, D. (2005). The boundaries of loss aversion. *Journal of Marketing Research*, *42*(2), 119–128. doi: 10/cvmz9d

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, *17*(4), 460–469. doi: 10.2307/3150499

Oliver, R. L., & Winer, R. S. (1987). A framework for the formation and structure of consumer expectations: Review and propositions. *Journal of Economic Psychology*, *8*(4), 469–499. doi: 10.1016/0167-4870(87)90037-7

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*(1), 277–294. doi: 10.1146/annurev-psych-010814-015148

Oulasvirta, A., Hukkinen, J. P., & Schwartz, B. (2009). When more is less: The paradox of choice in search engine use. In *Proceedings of the 32nd international acm SIGIR conference on research and development in information retrieval* (pp. 516–523). New York, NY: ACM. doi: 10/d7zbvn

Page, S. R., Johnsgard, T. J., Albert, U., & Allen, C. D. (1996). User customization of a word processor. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 340–346). New York, NY: ACM. doi: 10.1145/238386.238541

Payne, S. J., & Green, T. R. G. (1986). Task-action grammars: A model of the mental representation of task languages. *Human-Computer Interaction*, *2*(2), 93–133. doi: 10.1207/s15327051hci0202_1

Payne, S. J., & Howes, A. (2013). *Adaptive interaction: A utility maximization approach to understanding human interaction with technology*. San Rafael, CA: Morgan & Claypool. doi: 10.2200/S00479ED1V01Y201302HCI016

Peeters, G. (1971). The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology*, *1*(4), 455–474. doi: 10.1002/ejsp.2420010405

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, *1*(1), 33–60. doi: 10/b8jxpn

Picard, R. W., & Klien, J. (2002). Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers*, *14*(2), 141–169. doi: 10.1016/S0953-5438(01)00055-8

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675. doi: 10.1037/0033-295X.106.4.643

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual Review of Psychology*, *35*(1), 139–164. doi: 10/ctxm3d

Plott, C. R. (1986). Rational choice in experimental markets. *The Journal of Business*, *59*(4, Pt. 2), S301–S327. Retrieved from http://www.jstor.org/stable/2352762

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1–2), 122–136. doi: 10/b2vjdj

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1964). The foundations of decision under uncertainty: An elementary exposition. *Journal of the American Statistical Association*, *59*(306), 353–375. doi: 10/bhqp

Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*(3), 497–527. doi: 10.2307/2998573

Prelec, D. (2000). Compound invariant weighting functions in prospect theory. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 67–92). Cambridge, England: Cambridge University Press.

Preston, M. G., & Baratta, P. (1948). An experimental study of the auction-value of an uncertain outcome. *The American Journal of Psychology*, *61*(2), 183–193. doi: 10.2307/1416964

Putler, D. S. (1992). Incorporating reference price effects into a theory of consumer choice. *Marketing Science*, *11*(3), 287–309. doi: 10/bs62p5

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343. doi: 10.1016/0167-2681(82)90008-7

Quinn, P., Cockburn, A., & Delamarche, J. (2013). Examining the costs of multiple trajectory pointing techniques. *International Journal of Human-Computer Studies*, *71*(4), 492–509. doi: 10.1016/j.ijhcs.2012.12.001

Quinn, P., Cockburn, A., Räihä, K.-J., & Delamarche, J. (2011). On the costs of multiple trajectory pointing methods. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 859–862). New York, NY: ACM. doi: 10.1145/1978942.1979067

Quinn, P., & Zhai, S. (2016). A cost–benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 83–88). New York, NY: ACM. doi: 10/bhqr

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, *83*(5), 1281–1302. doi: 10/cr95qb

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, *36*(1), 11–46. Retrieved from http://www.jstor.org/stable/2564950

Rabin, M. (2000a). Diminishing marginal utility of wealth cannot explain risk aversion. In D. Kahneman & A. Tversky (Eds.), *Values, choices, and frames* (pp. 202–208). Cambridge, England: Cambridge University Press.

Rabin, M. (2000b). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292. doi: 10/c77qd6

Rabin, M. (2013). An approach to incorporating psychology into economics. *The American Economic Review*, *103*(3), 617–622. doi: 10/bhp8

Raiffa, H. (1961). Risk, ambiguity, and the Savage axioms: Comment. *The Quarterly Journal of Economics*, *75*(4), 690–694. doi: 10.2307/1884326

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London, England: Kegan, Paul, Trench, Trubner & Co. (Original work published 1926)

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. doi: 10/czjgb4

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain*, *104*(1–2), 187–194. doi: 10/dhs3ms

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56. doi: 10.1037/a0021150

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review*, *119*(2), 408–416. doi: 10.1037/a0027372

Roberts, H. V. (1963). Risk, ambiguity, and the Savage axioms: Comment. *The Quarterly Journal of Economics*, *77*(2), 327–336. doi: 10.2307/1884408

Rode, E., Rozin, P., & Durlach, P. (2007). Experienced and remembered pleasure for meals: Duration neglect but minimal peak, end (recency) or primacy effects. *Appetite*, *49*(1), 18–29. doi: 10.1016/j.appet.2006.09.006

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. doi: 10.1207/S15327957PSPR0504_2

Salvucci, D. D., & Lee, F. J. (2003). Simple cognitive modeling in a complex cognitive architecture. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 265–272). New York, NY: ACM. doi: 10.1145/642611.642658

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies*, *4*(2), 155–161. doi: 10.2307/2967612

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. doi: 10.1007/BF00055564

Savage, L. J. (1962). Subjective probability and statistical practice. In *The foundations of statistical inference* (pp. 9–35). London, England: Methuen & Co.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York, NY: Dover.

Schenkman, B. N., & Jönsson, F. U. (2000). Aesthetics and preferences of web pages. *Behaviour & Information Technology*, *19*(5), 367–377. doi: 10.1080/014492900750000063

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, *57*(3), 571–587. doi: 10.2307/1911053

Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis*. Champaign, IL: Human Kinetics.

Schoemaker, P. J. H. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, *20*(2), 529–563.

Schreiber, C. A., & Kahneman, D. (2000). Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology: General*, *129*(1), 27–42. doi: 10.1037/0096-3445.129.1.27

Schumm, G. F. (1987). Transitivity, preference and indifference. *Philosophical Studies*, *52*(3), 435–437. doi: 10.1007/BF00354058

Schwartz, B. (2004). *The paradox of choice: Why less is more*. New York, NY:

Ecco.

Sedley, A., & Müller, H. (2013). Minimizing change aversion for the Google Drive launch. In *CHI '13 extended abstracts on human factors in computing systems* (pp. 2351–2354). New York, NY: ACM. doi: 10/bhqh

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556. doi: 10.3758/BF03197186

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491–517. doi: 10.1146/annurev.psych.53.100901.135213

Shafir, E., & Tversky, A. (1995). Decision making. In E. E. Smith & D. N. Osherson (Eds.), *An invitation to cognitive science: Thinking* (2nd ed., Vol. 3, pp. 77–100). Cambridge, MA: MIT Press.

Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research*, *7*(2), 99–111. doi: 10.1086/208799

Siegel, S. (1956). A method for obtaining an ordered metric scale. *Psychometrika*, *21*(2), 207–216. doi: 10.1007/BF02289100

Simon, H. A. (1957). A behavioral model of rational choice. In *Models of man: Social and rational* (pp. 241–260). Boston, MA: John Wiley & Sons.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *The American Economic Review*, *49*(3), 253–283. Retrieved from http://jstor.org/stable/1809901

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, *68*(2), 1–16. Retrieved from http://jstor.org/stable/1816653

Simon, H. A. (1986). Rationality in psychology and economics. *The Journal of Business*, *59*(4, Pt. 2), S209–S224. Retrieved from http://www.jstor.org/stable/2352757

Slovic, P. (1966). Value as a determiner of subjective probability. *IEEE Transactions on Human Factors in Electronics*, *HFE-7*(1), 22–28. doi: 10/b828c3

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, *28*, 1–39. doi: 10/cc25w7

Slovic, P., & Lichtenstein, S. (1968). Relative importance of probabilities and payoffs in risk taking [Monograph]. *Journal of Experimental Psychology*, *78*(3, Pt. 2), 1–18. doi: 10.1037/h0026468

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *The American Economic Review*, *73*(4), 596–605. Retrieved from http://www.jstor.org/stable/1816560

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science*, *19*(6), 368–373. doi: 10.1002/bs.3830190603

Smith, A. (1976). *The theory of moral sentiments* (D. D. Raphael & A. L. Macfie, Eds.). Oxford, England: Clarendon Press. (Original work published

1759)

Solomon, J. (2014). Customization bias in decision support systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3065–3074). New York, NY: ACM. doi: 10.1145/2556288.2557211

Spector, A. J. (1956). Expectations, fulfillment, and morale. *The Journal of Abnormal and Social Psychology*, *52*(1), 51–56. doi: 10.1037/h0047881

Sprenger, C. (2010). *An endowment effect for risk: Experimental tests of stochastic reference points* (Tech. Rep.). University of California San Diego.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. doi: 10.1017/S0140525X00003435

Starmer, C. (1999). Cycling with rules of thumb: An experimental test for a new form of non-transitive behaviour. *Theory and Decision*, *46*(2), 139–157. doi: 10.1023/A:1004930205037

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332–382. doi: 10.1257/jel.38.2.332

Starmer, C., & Sugden, R. (1989). Violations of the independence axion in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operations Research*, *19*(1), 79–102. doi: 10/dqqz4f

Starmer, C., & Sugden, R. (1993). Testing for juxtaposition and event-splitting effects. *Journal of Risk and Uncertainty*, *6*(3), 235–254. doi: 10/bkhr9c

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881–911. doi: 10/fqx3hc

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. doi: 10.1016/j.cogpsych.2005.10.003

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General*, *132*(1), 23–46. doi: 10.1037/0096-3445.132.1.23

Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*(3), 687–705. doi: 10/bhqs

Stigler, G. J. (1950a). The development of utility theory. I. *Journal of Political Economy*, *58*(4), 307–327. Retrieved from http://www.jstor.org/stable/1828885

Stigler, G. J. (1950b). The development of utility theory. II. *Journal of Political Economy*, *58*(5), 373–396. Retrieved from http://www.jstor.org/stable/1825710

Sunstein, C. R. (2014). *Why nudge?* New Haven, CT: Yale University Press.

Sutcliffe, A. (2000). On the effective use and reuse of HCI knowledge.

ACM Transactions on Computer-Human Interaction, *7*(2), 197–221. doi: 10.1145/353485.353488

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67–85. doi: 10.1037/0033-2909.110.1.67

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *7*(1), 39–60. doi: 10/cmjq7d

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*(3), 199–214. doi: 10.1287/mksc.4.3.199

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206. doi: 10/d4njp3

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. New Haven, CT: Yale University Press.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2013). Choice architecture. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 428–439). Princeton, NJ: Princeton University Press.

Thielsch, M. T., Engel, R., & Hirschfeld, G. (2015). Expected usability is not a valid indicator of experienced usability. *PeerJ Computer Science*, *1*, e19. doi: 10.7717/peerj-cs.19

Todd, P. M., Rogers, Y., & Payne, S. J. (2011). Nudging the trolley in the supermarket: How to deliver the right information to shoppers. *International Journal of Mobile Human Computer Interaction*, *3*(2), 20–34. doi: 10.4018/jmhci.2011040102

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518. doi: 10.1126/science.1134239

Toomim, M., Kriplean, T., Pörtner, C., & Landay, J. (2011). Utility of human-computer interactions: toward a science of preference measurement. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 2275–2284). New York, NY: ACM. doi: 10/c87fzn

Tractinsky, N. (2004). A few notes on the study of beauty in HCI. *Human-Computer Interaction*, *19*(4), 351–357. doi: 10.1207/s15327051hci1904_3

Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies*, *64*(11), 1071–1083. doi: 10.1016/j.ijhcs.2006.06.009

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. doi: 10/c589n2

Treisman, M., Faulkner, A., Naish, P. L. N., & Brogan, D. (1990). The internal clock: Evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. *Perception*, *19*(6),

705–742. doi: 10.1068/p190705

Tuch, A. N., Trusell, R., & Hornbæk, K. (2013). Analyzing users' narratives to understand experience with interactive products. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2079–2088). New York, NY: ACM. doi: 10.1145/2470654.2481285

Turland, J., Coventry, L., Jeske, D., Briggs, P., & van Moorsel, A. (2015). Nudging towards security: Developing an application for wireless network selection for Android phones. In *Proceedings of the 2015 British HCI conference* (pp. 193–201). New York, NY: ACM. doi: 10.1145/2783446.2783588

Tversky, A. (1967a). Additivity, utility, and subjective probability. *Journal of Mathematical Psychology*, *4*(2), 175–201. doi: 10/dq57nn

Tversky, A. (1967b). Utility theory and additivity analysis of risky choices. *Journal of Experimental Psychology*, *75*(1), 27–36. doi: 10.1037/h0024915

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48. doi: 10.1037/h0026750

Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 713–717. doi: 10.1037/0278-7393.9.4.713

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*(2), 269–283. doi: 10.1037/0033-295X.102.2.269

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. doi: 10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi: 10/gwh

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. doi: 10/fj3z3r

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. doi: 10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, *59*(4, Pt. 2), S251–S278. Retrieved from http://www.jstor.org/stable/2352759

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. doi: 10.2307/2937956

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. doi: 10.1007/BF00122574

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*(3), 371–384. doi: 10/hbk

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference

reversal. *The American Economic Review*, *80*(1), 204–217. Retrieved from http://www.jstor.org/stable/2006743

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science*, *22*(2), 253–260. doi: 10.1177/0956797610396225

van Schaik, P., & Ling, J. (2008). Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers*, *20*(3), 419–432. doi: 10.1016/j.intcom.2008.03.001

van Schaik, P., & Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies*, *67*(1), 79–89. doi: 10.1016/j.ijhcs.2008.09.012

van Schie, E. C. M., & van der Pligt, J. (1995). Influencing risk preference in decision making: The effects of framing and salience. *Organizational Behavior and Human Decision Processes*, *63*(3), 264–275. doi: 10/cxxsfq

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making*, *5*(3), 169–185. doi: 10.1002/bdm.3960050303

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, *11*(4), 342–365. doi: 10/cdt

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, *15*(11), 546–554. doi: 10.1016/j.tics.2011.09.008

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

von Vierordt, K. (1868). *Der zeitsinn nach versuchen* [Empirical studies of time experience]. Tübingen, Germany: H. Laupp.

Wakker, P. P. (2015). *Annotated references on decisions and uncertainty*. Unpublished manuscript. Retrieved from http://people.few.eur.nl/wakker/refs/webrfrncs.pdf

Wakker, P. P., & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, *7*(2), 147–175. doi: 10/bx7vgk

Wallsten, T. S. (1970). Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology*, *88*(1), 31–40. doi: 10.1037/h0030669

Wang, Y., Leon, P. G., Acquisti, A., Cranor, L. F., Forget, A., & Sadeh, N. (2014). A field trial of privacy nudges for Facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2367–2376). New York, NY: ACM. doi: 10.1145/2556288.2557413

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Physiology*, *20*(3), 273–281. doi: 10.1080/14640746808400161

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*(1), 53–85. doi: 10/b9sk3g

Weber, M., & Camerer, C. F. (1998). The disposition effect in securities trading: an experimental analysis. *Journal of Economic Behavior & Organization*, *33*(2), 167–184. doi: 10.1016/S0167-2681(97)00089-9

Wedell, D. H., & Böckenholt, U. (1990). Moderation of preference reversals in the long run. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 429–438. doi: 10.1037/0096-1523.16.2.429

Weir, D., Pohl, H., Rogers, S., Vertanen, K., & Kristensson, P. O. (2014). Uncertain text entry on mobile devices. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2307–2316). New York, NY: ACM. doi: 10.1145/2556288.2557412

Weybrew, B. B. (1984). The Zeigarnik phenomenon revisited: Implications for enhancement of morale. *Perceptual and Motor Skills*, *58*, 223–226. doi: 10.2466/pms.1984.58.1.223

Wilfong, J. D. (2006). Computer anxiety and anger: The impact of computer use, computer experience, and self-efficacy beliefs. *Computers in Human Behavior*, *22*(6), 1001–1011. doi: 10.1016/j.chb.2004.03.020

Williams, A. W. (1987). The formation of price forecasts in experimental markets. *Journal of Money, Credit and Banking*, *19*(1), 1–18. doi: 10/bq8v7b

Woodworth, R. S. (1899). The accuracy of voluntary movement [Monograph]. *Psychological Review*, *3*(2), 1–114. doi: 10.1037/h0092992

Wright, C. E., & Lee, F. (2013). Issues related to HCI application of Fitts's law. *Human-Computer Interaction*, *28*(6), 548–578. doi: 10/bhp9

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, *42*(12), 1676–1690. doi: 10/c9vpd7

Wu, G., & Gonzalez, R. (1998). Common consequence conditions in decision making under risk. *Journal of Risk and Uncertainty*, *16*(1), 115–139. doi: 10.1023/A:1007714509322

Wu, G., Heath, C., & Larrick, R. (2008). *A prospect theory model of goal behavior*. Unpublished manuscript.

Yarmey, A. D. (2000). Retrospective duration estimations for variant and invariant events in field situations. *Applied Cognitive Psychology*, *14*(1), 45–57. doi: 10/dt56gg

Yi, Y. (1990). A critical review of consumer satisfaction. In V. A. Zeithaml (Ed.), *Review of marketing 1990* (pp. 68–123). Chicago, IL: American Marketing Association.

Zhang, Y., Bellamy, R. K., & Kellogg, W. A. (2015). Designing information for remediating cognitive biases in decision-making. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2211–2220). New York, NY: ACM. doi: 10.1145/2702123.2702239

Zhao, J. C., Fu, W.-T., Zhang, H., Zhao, S., & Duh, H. (2015). To risk or not to risk?: Improving financial risk taking of older adults by online social information. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 95–104). New York, NY: ACM. doi: 10.1145/2675133.2685033

EXPERIMENT CONSENT FORM

We (Philip Quinn and Andy Cockburn) are carrying out an investigation into user interface preferences. The experiment involves pointing to and dragging targets using a computer mouse. In total, your participation should take approximately fifteen minutes.

This experiment has been approved by the University of Canterbury Human Ethics Committee as part of program code HEC 2015/11/LR-PS.

This is not in any way a test of your competence with computers.

All references to participants in the investigation will be anonymous.

Thank you for your co-operation.

If you have any questions regarding this investigation, please contact us.

Philip Quinn
philip.quinn@canterbury.ac.nz

Andy Cockburn
andy@cosc.canterbury.ac.nz

I consent to act as a participant in an experiment that involves pointing to and dragging targets using a computer mouse. I agree to let the resulting data be used for analysis and presentation subject to the conditions below:

- only Philip Quinn and Andy Cockburn will have access to my data and be able to identify me from it;
- data presented or published will be stripped of my identity;
- I retain the right to stop my role as a participant at any time without question, and to have my data discarded.

Name:

Usercode:

Age:

Signature:

Date:

CHI 2016 PUBLICATION: TEXT ENTRY B

The following paper was presented at the ACM CHI conference in 2016. It describes an analysis of user interaction with text entry suggestions under varying levels of interface *assertiveness* (the frequency with which a user is interrupted with suggestions). The results found that although suggestions significantly impaired objective text entry performance, they were subjectively enjoyed by users. This can be understood within the context of the model presented in this thesis, but the work was not conducted or analysed under the same paradigm. The inceptive and experimental work occurred during a three-month internship at Google Inc. in 2012 under the supervision of Dr. Shumin Zhai; subsequent analysis of the data and authorship of the paper was completed by myself at the University of Canterbury in 2015.

> Quinn, P., & Zhai, S. (2016). A cost–benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 83–88). New York, NY: ACM. doi: 10.1145/2858036.2858305.

A Cost–Benefit Study of Text Entry Suggestion Interaction

Philip Quinn[1,2] **Shumin Zhai**[1]

[1]Google Inc.
Mountain View, CA, USA

[2]University of Canterbury
Christchurch, New Zealand

philip@quinn.gen.nz, zhai@google.com

ABSTRACT

Mobile keyboards often present error corrections and word completions (*suggestions*) as candidates for anticipated user input. However, these suggestions are not cognitively free: they require users to attend, evaluate, and act upon them. To understand this trade-off between suggestion savings and interaction costs, we conducted a text transcription experiment that controlled *interface assertiveness*: the tendency for an interface to present itself. Suggestions were either always present (*extraverted*), never present (*introverted*), or gated by a probability threshold (*ambiverted*). Results showed that although increasing the assertiveness of suggestions reduced the number of keyboard actions to enter text and was subjectively preferred, the costs of attending to and using the suggestions impaired average time performance.

Author Keywords

Interface assertiveness; predictive interfaces; mobile text entry.

ACM Classification

H.1.2 [**User/Machine Systems**] Human factors.

INTRODUCTION

A technique frequently employed to improve user performance is to observe a user's interactions, and detect potential errors or predict intended interactions from their behaviour. Corrections for these errors and candidates for these intentions (collectively *suggestions*) can be offered to the user as shortcuts that optimise away interactions the system can anticipate. However, although these suggestions are frequently employed, their *costs* and *benefits* are often not systematically studied [cf. 25]. That is, are the potential benefits of a suggestion worth the costs of interrupting the user, the time and effort for them to process it, decide if it is accurate, and ultimately accept or reject it?

A particularly intense use of suggestion interfaces is with mobile touch-keyboard interaction, where there are many opportunities to adapt the keyboard interface to enhance user performance [e.g. 5, 22, 32, 33]. Text entry suggestions attempt to relieve the burden of precise and tedious character input by offering predictions of the user's intended words or corrections for errors in words that have already been entered. The goal is to reduce the number of keystrokes or corrective actions required to compose text. However, their use is not perceptually or cognitively free: users must identify and attend to the display of the suggestions, evaluate their correctness, and possibly accept or reject them [13, 15].

A key consideration for these interfaces is ultimately *when* to present suggestions to the user. Many systems maintain an internal confidence value associated with each prediction (determined from a language model or dictionary) and only present those above a certain threshold. However, an aspect that is often overlooked is the overall *pay-off* or *utility* to the user – that is, does the user benefit over continuing to type the word or correcting the error manually? Benefits include the time from tapping on the requisite keys, but must be reconciled with the costs of attending and interacting with the interface.

In this paper, we term the tendency for an interface to present suggestions to a user its *assertiveness*, and examine its influence on user performance and satisfaction. Using such a control may improve user experience by reducing the amount of distraction and maximising the benefit of the suggestions. We examine the costs and benefits of text-entry suggestion interaction in an experiment that examines objective text-entry performance and subjective preference at different levels of assertiveness for word-completion suggestions. We find that increasing assertiveness decreases time performance but increases input economy and user preference; we discuss the implications for suggestion interface design.

BACKGROUND

Mobile soft keyboards are rendered and interacted with directly on a display, using either a finger on a touch-sensitive surface or a stylus with a digitiser [19, 21]. However, compared to their physical counterparts, the lack of strong tactile feedback limits users' ability to touch type [2, 10, 23, 28, 32].

Suggestion Systems

Prediction and error correction systems aim to reduce the number of actions required from a user to enter their desired text [4, 7, 8, 24, 31, 32]. Augmentative and Alternative Communication (AAC) research has investigated prediction systems to improve input speed for users with physical impairments, but has found that the benefits of suggestions are not always clear [8, 31]. Koester and Levine [13–15] suggested that the cognitive and motor costs of using suggestion interfaces may outweigh their benefits, and their evaluations of several AAC

CHI'16, May 7–12, 2016, San Jose, CA, USA.

ACM 978-1-4503-3362-7/16/05. . . $15.00.
DOI: http://dx.doi.org/10.1145/2858036.2858305

systems found either marginal performance gains for novice users, or performance impairments for experienced users.

Current commercial practice makes use of suggestion interfaces by default. Google's Android platform (version 5 'Lollipop') persistently presents up to three suggestions above the keyboard. In contrast, Apple's iOS platform (version 9) can display one suggestion below the text insertion cursor, or optionally, three in a bar above the keyboard. The suggestion below the cursor is transient, and is accepted by tapping on the spacebar (tapping on the suggestion will reject it).

Models of Interaction

Interaction with soft keyboards has been successfully modelled using elemental cognitive models (such as GOMS/KLM [3]) in a chain of key-tapping actions [e.g. 18, 29, 30]. These models have been extended to cover the actions involved in using suggestion interfaces [e.g. 11, 14, 17]. However, such models predict performance for expert users and focus on the mechanical aspects of interaction – with limited scope for the cognitive processes or novice users [cf. 1]. Models of attention [e.g. 12] capture issues of interruption and distraction, but have focussed on distractions from unrelated tasks (e.g. email notifications) rather than potentially helpful suggestions for the current task. However, recent work has begun to explore the interaction between a system's attempts at assistance and subjective user preferences [25].

INTERFACE ASSERTIVENESS

Suggestions are a problem for interaction design because of their probabilisitic nature: they are a *guess* about a user's intentions. Incorrect guesses will waste a user's time and lead them to spurn the interface (disabling or habitually ignoring its suggestions). To create a productive experience, it is therefore desirable to be maximally beneficial when suggestions are correct, with minimal costs when they are incorrect.

To use a suggestion, a user must: (a) recognise the presence of the suggestion(s), (b) evaluate the options presented, and (c) optionally take action upon them [13]. As the cost of these actions may be high and provide little benefit (distracting the user for potentially incorrect suggestions), determining precisely *when* to show suggestions should be assiduously considered. For example, suggestions may only be shown if they pass some confidence threshold (typically, according to an *a posteriori* estimation from a language model). In these cases, the user will only need to attend to the suggestions that system decides they are worth the cost. Similarly, a system may mediate suggestions with an H-metaphor control system that is responsive to changes in user behaviour [6].

A method for assessing these costs is the *utility* of the suggestions to be presented: a consideration of system's confidence that the suggestion is correct, weighted with the predicted time savings offered to the user (e.g. time saved tapping keys or editing the text) against the costs of attending to the suggestion. For example, completing the last character of a word does not offer much utility to a user as they may be able to hit the key for that character faster than they could attend to the suggestion interface, evaluate the suggestion, and confirm it.

Suggestion Presentation

The quality of suggestions depends on the system's method for generating them – such as a language model, adaptation from a user's behaviour [7], or characteristics of their input [6, 32]. However, while the quality of suggestions is a key factor in user performance, so too is how the system communicates them and how users interact with them.

Location. Where does the interface display suggestions? The user's awareness and attention to certain parts of the screen may influence whether they notice the suggestions, and alter the interaction requirements for using with them.

Presence. How many suggestions are shown at once? Each suggestion presented increases cognitive costs, as the user must evaluate its correctness. Showing only the top suggestion reduces this, but at the risk of ignoring good secondary options.

Default Behaviour. Are suggestions with a high probability of being correct automatically accepted? This eliminates the cost of explicitly accepting a suggestion, but at the cost of requiring the user to explicitly reject incorrect suggestions or manually correct accidental acceptances.

Stability. Are suggestions continually changing, or are they stable and minimise visual disruptions? Similarly, does the system learn from the user's use of them to help develop automaticity? High stability promotes rapid use of the interface and reduces cognitive costs, but risks hiding good suggestions.

These factors may in-turn influence the underlying model for choosing suggestions. For example, if only a single suggestion is presented, the system may want to ensure that there is a large confidence gap between the first and remaining options. Similarly, if the user has to explicitly reject poor suggestions, then the confidence threshold may be higher.

EXPERIMENT

We conducted an experiment to better understand the trade-offs involved in text entry suggestions in terms of user performance and subjective preference. Subjects copied simple phrases, with word completions shown at opportune times. Completions were ranked by a probability function that emulated current practice, and were presented to subjects in one of three assertiveness conditions: always (*extraverted*), never (*introverted*), and gated by a threshold (*ambiverted*).

Subjects & Apparatus

Seventeen (five female) volunteer subjects participated in the experiment; all had experience typing on mobile keyboards (from various vendors) and received a gift card for their time.

The experiment was run on an Apple iPod Touch (A1367) under iOS 5.1.1 (640×960 px screen at 326 ppi). All visual ornaments were hidden and default text editing features suppressed. Feedback was identical to the English iOS keyboard, but with the modifier keys removed (Figure 1). Letter keys had a visual size of 52×76 px, but their targetable area was expanded to consume the surrounding space ($\sim 65 \times 108$ px).

Stimuli & Tasks

Subjects copied phrases using the keyboard, with up to three suggestions occasionally appearing directly above it (Figure 1).

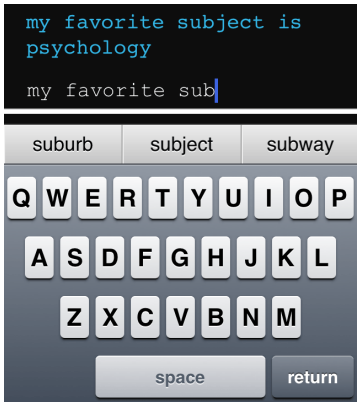


Figure 1. The experimental interface.

Any combination of correct letter keys and helpful suggestions could be used to complete the phrase.

Input. Suggestion interfaces are dependent upon a user's input accuracy to offer predictions: it difficult to generate good predictions for error-laden input, while error-free input eliminates any need for them. The input accuracy of subjects is not of interest here – only their interaction with the suggestion interface is. Therefore, we artificially constrained input to be error-free and only examined word-completion suggestions (not error corrections). This was accomplished by only accepting taps on the correct key or suggestion (ignoring all other taps with no feedback). Combined with the probability scoring (described below), this controlled the occurrence and content of suggestions.

Phrases & Suggestions. Phrases were randomly selected from a set of 500 [20] – modified for consistent spelling in all-lowercase, and without punctuation. Suggestions were generated from the open-source Google Android dictionary.[1]

Probability Scoring. Frequency information from the dictionary determined an estimate of suggestion probability: $(f/255) \cdot (1 - (c/s))$, where f is the frequency of the word (in the range $[0, 255]$), s is its length, and c is the number of letters completed by it.[2] The estimate was normalised by the sum of all candidate estimates. With the given dictionary, this has a theoretical average saving of 46% of the letters in each word (i.e. the correct suggestion appears in the top three after the user enters approximately half of the letters of a word).

Suggestions. After each key tap, the interface scored suggestions matching the entered word prefix. In the *extraverted* condition, or if any of the scores were above a threshold of 0.1 in the *ambiverted* condition, the top three suggestions (sorted by score) were displayed. If all words were below the threshold, any currently displayed suggestions remained. The suggestion bar was cleared between words.

Suggestions were selected by tapping on them, whereupon the characters it completed were inserted into the response.

[1] `https://goo.gl/wN4vHn`

[2] A simplified version of the scoring method used by Google Android: `https://goo.gl/vk4XNh`, lines 1098–1147.

Design & Procedure

A within-subjects design was used for the factor *assertiveness* ∈ *{extraverted, ambiverted, introverted}*, with the order of presentation counterbalanced. Subjects were instructed to enter the text however they felt was fast and comfortable, and should not feel compelled to use the suggestions.

Subjects first completed five practice phrases in an *extraverted* condition. For each condition, they completed 18 randomly selected phrases (without replacement), with the first three discarded – followed by a NASA-TLX questionnaire [9].

RESULTS

There are many metrics that can assess typing performance [19, 21]. We analyse task completion time measured in characters per second (CPS) for each phrase, but are also interested in the number of taps (on keys and suggestions) required per character of input (TPC), and the presentation/usage of the suggestion interface.

Performance

Figure 2a shows the mean CPS, and a one-way ANOVA for *assertiveness* revealed a significant effect ($F_{2,32} = 12.16$, $p < .001$, $\eta^2_G = 0.07$), with *introverted* (mean 3.09 CPS, 95% CI ±0.12) outperforming both *ambiverted* (2.81 ± 0.09) and *extraverted* (2.66 ± 0.09). For the number of taps required per character (TPC; Figure 2b), a Friedman test revealed a significant effect ($\chi^2_2 = 34$, $p < .001$), with *extraverted* requiring significantly fewer (mean .93 TPC, 95% CI [.92, .93]) taps than either *ambiverted* (.96 [.95, .96]) or *introverted* (1 [1, 1]); Bonferroni-corrected Wilcoxon signed-rank tests revealed significant differences ($p \leq .003$) between all pairs.

Suggestions

While suggestions in the *extraverted* condition were updated at every opportunity (after every letter key tap or suggestion selection), they were updated after only 26.72% of such opportunities in the *ambiverted* condition (based on the threshold). Figure 2c shows the proportion of those opportunities where suggestions were shown, those where helpful suggestions were shown,[3] and those where subjects utilised them.

The tap savings that subjects could have obtained if they selected the correct suggestion at the earliest opportunity and the savings actually realised are shown in Figure 2d. Wilcoxon signed-rank tests between the *extraverted* and *ambiverted* conditions revealed significant differences for both the savings offered and realised (both $W = 153$, $p < .001$). In both cases, *extraverted* provided higher potential (17.59% vs. 9.44%) and realised (9.44% vs. 4.66%) tap savings than *ambiverted*.

Subjective Responses

Responses from the NASA-TLX surveys are shown in Table 1. Friedman tests revealed significant differences for *physical demand* ($\chi^2_2 = 14.31$) and *effort* ($\chi^2_2 = 16.04$, both $p \leq .001$). Bonferroni-corrected Wilcoxon signed-rank tests found those ratings for *introverted* to be significantly higher.

DISCUSSION

Our results found that an *ambiverted* interface – guarding suggestions by a threshold – was faster, on average, than an

[3] Less than the theoretical 46% due to the relatively short words in the experimental phrases and consideration of the tap to select a suggestion.

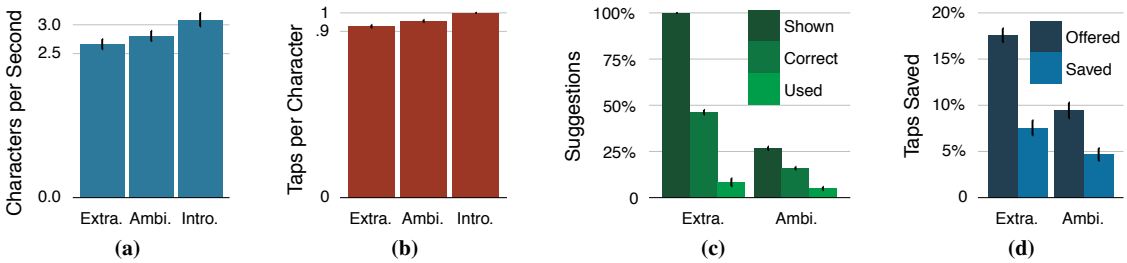


Figure 2. Mean performance metrics (±95% CIs) for each condition: (a–b) entry speed; (c) the proportion of suggestions bars shown, that contained a helpful suggestion, and that were actually used; and (d) the potential and obtained tap savings.

	Mental	**Phys.**	Temp.	Perf.	**Effort**	Frust.
Extra.	2.41 (0.94)	2.24 (0.83)	1.94 (0.56)	2.65 (1.06)	2.24 (0.56)	1.94 (0.75)
Ambi.	2.00 (1.06)	2.18 (0.88)	1.00 (0.79)	3.12 (1.17)	2.53 (1.12)	2.29 (0.99)
Intro.	2.12 (1.27)	3.41 (1.12)	2.47 (1.28)	3.06 (1.03)	3.59 (1.00)	2.59 (1.33)

Table 1. Mean (and std. dev.) NASA-TLX ratings (scale: 1–5).

extraverted interface that always made suggestions visible; and an *introverted* baseline that never made any suggestions was faster still. The performance loss with suggestions was in spite of the reduced average number of taps required to enter text. Nevertheless, suggestions were subjectively considered less physically demanding and effortful.

This result follows that of prior work in a different context (motor-control-impaired users on physical keyboards [14, 15]), which found it difficult for suggestions to gain a time advantage over the perceptual and cognitive costs they introduced. Our results are based on a modern touchscreen keyboard, and the constraint of perfect input from subjects arguably gives a greater time advantage to using suggestions. The maximum theoretical letter savings of 46% is similar to the current limit of production interfaces [7], but it is likely that developments in prediction systems will change this cost-benefit balance.

Although the time cost of attending, deciding on, and selecting suggestions lowered average speed, subjective comments following the experiment (in addition to TLX responses) indicated a strong dislike for the *introverted* condition: that it felt more demanding and time consuming. This suggests that the user experience benefits of suggestions lie beyond average time performance (e.g. spelling assistance or learned automaticity), and may be of greater psychological value than their objective costs imply. Therefore, we cannot conclude that that less assertiveness necessarily results in a better user experience [25].

Our results are also limited to word completion suggestions. N-gram word prediction and correction of input errors may provide greater savings at the same or similar costs of attending, deciding, and selecting them. Such suggestions may take advantage of natural pauses in user input, or the high costs of editing already entered text on mobile interfaces. The experimental task was also a copying task, where subjects could always see the exemplar text to be entered – and were not encumbered by decisions about how to *compose* the text, or uncertainty about spelling or grammar [16]. When a user is unsure about the correct spelling of a word, a correctly spelt suggestion may be substantially faster than trying out spellings and editing the text when it is discovered to be incorrect.

Suggestion interfaces may also have an *overhead cost* in addition to the costs of using them. That is, the anticipation of making a decision, may introduce a persistent cost to the use of the interface. Such overhead costs are likely to be interface and user-dependent, but are critical to understand if systems are to be designed to overcome them [e.g. 26, 27]. Conversely, common suggestions for habitual errors may be learned and become an automatic part of a user's behaviour and removing them may impair performance and overall experience. Teasing apart these factors will require detailed and precise experimentation.

Our results have shown that suggestion interfaces face significant challenges in enabling users to improve their input performance. However, the costs associated with suggestions deserve more attention, as do more principled approaches for managing assertiveness (e.g. with models of economic risk and utility). Researchers and practitioners developing systems with such trade-offs should be cognisant of these costs, but also that benefits may be realised in other types of assistance.

CONCLUSIONS & FUTURE WORK

We have presented an empirical analysis of interaction with text input suggestions, focussing on how eager interfaces should be to show word completions (their *assertiveness*). A moderately assertive (*ambiverted*) interface that guarded completions with a confidence threshold afforded faster performance than a more assertive (*extraverted*) interface that always presented the best three. While both types of interface were, on average, slower than never presenting completions at all (*introverted*), the suggestions were utilised to save taps and were subjectively preferred – indicating benefits not measured by time. Future work will examine broader text entry tasks (e.g. composition and error correction), and model this trade-off by examining the *utility* of text entry suggestions.

REFERENCES

1. Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2010. Predicting the cost of error correction in character-based text entry technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, 5–14. DOI: http://dx.doi.org/10.1145/1753326.1753329
2. Shiri Azenkot and Shumin Zhai. 2012. Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services (MobileHCI '12)*. ACM, New York, NY, 251–260. DOI:http://dx.doi.org/10.1145/2371574.2371612
3. Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, London.
4. Mark Dunlop and John Levine. 2012. Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, 2669–2678. DOI:http://dx.doi.org/10.1145/2207676.2208659
5. Leah Findlater and Jacob Wobbrock. 2012. Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, 815–824. DOI: http://dx.doi.org/10.1145/2207676.2208520
6. Frank O. Flemisch, Catherine A. Adams, Sheila R. Conway, Ken H. Goodrich, Michael T. Palmer, and Paul C. Schutte. 2003. *The H-Metaphor as a Guideline for Vehicle Automation and Interaction*. Technical Report NASA/TM-2003-212672. NASA Langley Research Center, Hampton, VA. http://ntrs.nasa.gov/search.jsp?R=20040031835
7. Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2015. Effects of Language Modeling and Its Personalization on Touchscreen Typing Performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, 649–658. DOI:http://dx.doi.org/10.1145/2702123.2702503
8. Nestor Garay-Vitoria and Julio Abascal. 2006. Text prediction systems: a survey. *Universal Access in the Information Society* 4 (2006), 188–203. DOI:http://dx.doi.org/10.1007/s10209-005-0005-9
9. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. Elsevier, North Holland, 139–183. DOI:http://dx.doi.org/10.1016/S0166-4115(08)62386-9
10. Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, 2659–2668. DOI:http://dx.doi.org/10.1145/2207676.2208658
11. Heidi M. Horstmann and Simon P. Levine. 1990. Modeling of user performance with computer access and augmentative communication systems for handicapped people. *Augmentative and Alternative Communication* 6, 4 (1990), 231–241. DOI:http://dx.doi.org/10.1080/07434619012331275514
12. Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of attention in computing and communication: from principles to applications. *Commun. ACM* 46, 3 (March 2003), 52–59. DOI: http://dx.doi.org/10.1145/636772.636798
13. Heidi Horstmann Koester and Simon P. Levine. 1993. A model of performance cost versus benefit for augmentative communication systems. In *Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 1303–1304. DOI:http://dx.doi.org/10.1109/IEMBS.1993.979147
14. Heidi Horstmann Koester and Simon P. Levine. 1994. Modeling the Speed of Text Entry with a Word Prediction Interface. *IEEE Transactions on Rehabilitation Engineering* 2, 3 (1994), 177–187. DOI: http://dx.doi.org/10.1109/86.331567
15. Heidi Horstmann Koester and Simon P. Levine. 1996. Effect of a word prediction feature on user performance. *Augmentative and Alternative Communication* 12, 3 (1996), 155–168. DOI:http://dx.doi.org/10.1080/07434619612331277608
16. Per Ola Kristensson and Keith Vertanen. 2012. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, 29–32. DOI: http://dx.doi.org/10.1145/2166966.2166972
17. Ying Liu and Kari Jouko Räihä. 2010. Predicting Chinese text entry speeds on mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, 2183–2192. DOI:http://dx.doi.org/10.1145/1753326.1753657
18. I. Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. 2001. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '01)*. ACM, New York, NY, 9–16. DOI: http://dx.doi.org/10.1145/365024.365028
19. I. Scott MacKenzie and R. William Soukoreff. 2002. Text Entry for Mobile Computing: Models and Methods, Theory and Practice. *Human-Computer Interaction* 17, 2–3 (2002), 147–198. DOI:http://dx.doi.org/10.1080/07370024.2002.9667313
20. I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in*

Computing Systems (CHI EA '03). ACM, New York, NY, 754–755. DOI:http://dx.doi.org/10.1145/765891.765971

21. I. Scott MacKenzie and Kumiko Tanaka-Ishii (Eds.). 2007. *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann, San Francisco, CA.
22. Toshiyuki Masui. 1998. An efficient text input method for pen-based computers. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '98)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, 328–335. DOI:http://dx.doi.org/10.1145/274644.274690
23. Hugo Nicolau and Joaquim Jorge. 2012. Touch typing using thumbs: understanding the effect of mobility and hand posture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, 2683–2686. DOI:http://dx.doi.org/10.1145/2207676.2208661
24. Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. Improving two-thumb text entry on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, 2765–2774. DOI:http://dx.doi.org/10.1145/2470654.2481383
25. Philip Quinn and Andy Cockburn. 2016. When Bad Feels Good: Assistance Failures and Interface Preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY. DOI:http://dx.doi.org/10.1145/2858036.2858074
26. Philip Quinn, Andy Cockburn, and Jérôme Delamarche. 2013. Examining the costs of multiple trajectory pointing techniques. *International Journal of Human-Computer Studies* 71, 4 (2013), 492–509. DOI:http://dx.doi.org/10.1016/j.ijhcs.2012.12.001
27. Philip Quinn, Andy Cockburn, Kari-Jouko Räihä, and Jérôme Delamarche. 2011. On the costs of multiple trajectory pointing methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, 859–862. DOI:http://dx.doi.org/10.1145/1978942.1979067
28. Andrew Sears. 1991. Improving Touchscreen Keyboards: Design issues and a comparison with other devices. *Interacting with Computers* 3, 3 (1991), 253–269. DOI:http://dx.doi.org/10.1016/0953-5438(91)90016-U
29. Andrew Sears, Julie A. Jacko, Josey Chu, and Francisco Moro. 2001. The role of visual search in the design of effective soft keyboards. *Behaviour & Information Technology* 20, 3 (2001), 159–166. DOI:http://dx.doi.org/10.1080/01449290110049790
30. R. William Soukoreff and I. Scott MacKenzie. 1995. Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard. *Behaviour & Information Technology* 14, 6 (1995), 370–379. DOI:http://dx.doi.org/10.1080/01449299508914656
31. Keith Trnka, John McCaw, Debra Yarrington, Kathleen F. McCoy, and Christopher Pennington. 2009. User Interaction with Word Prediction: The Effects of Prediction Quality. *ACM Transactions on Accessible Computing* 1, 3, Article 17 (Feb. 2009), 34 pages. DOI:http://dx.doi.org/10.1145/1497302.1497307
32. Daryl Weir, Henning Pohl, Simon Rogers, Keith Vertanen, and Per Ola Kristensson. 2014. Uncertain Text Entry on Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, 2307–2316. DOI:http://dx.doi.org/10.1145/2556288.2557412
33. Shumin Zhai, Michael Hunter, and Barton A. Smith. 2002. Performance Optimization of Virtual Keyboards. *Human-Computer Interaction* 17, 2–3 (2002), 229–269. DOI:http://dx.doi.org/10.1080/07370024.2002.9667315