The Temporal Dynamics of Speeded Decision Making

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 13 maart 2012, te 12:00 uur

door
Gilles Dutilh

geboren te Warnsveld
Promotiecommissie

Promotor: Prof. Dr. H. L. J. van der Maas
Copromotor: Dr. E.–J. Wagenmakers

Overige leden: Prof. Dr. G. D. Logan
               Prof. Dr. J. M. J. Murre
               Prof. Dr. K. R. Ridderinkhof
               Prof. Dr. F. Tuerlinckx
               Dr. R. P. P. P. Grasman
               Dr. H. D. van Rijn

Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam

Cover design by Janwillem van der Veen: An impression of the decomposition that helps us to understand the dynamics of correct and error response times.

This dissertation is typeset in \LaTeX{} using the Memoir class.
# Contents

## Contents

1 Introduction
   1.1 Response Times and Their Role in Inferring Mental Organization
   1.2 Why a Model?
   1.3 The Diffusion Model
   1.4 Exploring Virgin Territory and Beyond
   1.5 Practice Effects: A Unitary Phenomenon?
   1.6 Sequential Effects
   1.7 Accurate Responding and Guessing: Two Extremes of the Same Process?
   1.8 Temporal Dynamics of Speeded Decision Making

2 A Diffusion Model Decomposition of the Practice Effect
   2.1 Introduction
   2.2 Power Law Theories of Automatization
   2.3 The Power Law Repealed?
   2.4 Interim Conclusion: State of the Field
   2.5 The Ratcliff Diffusion Model
   2.6 Experiment
   2.7 Discussion
   2.A The Bayesian Ratcliff Diffusion Model
   2.B Results for the Variability Parameters

3 Task–Related vs. Stimulus–Specific Practice: A Diffusion Model Account
   3.1 The Diffusion Model
   3.2 Method
   3.3 Results

4 Testing Theories of Post–Error Slowing
   4.1 Explanations for Post–Error Slowing
   4.2 A Drift Diffusion Model Decomposition of Response Times
   4.3 From Process to Parameter: A Drift Diffusion Model Perspective on Post–Error Slowing
   4.4 Method
   4.5 Results and Discussion
# Contents

4.6 Concluding Comments ........................................... 59

5 A Diffusion Model Account of Age Differences in Post–Error Slowing
  5.1 Introduction ......................................................... 63
  5.2 Response Speed and Age .......................................... 64
  5.3 Accounting for Speed and Accuracy Simultaneously .......... 65
  5.4 Quantification of Post–Error Slowing ......................... 67
  5.5 Method ............................................................. 68
  5.6 Results ............................................................. 69
  5.7 Discussion ......................................................... 75

6 How to Measure Post–Error Slowing: A Confound and a Simple Solution
  6.1 Introduction ......................................................... 79
  6.2 The Measurement of Post–Error Slowing .................... 81
  6.3 A Confound ........................................................ 81
  6.4 Simulation Studies ............................................... 83
  6.5 A Simple Solution ............................................... 87
  6.6 Empirical Illustration: The Confound is Real ............... 88
  6.7 Concluding Comments .......................................... 89

7 A Phase Transition Model for the Speed–Accuracy Trade–Off in Response Time Experiments
  7.1 Introduction ......................................................... 92
  7.2 The Speed–Accuracy Trade–off ................................ 93
  7.3 Current Models of the Speed–Accuracy Trade–Off .......... 96
  7.4 The Phase Transition Model .................................... 101
  7.5 Experiment 1: Hysteresis ....................................... 102
  7.6 Interim Conclusion ............................................... 105
  7.7 Experiment 1c: Modified Method of Limits ................. 107
  7.8 Experiment 2: Bimodality ....................................... 109
  7.9 Concluding Remarks ............................................ 124

8 Summary and Discussion ........................................... 129
  8.1 Summary of Results ............................................. 129
  8.2 Discussion ......................................................... 132
  8.3 Concluding Remarks ............................................ 136

References .............................................................. 137

9 Nederlandstalige Samenvatting ................................ 149
  9.1 Reactietijden en het Belang van een Formeel Model .... 149
  9.2 Oefening Baart (niet alleen) Kunst .......................... 151
  9.3 Leren van Fouten? .............................................. 151
  9.4 Van Accuraat Reageren tot Gokken ......................... 152
  9.5 Conclusie .......................................................... 153

Dankwoord ............................................................. 155
Chapter 1

Introduction

1.1 Response Times and Their Role in Inferring Mental Organization

Brains think. The more brains think, the more time they need to do so. This simple fact is the basis of the wide-spread use of response times in psychology. Many years ago, psychologists already realized that response times are informative about the mental processes instantiated in the brain. Or, to put it negatively, “we surely do not understand a choice process very thoroughly until we can account for the time required for it to be carried out” (Luce, 1986, page vii). F. C. Donders was one of the first who realized that with good experimental design, response times can be used to infer the duration of different components of mental processing (Donders, 1869). Since then, experimental psychologists have developed many experimental tasks that have response time as their main observed variable, such as the Stroop task (Stroop, 1935), the lexical decision task (e.g., Wagenmakers, Ratchi, Gomez, & McKoon, 2008), the stop signal task (Logan, Cowan, & Davis, 1984), the flanker task (Eriksen & Eriksen, 1974), the Simon task (Craft & Simon, 1970), the random dot motion task (Britten, Shadlen, Newsome, & Movshon, 1992), the implicit association test (Greenwald, McGhee, & Schwartz, 1998), and many others. In these tasks, response time (from now on: RT) are of interest not only because they allow us to learn about how people process information, but also because they allow us to quantify individual differences in performance.

A large class of RT tasks are speeded two choice decision tasks. In such tasks, a participant has to choose between two response alternatives on the basis of a stimulus. A popular task, for example, is the “lexical decision task”, in which participants have to quickly decide whether a presented letter string is an existing word (such as “house”) or a nonword (or pseudoword, such as “drapa”). Such two-choice tasks come with a second variable: the accuracy of the response. Often, accuracy has a strong relationship to the RT in the task, a phenomenon known as the speed–accuracy trade–off (Schouten & Bekker, 1967; R. G. Pachella, 1973; Wickelgren, 1977). This speed–accuracy trade–off is largely under the control of the participant, who is most often simply asked to respond “as fast and accurately as possible”.

1
1. Introduction

1.2 Why a Model?

Since the speed–accuracy trade–off lies largely in hands of the participant, the use of either mean RT or accuracy as dependent measure is problematic. To understand the problem, consider two participants who perform the same RT task. One of them turns out to be much quicker, but also much less accurate than the other. In this situation, it is difficult to tell which of the two participants is better. In many studies, this problem is ignored and either RT or accuracy are taken as the dependent measure of interest. Although this choice is in some cases defendable, in most cases, ignoring either accuracy or RT means to ignore an important source of information in the data.

A second source of information that is ignored by the standard analysis of accuracy and mean RT is the shape of the RT distribution. Two participants can have the same mean RT but vary enormously in the spread of their RTs. Moreover, RT distributions have a pronounced right skew; this skew increases with task difficulty such that the slow RTs may be most informative about the efficiency of processing.

In order to account for accuracy, mean RT, and the distributional form of RT, several models of RT have been proposed over the years. The most popular class of RT models consists of the sequential sampling models. All sequential sampling models share the same basic assumption: When a stimulus is presented, a participant starts to sequentially sample units of information from that stimulus. These units of information can be either evidence in favor of say, response A or in favor of response B. At every step in this sampling process, the sampled evidence is integrated with the evidence already collected and the evidence in favor of response A and B is evaluated. Whenever the evidence for one response over the other reaches a pre–set criterion, the corresponding response is initiated. The different sequential sampling models differ amongst each other in 1) the way information is sampled (discretely or continuously), 2) the way the sampled information is integrated, and 3) the way the decision criterion is implemented.

1.3 The Diffusion Model

In 1978, Ratcliff proposed the diffusion model for speeded two choice decision making (Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002). The diffusion model is a typical sequential sampling model. The model assumes that noisy information is continuously sampled from the stimulus. In the model, the difference in evidence accumulated in favor of response A versus B is continuously compared to two boundaries (Figure 1.1). When the upper boundary is reached, enough information has accumulated to respond “A”. When the lower boundary is reached, enough information has accumulated to respond “B”. When an A stimulus is presented, most of the sampled information will be in favor of response “A”. However, due to the noisy nature of the accumulation process, every now and then evidence in favor of response “B” is added to the balance. The result is that most evidence sample paths will reach the correct A boundary (grey sample path in Figure 1.1), but sometimes the B boundary will be hit, resulting in an error response (black sample path). The noise in the information accumulation does not only result in occasional errors, but also produces the variability in RT (see histograms in Figure 1.1).

From these basic assumptions, the diffusion model provides a detailed and comprehensive account of performance in speeded two–choice tasks (Ratcliff & McKoon, 2008; Wagenmakers, 2009). The model describes how accuracy and the entire distribution of RT for both correct and error responses result from unobserved psychological processes that are represented by the model’s parameters. In addition, the model provides an ac-
1.3. The Diffusion Model

Figure 1.1: Graphical illustration of the diffusion model. The two example sample paths represent the accumulation of evidence from an A stimulus, resulting in one correct response (light line) and one error response (dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper A boundary is reached more often than the incorrect, lower B boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a non–decision component that represents the time needed for processes such as stimulus encoding and response execution.

In order to apply the diffusion model for the analysis of empirical data, several of its components can vary from one experimental situation to the next. The precise contribution of each of those components is defined by the model’s parameters. Fitting the diffusion model to data means to optimize those parameters to describe the data of an experimental situation. In the version of the model that is used throughout most of this dissertation, the diffusion model has 7 parameters.

1. Mean drift rate ($v$). Drift rate quantifies the deterministic component in the information–accumulation process. This means that when the absolute value of drift rate is high, decisions are fast and accurate; Thus, $v$ indexes task difficulty or subject ability.

2. Across–trial variability in drift rate ($\eta$). This parameter reflects the fact that drift rate may fluctuate from one trial to the next, according to a normal distribution with mean $v$ and standard deviation $\eta$. The parameter $\eta$ allows the diffusion model to account for data in which error responses are systematically slower than correct responses (Ratcliff, 1978).

3. Boundary separation ($a$). Boundary separation quantifies response caution and modulates the speed–accuracy tradeoff: At the price of an increase in RT, participants can decrease their error rate by widening the boundary separation. Thus, boundary separation $a$ indexes a participant’s response caution.
4. Mean starting point (z). Starting point reflects the a priori bias of a participant for one or the other response. This parameter is usually manipulated via payoff or proportion manipulations (Edwards, 1965; Wagenmakers et al., 2008; but see Diederich & Busemeyer, 2006).

5. Across–trial variability in starting point (sz). This parameter reflects the fact that starting point may fluctuate from one trial to the next, according to a uniform distribution with mean z and range sz. The parameter sz also allows the diffusion model to account for data in which error responses are systematically faster than correct responses.

6. Mean of the nondecision component of processing (Ter). This parameter encompasses the time spent on common processes, i.e., processes executed irrespective of the decision process. The diffusion model assumes that the observed RT is the sum of the nondecision component and the decision component (Luce, 1986):

\[ RT = DT + T_{er}, \quad (1.1) \]

where DT denotes decision time. Therefore, nondecision time Ter does not affect response choice and acts solely to shift the entire RT distribution.

7. Across–trial variability in the nondecision component of processing (st). This parameter reflects the fact that nondecision time may fluctuate from one trial to the next, according to a uniform distribution with mean Ter and range st. The parameter st also allows the model to capture RT distributions that show a relatively shallow rise in the leading edge.

Validity of the Interpretation of the Diffusion Model’s Parameters

One main advantage of the diffusion model is that it allows researchers to draw conclusions based on the psychological components of processing that are postulated by the model, rather than on accuracy and RT alone. Such conclusions can only be drawn when the mapping of the hypothesized psychological components on the model’s parameters is valid.

Fortunately, several studies attest to the validity of the diffusion model parameters. One such validation study was performed by Voss, Rothermund, and Voss (2004) who experimentally manipulated the underlying psychological components that are associated with the main four parameters of the diffusion model. Subsequently, Voss et al. (2004) studied whether these manipulations were reflected in the model’s parameters when fit to the data. For instance, Voss et al. (2004) found that (1) increasing participants’ motivation to respond accurately resulted in higher boundary separation estimates, (2) difficult stimuli resulted in lower drift rates than easy stimuli, (3) a less convenient way of pressing the response button resulted in higher non–decision times and (4) unequal rewards resulted in a shift in the starting point toward the rewarded response. These results strongly support the link between the parameters of the diffusion model and the psychological processes that they are thought to represent. Converging evidence comes not just from validation studies, but also from empirical studies that repeatedly find the same mapping (e.g., Ratcliff & Rouder, 1998; Wagenmakers et al., 2008).
1.4 Exploring Virgin Territory and Beyond

Although the diffusion model has become increasingly popular over the last 15 years, many phenomena from fields of research that use RT data are still virgin territory as formal modeling is concerned. The application of the diffusion model to explore such virgin phenomena can turn out enlightening. The first four chapters of this thesis are reports of two such explorations (i.e., practice effects and post–error slowing). Chapter 6 describes a methodological obstacle that we encountered when studying post–error slowing, and offers a simple solution to avoid it. In the chapter 7, we take some distance and question the comprehensiveness of the diffusion model. The next three paragraphs give a short introduction to each of the chapters in this thesis.

1.5 Practice Effects: A Unitary Phenomenon?

When participants repeatedly perform the same task, their RTs invariably decrease. This benefit of practice is generally strongest at the beginning of practice and diminishes over time. In the literature on the practice effect, the focus of debate has mostly been on what mathematical function best describes the speed–up of mean RT with practice. This focus on mean RT has largely ignored effects on accuracy and the shape of RT distributions. In the study described in the chapter 2, we applied the diffusion model to take into account all these sources of information and decompose the effect of practice into underlying psychological constructs. To this end, a 10,000 trial lexical decision task was administered and analyzed with the diffusion model. The results of this study showed that the effect of practice could not be attributed to a single psychological process, but instead reflects changes in at least three qualitatively different processes. However, after this study, we were left with uncertainty about the origin of the effects. In particular, we wondered whether the effects were due to increased familiarity with the word and nonword stimuli or to increased familiarity with the task in general. Chapter 3 of this dissertation addresses this question.

1.6 Sequential Effects

In RT research, consecutive trials are generally assumed to be independent. However, performance on the present trial often depends on performance on the previous trials. One sequential effect that has received considerable attention, in particular in the last decade, is post–error slowing. Post–error slowing refers to the phenomenon that after committing an error, participants tend to slow down on subsequent trials. Several explanations have been proposed for this phenomenon. The most popular explanation is that participants monitor their performance and interpret an error as a sign that they should respond more carefully on the next trial. Although this explanation in terms of strategic adjustments is very popular, several other plausible hypotheses exist. In chapter 4, we show that the diffusion model is an ideal tool to tell apart the different explanations that have been proposed. The analysis of a 1,094,886 trial lexical decision data set allowed us to conclude that, for this data set, strategic adjustments of response caution indeed underlie the post–error slowing effect.

One of the fields in which post–error slowing plays a role is the psychology of aging. Elderly participants, who are generally found to respond more cautiously than young participants, have also been found to slow down more strongly after committing an error. To study what processes underly this increased post–error slowing with age, we applied
1. Introduction

the diffusion model to a perceptual discrimination data set administered to young and elderly participants in chapter 5. The results show that both young and elderly participants wasted time on irrelevant processes after committing an error. Furthermore, the occurrence of an error caused elderly participants to process information more slowly and to become more cautious.

In the chapter 6, we describe a problem we encountered when analyzing post–error slowing effects. It turned out that the standard and wide–spread method to analyze post–error slowing is vulnerable to a confound of global changes in behavior over the course of a task. Fortunately, the solution we offer is simple and effective.

1.7 Accurate Responding and Guessing: Two Extremes of the Same Process?

The diffusion model describes how the speed–accuracy trade–off is governed by the separation of response boundaries. Wide boundaries result in decisions that are slow but accurate; narrow boundaries lead to decisions that are fast but error–prone. This interpretation of the model suggests that, with increasing emphasis to respond quickly, participants can bring response boundaries together closer and closer, until—with a boundary separation of zero—their performance is very fast and at chance accuracy. In chapter 7 we study this limiting case and focus on what happens when participants are forced to get faster and faster, until they are guessing. We show that two different regimes appear to govern behavior: one fast–guessing regime and an accurate, stimulus controlled regime. We show that both regimes are stable and that the transitions between these regimes are abrupt rather than continuous. From catastrophe theory, we derive testable predictions that are indicators of such discrete phase transitions. Three experimental studies test these predictions.

1.8 Temporal Dynamics of Speeded Decision Making

Although the chapters of this thesis cover a wide range of topics, they address a common question: how do people adjust the way they perform a task, given changes in internal and external constraints? Throughout this dissertation, you will find that performance of participants is far from static. Past experience with stimulus material, their own performance, and task demands all allow participants to strategically adjust the way they process information and make decisions.
Chapter 2

A Diffusion Model Decomposition of the Practice Effect

An excerpt of this chapter has been published as:
A Diffusion Model Decomposition of the Practice Effect

Abstract

When people repeatedly perform the same cognitive task, their response times (RTs) invariably decrease. The mathematical function that best describes this decrease has been the subject of intense theoretical debate. Most current theories start from the assumption that the practice effect is driven by changes in a single psychological process. To test this assumption, we set out to decompose the effect of practice with the Ratcliff diffusion model. The diffusion model accounts for proportion correct as well as the shape of RT distributions for both correct and error responses. Moreover, the estimation of the model parameters allows a decomposition of the practice effect in terms of its constituent psychological processes. Diffusion model analysis of data from a 10,000–trial lexical decision task demonstrate that practice not only affects speed of information processing, but also affects response caution, response bias, and peripheral processing time. We conclude that the practice effect consists of multiple subcomponents, and that it may be hazardous to abstract the interactive combination of these subcomponents in terms of a single output measure such as mean RT for correct responses. Supplementary materials may be downloaded from www.psychonomic.org/archive.

2.1 Introduction

When people repeatedly perform the same task, their performance becomes fast, accurate, and relatively effortless. For example, you are able read this text quickly, virtually without errors, and, hopefully, without investing too much effort. The difference between your
2. Decomposition of Practice

performance now and when you first learned to read is staggering; from a slow, error
prone, and effortful endeavor, your reading has matured into automatized skill.

Traditionally, researchers in the field of skill acquisition have quantified the effect of
practice primarily in terms of reduction in the time to execute a given task (i.e., response
time or RT; Logan, 1988, 1992; A. Newell & Rosenbloom, 1981; Woodworth & Schlosberg,
1954). Almost every study has shown that the RT benefits due to practice are largest at
the start of training, and then slowly diminish over time.

This ubiquitous result, many researchers argued, is best captured by a power function
that relates mean RT to practice via the equation

\[ MRT = a + bN^{-c}, \]  

(2.1)

where \( MRT \) is the mean RT for correct responses, \( a \) quantifies asymptotic performance,
\( b \) quantifies the difference between initial and asymptotic performance, \( N \) represents
the amount of practice, and \( c \) is the rate parameter that determines the shape of the power
law. Empirical support for the power function relation between RT and practice has been
reported across a range of tasks, for instance in cigar rolling, maze solving (Crossman,
1959), fact retrieval (Pirolli & Anderson, 1985), and a variety of standard psychological
tasks (Logan, 1992). Support for the power function relation appeared so strong that
the relation has often been referred to as a law (e.g., “the ubiquitous law of practice”,

Nevertheless, some researchers have questioned whether the speed up with practice
is really governed by a power function. In particular, Heathcote, Brown, and Mewhort
(2000) have argued that the power law is an artifact of averaging practice functions over
participants (see also R. B. Anderson & Tweney, 1997; Myung, Kim, & Pitt, 2000). Heathcote
et al. (2000) showed that for data of many experiments on skill acquisition,
individual learning curves were better described by an exponential function that relates
mean RT to practice via the equation

\[ MRT = a + b \exp (-cN), \]  

(2.2)

where the interpretation of the parameters is the same as in Equation (2.1).

Regardless of the specific shape of the function that relates the amount of practice to
mean RT, the previous discussion illustrates how most empirical studies on the practice
effect have focused on the decrease in mean RT for correct responses. By doing so, the field
has largely neglected two other important sources of information, namely accuracy (i.e.,
proportion correct responses) and the distribution of RTs (e.g., spread and skewness).
Those researchers that do take response accuracy into account tend to ignore RT (e.g.,
B. A. Dosher & Lu, 2007, but see Nosofsky & Alfonso-Reese, 1999), or present both RT
and accuracy as separate output variables—even though it is well known that RT and
accuracy are intimately related (e.g., R. G. Pachella, 1974; Schouten & Bekker, 1967;

In this article we seek a detailed understanding of the effect of practice by taking into
account simultaneously the changes in accuracy and the changes in RT distributions, both
for correct responses and for error responses. In order to do so, we follow Ratcliff, Thapar,
and McKoon (2006b) in applying the Ratcliff diffusion model (Ratcliff, 1978; Ratcliff &
McKoon, 2008; Wagenmakers, 2009) to the field of automatization in cognitive tasks.
The diffusion model uses all of the information in the data, and is able to decompose the
practice effect into its constituent psychological processes.

To foreshadow our main theoretical result, the diffusion model analyses demonstrate
that the effect of practice is driven by changes in several processes: In our study, practice
effects could be decomposed into increases in speed of information processing, changes in response caution and a priori bias, and large changes in peripheral processing time. Thus, different processes appear to be involved in the improvement due to practice, and these processes furthermore interact in non–trivial ways to yield the observed data. This suggests that it may be hazardous to draw strong conclusions from practice–induced changes in mean or spread of RT alone. Theories of practice should address changes in these components.

The outline of this paper is as follows. The first section provides a selective overview of quantitative models that account for the power law of practice, and the second section briefly states the argument for an exponential law of practice. These sections provide a summary of the theoretical framework against which our approach may be contrasted. The third section introduces the diffusion model, the fourth section describes our 10,000–trial lexical decision experiment, and the fifth section discusses the results, both in terms of the observed data and in terms of the model parameters. The sixth section concludes.

2.2 Power Law Theories of Automatization

Research on practice has a long history of large experiments with a variety of tasks. The first comments on the power function as a general law of practice were already made in 1926 by Snoddy (as cited by A. Newell & Rosenbloom, 1981) and from the 1950’s onward, belief in the ubiquity of the power function has steadily grown (see for example Fitts & Posner, 1967). A. Newell and Rosenbloom (1981, p. 3) even state that the power law “holds for learning of all kinds”, and exemplify this by learning of motor skills, perceptual skills, complex routines, and all kinds of problem solving.

Several theories have tried to explain why the speed–up with practice follows a power function. A. Newell and Rosenbloom (1981), for example, conceptualize a task as being composed of different subtasks, each with a different learning rate. The subtasks that learn at the highest rate initially dominate the practice effect. When the ability for these subtasks to learn has gone down sufficiently, subtasks with a slower learning rate will begin to dominate the speed–up, as their relative contribution to the total processing time has increased. Thus, with practice, the number of components that contribute significantly to the practice effect decreases and the mean contribution of each component diminishes, causing a power function shaped speed–up (for a similar account see MacKay, 1982, p. 493). Another theory was proposed by Crossman (1959). According to this theory, a participant chooses among various methods when performing a task. With practice, faster methods are more and more likely to be selected. These two assumptions can account for a power shaped practice effect (A. Newell & Rosenbloom, 1981).

None of the above mentioned theories are very concrete about the cognitive processes involved. In the following section we provide a selective overview of theories and models that are more specific with respect to the underlying processes that drive the practice effect.

Logan’s Instance Theory

Logan’s Instance Theory of Automatization (ITA; Logan, 1988) outlines a specific mechanism by which practice affects cognitive performance. Instance theory assumes that the first time someone engages in a cognitive task, he or she uses a general algorithm to select the correct response. In addition, the first encounter with a particular stimulus leads to the formation of a new episodic memory trace (i.e., an instance) that represents
2. Decomposition of Practice

A direct link between the stimulus and the response. Upon a future encounter with the same stimulus, this memory trace enters a race with the general algorithm. Performance is determined by the process that finishes first. Every encounter with a stimulus leads to the formation of a new memory trace, and hence the number of traces that compete in the race grows with practice. After several presentations of the same stimulus, the probability that the algorithm finishes first has decreased considerably, and after many presentations virtually all responses originate from memory retrieval—this is when performance is said to be automatized \cite{Logan1988, Logan1990}.

The finishing time of each trace is assumed to be sampled from a single distribution of finishing times. Only the minimum sample of this distribution (i.e., the trace that wins the race) is empirically observed. \cite{Logan1988} shows that, for a broad range of reasonable distributions of finishing times, this minimum decreases as a power function of the number of traces in the race (for a discussion see \cite{Colonius1995, Cousineau2002, Goodman2002, Logan1995}). This means that the ITA is able to predict the power law of practice from a simple set of assumptions.

The ITA not only predicts a power law decrease of mean RT, but it also predicts a power law decrease of the entire RT distribution. Specifically, \cite{Logan1992} showed that adding memory traces in the ITA leads to a power law decrease of all percentiles of an RT distribution. This power law decrease is characterized by the same exponent (i.e., learning rate) for the RT mean, for the RT standard deviation, and for all RT quantiles. This detailed and non–trivial prediction was confirmed by \cite{Logan1992}.

Logan’s ITA was inspirational to the development of Nosofsky and Palmeri’s exemplar based random walk model (EBRW; \cite{Nosofsky1991, Palmeri1991}). In the EBRW, assumptions of instance theory are integrated in Nosofsky’s generalized context model (GCM, \cite{Nosofsky1991}). GCM assumes that categories are represented by exemplars stored in memory. The probability of classification of an object into a certain category depends on how similar the object is to exemplars from that category, relative to exemplars from other categories. In order to not only account for classification proportions, but also RT, in EBRW, Nosofsky and Palmeri added the race principle of ITA to GCM. In contrast to ITA, where only instances identical to the current stimulus compete in the race, in EBRW, all exemplars compete in the race, with rates proportional to their similarity to the presented object. Every race is won by one exemplar, which then drives a random walk process. The duration of a step in the random walk is thus dependent on the exemplar retrieval time, and the direction is defined by which category the retrieved exemplar belongs to. The fact that all exemplars compete in the race, allows EBRW to account for interactive effects of between and within category similarity and practice on response times. In simulations, \cite{Palmeri1997, p. 28–31} shows that the EBRW model is able to predict a power shaped speed up with practice.

In 2002, Logan proposed the instance theory of attention and memory (ITAM). This theory is an elaboration of ITA that incorporates ideas from EBRW such as the random walk decision process and the sensitivity to similarity. Therefore, ITAM is one of the most general theories that predict categorization, memory retrieval, and automatization from a small set of key assumptions, the most prominent one being the episodic, instance–based representation of items in memory.

According to Logan, ITAM “places less emphasis on the power law” \cite[page 391]{Logan2002} than did ITA. In ITAM, the races between the memory traces drive a random walk response selection process. This means that RT in ITAM is the result of the duration for all races plus some additional “bookkeeping time” \cite[p. 391]{Logan2002}. Logan argues that this “combination of effects” \cite[p. 391]{Logan2002} makes it difficult to derive from ITAM mathematical predictions about the effect of practice on RT.
Rickard’s Component Power Laws Model

Just as ITA, Rickard’s component power laws theory (CMPL; [Rickard, 1997, 2004]) describes automatization as a practice-induced shift from algorithm retrieval to memory retrieval. Nevertheless, the CMPL and the ITA make fundamentally different assumptions concerning the nature of information processing and the effect of practice.

In CMPL, algorithm retrieval is defined as a sequence of steps in each of which information is retrieved from memory. On trial outset, the algorithm’s first step competes with memory retrieval. The process that wins this competition is selected, and the other process is suppressed via inhibitory connections (e.g., [Rickard, 1997, p. 292]). However, in contrast to the ITA, the CMPL does not assume that every encounter with a stimulus leads to the formation of a new memory trace. Instead, the CMPL assumes a network of nodes in which practice strengthens existing connections between the problem node, memory retrieval nodes, the algorithm node, and response nodes.

The nodes that correspond to memory retrieval are strengthened every time a stimulus is associated with a particular response, and regardless of whether the response was produced by means of the algorithm strategy or by means of the memory retrieval strategy. The nodes that correspond to the algorithm strategy are strengthened only when the algorithm is executed.

From these assumptions follow several key predictions. First, the model predicts that early in practice, the algorithm strategy will dominate. Because the algorithm is strengthened when it is used, RT will decrease during this phase, even in the absence of any decisions based on memory retrieval. As practice progresses, the memory retrieval nodes are strengthened, and at some point the memory retrieval nodes will start to determine performance. This change from algorithm-based performance to memory-retrieval based performance is typified by an increase in RT variance and a decrease in RT mean (e.g., [Wagenmakers & Brown, 2007, p. 837]). After sufficient practice, memory retrieval will completely dominate the execution of the algorithm.

According to CMPL, the power law does not hold over the entire range of practice, but [Rickard, 1997] shows both formally and in simulations that the power law does hold separately within the two processing regimes, one characterized by a dominant algorithm strategy and one characterized by a dominant memory retrieval strategy (hence, component power laws theory; [Rickard, 1997, pp. 293–296]).

Models that are similar to the CMPL—but that will not be discussed here—include Siegler’s distribution of associations model ([Siegler, 1988]), in which memory retrieval is always attempted first, whereafter the algorithm is applied only when needed, and the Schunn et al. source activation confusion model, in which strategy choice is governed by a higher level familiarity-based feeling-of-knowing process ([Schunn, Reder, & Nhouyvanisvong, 1997]).

Anderson’s ACT–R Model

The ACT–R (Adaptive Control of Thought—Rational; e.g., [J. R. Anderson et al, 2004]) theory attempts to account for the complexities of human cognition from within a single, integrated architecture. The architecture contains, among others, goals, declarative memory “chunks” (i.e., stored units of information), and production rules that compete for execution. Over the course of several decades, the ACT theory has undergone several changes and refinements; nevertheless, the account of practice has remained the same (e.g., [J. R. Anderson, 1982, 1992, J. R. Anderson et al, 2004; Pirolli & Anderson, 1985]).
2. Decomposition of Practice

In ACT–R, practice effects originate from changes in the declarative memory module, which is responsible for storing and reproducing factual information. Importantly, the speed with which information can be recalled from declarative memory is determined by the activation level of the relevant chunk. The activation level $A_i$ of a chunk $i$ (e.g., four plus eight equals twelve, or a robin is a bird) can be decomposed into the additive combination of a baseline level of activation $B_i$ and an associative activation, that is, $A_i = C + B_i$. The associative activation $C$ is context–specific, whereas the baseline activation $B_i$ reflects the “general usefulness” of the chunk in the past.

ACT–R produces practice effects because repeated use of a particular chunk $i$ increases its baseline level of activation $B_i$. As described in J. R. Anderson et al. (2004, p. 1042), $B_i$ is an inverse function of the time that has elapsed since the previous presentations of the item. Based on an analysis of information reoccurrence in a variety of natural environments (J. R. Anderson & Schooler, 1991), the equation for $B_i$ is

$$B_i = \log \left( \sum_{j=1}^{n} t_j^{-d} \right), \quad (2.3)$$

where $t_j$ represents the time since the $j^{th}$ practice of an item that elicits retrieval of chunk $i$, and $d$ is the decay rate. In other words, the activation level of a chunk rises with practice and falls with delay. Finally, ACT–R predicts RT to be an inverse function of association strength:

$$T_i = M + F e^{-A_i}, \quad (2.4)$$

where $F$ is a scale factor and $M$ is the residual RT that is unrelated to retrieval. As shown by J. R. Anderson (1982, pp. 399–400) and J. R. Anderson, Fincham, and Douglass (1999, p. 1122 and Footnote 1), Equation (2.4) yields a power law of practice.

Cohen, Dunbar, and McClelland’s PDP Model

Cohen, Dunbar, and McClelland (1990) proposed a parallel distributed processing (PDP; Rumelhart, Hinton, & McClelland, 1986) model for the Stroop task. In the model, information is represented as a pattern of activation over elementary units. The units send information upwards, from an input level, via a “hidden units” level, to an output level. The output of the model then drives a diffusion–like process that eventually determines the response.

One of the phenomena that the PDP model can account for is the power law of practice. Specifically, Cohen et al. (1990, pp. 345–346) show by simulation that their PDP model produces a power law decrease in both RT mean and RT standard deviation. The authors argue that this result is due to the combination of two effects. First, the model uses an error–driven learning mechanism (i.e., backpropagation; Rumelhart, Hinton, & Williams, 1986). This means that the connection weights of individual units are adjusted as a function of the distances between the actual activation of the output units and the desired, target activation (which could be either 0 or 1). Early in training, the distances to the target values will be high, and hence the adjustments will also be high. As training progresses, the model will learn the appropriate connection weights, and the size of the adjustments will decrease.

Second, the PDP model uses a logistic transformation to map, for each unit $i$, the net input $net_i$ to the activation $A$:

$$A_i = \frac{1}{1 + e^{-net_i}}, \quad (2.5)$$
As long as the net input is finite, \( A_i \) will be unequal to the target values of 0 or 1, effectively ensuring that the process of strengthening connection weights will continue throughout the learning process, even though the absolute size of the adjustments will become smaller and smaller. The combination of diminishing adjustments and the diminishing effect of these adjustments causes a power function speed up with practice.\(^1\)

### 2.3 The Power Law Repealed?

Over the course of the last decade, several researchers have begun to question the ubiquity of the power law. For instance, Rickard (1997) argued that the power law holds when the speed up in RT is due completely to speed up in algorithm execution or to speed up in memory retrieval, but that the power law does not hold when performance includes the transition from algorithm execution to memory retrieval. B. A. Dosher and Lu (2007) showed that learning in a very simple perceptual task can best be described by an exponential function rather than a power function. Finally, Heathcote et al. (2000) demonstrated that the relatively good fit of the power function can be caused by averaging artifacts (see also R. B. Anderson & Tweney, 1997; Myung et al., 2000), and that—on the level of the individual participant—the exponential function may provide a better description of how RT decreases with practice. To account for both power and exponential effects, Heathcote et al. (2000) introduced the APEX function, which describes the learning rate as the sum of power and exponential rates. Brown and Heathcote (2005) proposed that the effect of practice could involve decreased leakage in the process of information accumulation, and show that this mechanism yields an APEX–like practice function.

### 2.4 Interim Conclusion: State of the Field

As the foregoing makes clear, most models assume that practice effects originate from changes in a single underlying process. Although parsimonious and arguably plausible, the empirical veracity of this assumption has not been subjected to a rigorous test. In addition, most researchers focus on how practice decreases mean RT or on how it increases accuracy. Seldom do researchers use a model that accounts for practice–induced changes in mean RT and accuracy simultaneously. Finally, most researchers that study the effects of practice on RT have concentrated on finding the shape—power, exponential, or other—with which mean RT for correct responses changes as a function of practice.

In this article, we take a quite different approach, as we intend to decompose the practice–induced effects on RT and accuracy into its constituent cognitive components. To do so, we apply the Ratcliff diffusion model to the data of a large practice experiment. The results show that the single–process assumption is most likely false, and that the traditional focus on RT for correct responses may not be as informative as previous studies have suggested it to be.

### 2.5 The Ratcliff Diffusion Model

The Ratcliff diffusion model provides a detailed and comprehensive account of performance in speeded two–choice tasks (e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff and Cohen et al. (1990) suggest that for tasks that require the construction of an intermediate representation of the problem, the power law may not hold.
2. Decomposition of Practice

& Tuerlinckx, 2002; Tuerlinckx, 2004; Voss et al., 2004; for recent reviews see Ratcliff & McKoon, 2008; Wagenmakers, 2009). In the Ratcliff diffusion model, from now on simply “diffusion model”, several unobserved psychological processes or parameters give rise to observed behavior, that is, proportion correct, entire RT distributions for correct responses, and entire RT distributions for incorrect responses. In addition, the model provides several non–trivial predictions that have been confirmed by experiment, such as those about the relative speed of correct versus incorrect RTs and its interaction with the speed–accuracy tradeoff (e.g., Swensson, 1968; Ratcliff & Smith, 2004; Wagenmakers et al., 2008).

In the past, the diffusion model has been successfully applied to many two–choice RT paradigms, including lexical decision, short–term and long–term recognition memory tasks, same/different letter–string matching, numerosity judgments, visual–scanning tasks, brightness discrimination, and letter discrimination (e.g., Wagenmakers et al., 2008; Ratcliff & Rouder, 1998, 2000; Ratcliff, 2002; Ratcliff, Van Zandt, & McKoon, 1999). By virtue of its precise, quantitative account of a wide range of observed data in terms of unobserved psychological processes, the diffusion model has effectively raised the bar in every paradigm to which it has been applied (see Wagenmakers, 2009 for a list of specific examples).

Here we describe the diffusion model as it applies to the lexical decision task, a task in which participants have to quickly decide whether a letter string is a word such as party or a nonword such as drapa (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers et al., 2008).

The Wiener Diffusion Process

The core of the diffusion model is the Wiener diffusion process that describes how the relative evidence for one of two response alternatives accumulates over time. The meandering lines in Figure 2.1 illustrate the continuous and noisy accumulation of evidence for a word response over a nonword response, when a word is presented. When the amount of diagnostic evidence for one of the response options reaches a predetermined response threshold (i.e., one of the horizontal boundaries in Figure 2.1), the corresponding response is initiated. The dark line in Figure 2.1 shows how the noise inherent in the accumulation process can sometimes cause the process to end up at the wrong (i.e., nonword) response boundary.

Mathematically, the Wiener diffusion process is described by the following stochastic differential equation (e.g., Gardiner, 2004; P. L. Smith, 2000):

\[
\frac{dX(t)}{dt} = \xi dt + s dW(t),
\]

where \(dX(t)\) is the change in the accumulated evidence \(X\) for a small time interval \(dt\), \(\xi\) is drift rate (i.e., the deterministic component of the noisy process) and \(sdW(t)\) are zero–mean random increments with infinitesimal variance \(s^2 dt\). The factor \(W(t)\) represents the Wiener noise process (i.e., idealized Brownian motion). Thus, the amplitude of the noise in the information accumulation process is governed by parameter \(s\). This parameter is a scaling parameter, which means that if \(s\) doubles, other parameters in the model can be doubled to obtain exactly the same result. Therefore, the choice of a specific value for \(s > 0\) is arbitrary; for historical reasons, \(s\) is usually fixed at 0.1.
2.5. The Ratcliff Diffusion Model

Figure 2.1: The diffusion model as it applies to the lexical decision task. A word stimulus is presented (not shown) and two example sample paths represent the accumulation of evidence resulting in one correct response (light line) and one error response (dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper word boundary is reached more often than the incorrect, lower nonword boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a non–decision component that represents the time needed for processes such as stimulus encoding and response execution.

The Diffusion Model Parameters

The version of the diffusion model that we apply in this article has seven parameters. These are:

1. Mean drift rate ($v$); the across–trial mean of $\xi$. As is evident from Equation 2.6, drift rate quantifies the deterministic component in the information–accumulation process. This means that when the absolute value of drift rate is high, decisions are fast and accurate; when the absolute value of drift rate is low, however, processing is driven to a large extent by noisy fluctuations, and as a result decisions are slow and inaccurate. In the lexical decision task, for example, classification performance for high–frequency words such as chair is better than for low–frequency words such as fume. The diffusion model accommodates this result through a change in drift rate: high–frequency words have a higher drift rate than low–frequency words. Thus, $v$ indexes task difficulty or subject ability.

2. Across–trial variability in drift rate ($\eta$). This parameter reflects the fact that drift rate may fluctuate from one trial to the next, according to a normal distribution with mean $v$ and standard deviation $\eta$. The parameter $\eta$ allows the diffusion model to account for data in which error responses are systematically slower than correct responses (Ratcliff, 1978; Ratcliff et al., 1999).
3. Boundary separation \((a)\). Boundary separation quantifies response caution and modulates the speed–accuracy tradeoff: When the participant is careful not to make a mistake, the boundaries are set wide apart—as a result, the noisy fluctuations inherent in the accumulation of evidence are less likely to result in an incorrect response. The price that has to be paid for this decrease in error rate is an increase in response time. Thus, when boundary separation is large, decisions are slow and accurate; and when boundary separation is small, decisions are fast and inaccurate.

4. Mean starting point \((z)\). Starting point reflects the *a priori* bias of a participant for one or the other response. This parameter is usually manipulated via payoff or proportion manipulations (Edwards, 1965; but see Diederich & Busemeyer, 2006). For instance, Wagenmakers et al. (2008), Experiment 2, used a proportion manipulation in which particular blocks of trials featured three times as many words as nonwords. In the diffusion model, this manipulation causes the starting point to shift towards the word boundary. Such a shift would lead to relatively fast and accurate responding for word stimuli, but relatively slow and inaccurate responding for nonword stimuli.

5. Across–trial variability in starting point \((s_z)\). This parameter reflects the fact that starting point may fluctuate from one trial to the next, according to a uniform distribution with mean \(z\) and range \(s_z\). The parameter \(s_z\) also allows the diffusion model to account for data in which error responses are systematically faster than correct responses.

6. Mean of the nondecision component of processing \((T_{er})\). This parameter encompasses the time that is spent on common processes, i.e., processes that are executed regardless of the stimulus that is presented. Examples of such processes include the time needed to read the word or nonword stimulus in a lexical decision task, or the time needed to execute the key–press motor command that is issued after the decision process has reached a response boundary. The diffusion model assumes that the observed RT is the sum of the nondecision component and the decision component of processing (Luce, 1986):

\[
RT = DT + T_{er},
\]

where \(DT\) denotes decision time. Therefore, the nondecision time \(T_{er}\) does not affect response choice and acts solely to shift the entire RT distribution by a constant amount.

7. Across–trial variability in the nondecision component of processing \((s_t)\). This parameter reflects the fact that nondecision time may fluctuate from one trial to the next, according to a uniform distribution with mean \(T_{er}\) and range \(s_t\). The parameter \(s_t\) also allows the model to capture RT distributions that show a relatively shallow rise in the leading edge.

Many experiments attest to the validity and specificity of the parameters of the diffusion model. For instance, Voss et al. (2004), Ratcliff and Rouder (1998), and Wagenmakers et al. (2008) show that accuracy instructions increase boundary separation, easier stimuli have higher drift rates, and unequal reward rates or presentation proportions are associated with changes in starting point. These and other experiments justify the psychological interpretation of the diffusion model parameters in terms of underlying cognitive processes and concepts.
The Diffusion Model and the Effect of Practice

In this article, we study the extent to which practice affects the parameters of the diffusion model, as this allows us to draw conclusions in terms of the processes postulated by these parameters. The characteristic speed–up with practice could be captured by several parameters of the diffusion model, but the prime candidate for capturing the effect of practice is drift rate. Drift rate quantifies the signal–to–noise ratio in information accumulation and hence reflects the ease with which people process stimuli. In addition, Wagenmakers, Grasman, and Molenaar (2005) have shown that an increase in drift rate causes RT mean and RT standard deviation to decrease at the same rate, and Ratcliff and McKoon (2008, Figure 8) recently showed that an increase in drift rate also causes all RT quantiles to decrease at the same rate, i.e., quantiles from a condition with low drift rate are a linear function of the quantiles of a condition with high drift rate. Identical predictions (i.e., a linear relation between RT mean and RT standard deviation with practice, and a linear relation between RT quantiles with practice) were derived from the ITA (Logan, 1992). In addition, Logan’s ITAM features a random walk decision process that is driven by the sequence of retrieved memory traces (Logan, 2002, page 393); a speed–up in memory retrieval should map on to an increase in drift rate.

In an earlier experiment on aging, Ratcliff et al. (2006b) used the diffusion model to analyze practice effects separately for older and younger adults. Ratcliff et al. (2006b) used two different tasks, a letter discrimination task and a brightness discrimination task, and alternated blocks with speed instructions with blocks with accuracy instructions. As expected, Ratcliff et al. (2006b) found that drift rate increased with practice. However, the results also showed that boundary separation decreased (i.e., participants became less cautious with practice, an effect that was particularly pronounced for older adults). Together, the changes in drift rate and boundary separation accounted for the observed changes in performance with practice. Because each block contained only 96 stimuli and several experimental conditions, the final analyses required that the data be averaged both across participants and across several blocks, such that the practice effect could only be assessed by a comparison of three or four sessions.

In order to apply the diffusion model across many practice blocks and on the level of individual participants, it is necessary to collect a lot of data. Here we used a 10,000–trial lexical decision experiment with 25 blocks of 400 trials each. Two participants were instructed to pay attention primarily to speed, and two participants were instructed to pay attention primarily to accuracy.

2.6 Experiment

Method

Participants

Four undergraduate psychology students participated for course credit. All participants were native Dutch speakers.

Stimulus materials and design

In a lexical decision task, participants were presented with letter strings that had to be classified as “word” (e.g., fume) or “nonword” (e.g., drapa). For this lexical decision task, we selected 200 low–frequency Dutch words, whose frequencies ranged from 0.31 to 5.48 per million (mean frequency = 3.44 × 10^{-6}, SD = 1.29 × 10^{-6}, R. Baayen, Piepenbrock)
A set of 200 pronounceable nonwords was created by replacing a single letter in an existing Dutch word (vowels were replaced by vowels, and consonants by consonants. The words that were used to generate the nonwords were not used as word stimuli). Words and nonwords were approximately matched in length. The set of 200 words and 200 nonwords was the same for all participants and in all blocks of the experiment.

Participants completed five blocks per day on five consecutive days. The 25 blocks of 400 stimuli thus constitute 10,000 trials per participant. Before each block, the same instructions were given. For the accuracy condition, participants A1 and A2 were instructed to respond as quickly and accurately as possible. Feedback was directed towards accurate responding. For the speed condition, participants S1 and S2 were instructed to be fast, but still accurate. Here, feedback was directed towards fast responding.

Procedure
Stimuli were presented on a 17” CRT screen, at about 40 cm distance from the participant, using Presentation for Windows (version 10.3). Letters were presented in lower case, 6 mm in height, white on black. Responses were registered using a two–button response device attached to the computer’s parallel port, to achieve maximum timing accuracy. The experimenter was in the same room as the participants for the entire duration of the experiment.

For participants in the accuracy condition, RTs longer than 2000 ms were followed by the feedback message “TE LANGZAAM!” (i.e., “too slow!”), and RTs shorter than 200 ms were followed by the feedback message “TE SNEEL” (i.e., “too fast”). For RTs in the 200–2000 ms time window, incorrect responses were followed by the feedback message “FOUT” (i.e., “error”), whereas correct responses did not trigger a feedback message. The duration of all feedback messages was 1200 ms. Every trial started with a blank screen that was presented for 250 ms.

For participants in the speed condition, RTs longer than 750 ms were followed by the “too slow” feedback message. No feedback on accuracy was given. In all other respects, the speed condition was identical to the accuracy condition.

For every participant, the series of blocks was preceded by a short training block (with corresponding instructions), consisting of 15 words and 15 nonwords, none of which were also present in the 400 trials of the main experiment. The order of the stimuli was randomized before each block. Participants were given a four–minute break after each block. Each five–block session lasted approximately 60 minutes.

Results
In this section, we first summarize the data using descriptive measures of RT and accuracy, and then analyze the data using the diffusion model. All analyses were conducted on the level of individual participants.

Preprocessing of RT data
Lower bounds for acceptable RTs were determined by visual inspection, which revealed that 250 ms was a reasonable cut–off to eliminate fast guesses. RTs longer than 2000 ms were designated as slow outliers. This filtering resulted in the elimination of 2, 2, 37, and 126 fast guesses for participants A1, A2, S1, and S2, respectively—this means that, for
participant S2, only 1.26% of the data were discarded. In the entire experiment, only a single trial was classified as a slow outlier.

**Descriptives: Emphasis on Accuracy**

The upper panels of Figure 2.2 show the effects of practice on accuracy and RT for participants A1 and A2 (i.e., the accuracy condition). The upper panels show the proportion of correct responses and the lower panels show RT quantiles (.1, .3, .5, .7, and .9) calculated for each block. Additionally, the figures show 95% bootstrap confidence intervals. Accuracy for words was largely constant over blocks, although accuracy for nonwords increased somewhat for participant A1. As expected, both the mean and the spread of the RT distributions showed marked decrease with practice. This pattern was evident both for correct responses (shown in Figure 2.2) and for error responses (shown in the archived materials).

**Descriptives: Emphasis on Speed**

The lower panels of Figure 2.2 show the effects of practice on accuracy and RT for participants S1 and S2 (i.e., the speed condition). Participant S1 clearly improved on accuracy, with constant RT from the fourth block onward. Participant S2 appears to speed up in the first ten blocks and then showed stable RT with increasing accuracy. For both participants, the spread of the RT distributions decreased with practice.

In summary, accuracy–stressed participants A1 and A2 improved mainly on speed, whereas speed–stressed participants S1 and S2 improved mainly on accuracy. For all participants, variability in RT decreased. Note that for all participants, performance in the final blocks was both very fast and, for all but participant S2, very accurate. Accuracy–stressed participants seem to start near maximal accuracy and speed–stressed participants reached their maximum speed after a few blocks of practice.

**Diffusion Model Analyses**

In order to decompose the large changes in RT and accuracy into the latent psychological processes hypothesized by the diffusion model, the model needs to be fit to the data. This can be easily accomplished using one of several freely available software packages, primary among them DMAT (Vandekerckhove & Tuerlinckx, 2007, 2008) and fast–dm (Voss & Voss, 2007, 2008). We analyzed our data with both packages, but—although the results were generally very similar—we obtained the most satisfactory results from a Bayesian implementation of the diffusion model (Vandekerckhove, Tuerlinckx, & Lee, 2008). In particular, we found that the Bayesian estimation routine was able to deal with sparse data, i.e., conditions with relatively few errors, without producing occasional outlier results. In a Bayesian analysis, probability distributions quantify uncertainty about the values of the model parameters. One generally starts with a vague “prior” distribution, which then gets updated by means of the data to yield a “posterior” distribution. This posterior distribution reflects the knowledge about the model parameter after having seen the data (see, e.g., Gelman, Carlin, Stern, & Rubin, 2004).

In the estimation procedure, all parameters were allowed to vary freely across practice blocks, reflecting the exploratory nature of our analysis and the fact that we did

---

2Our Bayesian analysis experienced problems of numerical stability only for the first block of participant A2. This explains why the following sections and graphs do not report any parameter estimates for this particular block of trials.
2. Decomposition of Practice

Figure 2.2: Mean accuracy and RT quantiles (.1, .3, .5, .7, .9) for correct responses per practice block. Grey lines in RT quantiles show 95% bootstrap confidence intervals. Accuracy stressed participants improved on speed. Speed stressed participants mainly improved on accuracy.

not want to commit ourselves to a particular functional form of the practice effect. Within each practice block, drift rates were allowed to vary between words and non-words (Wagenmakers et al., 2008; Ratcliff, Gomez, & McKoon, 2004), and so were the associated trial–to–trial variabilities in drift rate (i.e., the \( \eta \)'s); this latter modeling choice was motivated by the intuition, confirmed in early exploratory analyses, that nonwords, who by definition have no meaning, are more similar to each other than words, who all have different meanings and frequencies. Starting point was modeled as the bias \( (B) \) in favor of words over nonwords, i.e., \( z/a \). More details about the statistical modeling can be found in the archived materials. Information on model fit can be found in the archived materials.
Below, we describe the modeling results, discussing in turn each of the four most important parameters (drift rate $v$, boundary separation $a$, bias $b$, and non-decision time $T_{ed}$) for all participants. We will only briefly mention results for the variability parameters $s_z$, $s_t$, and $\eta$—more extensive results can be found in the archived materials.

**Diffusion model inference: Drift rate**

Figure 2.3 and 2.4 show the posterior distributions of the drift rate parameter ($v$) for each of the four participants, for words and nonwords separately. In these figures and the ones that follow, we visualize the posterior distribution through a color coding scheme; high density regions of the posterior have a darker color than low density regions.

All participants but A2 showed a clear increase in drift rate for both words and nonwords (Figure 2.3 and 2.4). Note that for the participants in the speed condition, drift rate increased even in the later practice blocks.

![Emphasis on Accuracy](image)

Figure 2.3: Posterior distributions of the drift rate parameter across practice blocks (accuracy stressed participants). Dark colors represent high density. White lines are cubic smoothed splines through the medians of the posterior distributions.
Figure 2.4: Posterior distributions of the drift rate parameter across practice blocks (speed stressed participants). Dark colors represent high density. White lines are cubic smoothed splines through the medians of the posterior distributions.

**Diffusion model inference: Boundary separation**

Figure 2.5 shows the posterior distributions of the boundary separation parameter $a$. Participants A1 and A2 from the accuracy condition decreased their response caution throughout the experiment. This decrease in response caution combines with the increase in drift rate to explain why accuracy was approximately constant across practice (cf. Figure 2.2) whereas RT mean and RT variability noticeably decreased.

One might argue that, at least for participants A1 and A2, the errors are mainly caused by attentional lapses. Furthermore, one might argue that the probability of making an error due to an attentional lapse is approximately constant over practice. When RT decreases with practice, these attentional lapses would then lead to a systematic underestimation of boundary separation. According to this account, the observed difference in boundary separation is a statistical artefact caused by model misspecification.
This misspecification account is vulnerable to at least two counter-arguments. First, the archived materials show that both correct RTs and error RTs decrease over time, and in a similar fashion (Correlations of mean correct and error RT over blocks are $r = 0.95$ for A1, and $r = 0.92$ for A2). Second, the RT distributions for error RTs are skewed to the right, and this skew decreases with practice. These phenomena are predicted by the diffusion model, but not by the misspecification account.

For the participants in the speed condition, S1 shows little or no systematic changes in boundary separation, but—just as the participants from the accuracy condition—S2 does show a clear decrease in boundary separation with practice. Also note that, in line with the instructions, the participants in the accuracy condition have larger boundary separation (i.e., more response caution) than the participants in the speed condition.

![Emphasis on Accuracy](image1)

![Emphasis on Speed](image2)

Figure 2.5: Posterior distributions of the boundary separation parameter across practice blocks. Dark colors represent high density. White lines are cubic smoothed splines through the medians of the posterior distributions.
2. Decomposition of Practice

Diffusion model inference: Response bias

Figure 2.6 shows the posterior distributions of the response bias parameter $B$. This parameter gives the height of the starting point $z$ as a proportion of the boundary separation $a$, so that $B = z/a$. Thus, values of $B > .5$ indicate an a priori bias towards the word response, and values of $B < .5$ indicate an a priori bias towards the nonword response. Both participants in the accuracy condition started the experiment with a slight bias in favor of a word response and, over the practice sessions, developed a reverse preference for nonword over word responses. For the participants in the speed condition the practice–induced changes in bias are less systematic.

![Emphasis on Accuracy](image1)

![Emphasis on Speed](image2)

Figure 2.6: Posterior distributions of the response bias parameter across practice blocks. Here, response bias $B$ is defined as $B = z/a$. Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.
Diffusion model inference: Non–decision time

Figure 2.7 shows the posterior distributions of the non–decision time parameter $T_{er}$. For both participants in the accuracy condition, $T_{er}$ decreases with practice (i.e., about 100 ms for both participants). These decreases in $T_{er}$ account for approximately 40% of the total practice–induced decrease in mean RT, which is about 250 ms for both A1 and A2. The participants in the speed condition did not show a systematic decrease in $T_{er}$ with practice, but they did display large block–to–block fluctuations in $T_{er}$ that cover a range of about 100 ms.

![Emphasis on Accuracy](image)

![Emphasis on Speed](image)

Figure 2.7: Posterior distributions of the non–decision time parameter across practice blocks. Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.

Diffusion model inference: Variability parameters

In our Bayesian analyses, we estimated all parameters of the Ratcliff diffusion model, including the parameters that represent trial–to–trial variability in drift rate (i.e., $\eta$),
starting point (i.e., $s_z$), and non–decision time (i.e., $s_t$). The results showed that, as expected, $\eta$ is much higher for words than for nonwords. Also, $\eta$ appears to decrease with practice for word stimuli in the accuracy condition. $s_t$ decreases with practice for all participants but S1. For $s_z$, no structural effects of practice were found. Detailed results regarding the variability parameters can be found in the archived materials.

2.7 Discussion

In contrast to what many models have assumed in the past, our analysis shows that practice may not be a unitary phenomenon that can be attributed to a single underlying mechanism. According to our diffusion model decomposition, practice leads to an increase in the rate of information processing, a decrease in response caution, and a decrease in non–decision time. In addition, participants also exhibit systematic changes in a priori bias.

As mentioned in the introduction, the finding that drift rate increases with practice is consistent with Logan’s Instance Theory of Attention and Memory. It is plausible that this increase in drift rate leads to a decrease in boundary separation—this would allow participants to greatly decrease RT while keeping accuracy relatively constant.

In contrast, the practice–induced reduction of the non–decision component and the fluctuations in response bias were both pronounced and unexpected. It is possible that the reduction in non–decision time is task–specific rather than stimulus–specific, hence reflecting an increased familiarity with the general task–requirements, response buttons, and the processing of visual input and feedback displayed on the computer screen. To examine this possibility, our current work focuses on transfer effects by including both old and new stimuli in the same task.

Power Law Theories Revisited

The goal of our diffusion model analysis was to decompose the practice effect into its constituent processes. However, the diffusion model is not an explanatory model of practice, and it does not describe how practice alters or adds memory representations. Ideally, then, one would like to fit to the data the power law theories discussed in the introduction (i.e., Logan’s instance theory, its successor ITAM, Rickard’s component power laws model, Anderson’s ACT–R, and Cohen et al.’s PDP model). Unfortunately, many of these models are not as explicit about the decision process as the diffusion model. The diffusion model is able to fit entire RT distributions, both for correct and error responses, and it is able to separately estimate components of processing such as non–decision time, response bias, and boundary separation. It is likely that the power law models can be extended to match the performance of the diffusion model, but this is presently not the case.

Note, however, that both ITAM (the Instance Theory of Attention and Memory, Logan, 2002) and EBRW (the Exemplar Based Random Walk model, Nosofsky & Palmeri, 1997; Palmeri, 1997) are very similar to the diffusion model in that they characterize the decision process as a noisy accumulation of relative evidence that terminates when a predefined threshold is reached. When models such as ITAM are extended so as to fit the speed and proportion of error responses, we expect close qualitative agreement between results from the diffusion model and those from ITAM.
Concluding Comments

Our diffusion model analysis suggests that the traditional methods of analysis might provide a false sense of security. Most traditional methods focus on improvements in either mean RT for correct responses or in response accuracy, without any recourse to changes in the underlying processes. Our analysis strongly suggests that the practice effect is the interactive combination of several underlying processes—people do not only improve on stimulus processing, but, at the same time, they are able adjust their response strategy. In combination with changes in non–decision time, these processes generate a data pattern that cannot be usefully abstracted in terms of mean RT. Instead of focusing on the mathematical function that relates practice to mean RT for correct responses (i.e., power, exponential, or APEX), we feel that a model–driven analysis of underlying processes is both more appropriate and more insightful.

2.A The Bayesian Ratcliff Diffusion Model

In Appendix A, we provide details regarding the Bayesian implementation of the Ratcliff diffusion model.

Notation

We will use the following notation. The “twiddles” symbol $\sim$ means “is distributed according to”, and the function $W(x, t|a, b, t_{er}, \xi)$ is the bivariate probability density function (PDF) of the first–passage times of a Wiener diffusion process. In this function, $a$ is the boundary separation, $t_{er}$ is the nondecision time, $\xi$ is the drift rate, and $b$ is the ratio of the starting point of the diffusion to the distance between the boundaries, so that the starting point $z_0 = ab$. The full expression for the PDF can be found in Tuerlinckxs (2004). We will use indexes $p$ for participants, $j$ for blocks, $i$ for items, and $c_i$ for the word/nonword value of item $i$. To avoid confusion with other subscripts, indexes are always put between round brackets. In general, therefore, the vector $y_{(p_ji_c_j)}$, which consists of the accuracy score $x_{(p_ji_c_j)}$ and the response time $t_{(p_ji_c_j)}$, adheres to:

$$p\left(y_{(p_ji_c_j)}\right) = W\left(a_{(p_ji_c_j)}, b_{(p_ji_c_j)}, t_{er(p_ji_c_j)}, \xi_{(p_ji_c_j)}\right).$$

Model

To arrive at the Ratcliff diffusion model, we now add random trial–to–trial variability to the $t_{er}$, $\xi$, and $b$ parameters according to the following mixing distributions:

$$t_{er(p_ji_c_j)} \sim N\left(T_{er(p_j)}, s_{t(p_j)}^2\right),$$

$$b_{(p_ji_c_j)} \sim N\left(B_{(p_j)}, s_{b(p_j)}^2\right),$$

and

$$\xi_{(p_ji_c_j)} \sim N\left(v_{(p_ji_c_j)}, \eta_{(p_ji_c_j)}^2\right).$$

With the addition of these mixing distributions, the model now has nine parameters per block per participant:

- $a$ for the boundary separation
- $B$ for the across–trial mean bias, meaning that the mean starting point $z = aB$
2. Decomposition of Practice

- \( s_b^2 \) for the variability in bias, so that the variance in starting point \( s_z^2 = a^2 s_b^2 \).
- \( T_{cr} \) for the across–trial mean nondecision time
- \( s_t^2 \) for the across–trial variance of the nondecision time
- two \( \nu \) parameters for the mean drift rate of word stimuli and of nonword stimuli
- two \( \eta \) parameters for the trial–to–trial variance in drift rate of word stimuli and of nonword stimuli

**Bayesian methods**

In Bayesian statistics, one makes inferences about the different model parameters through their posterior distributions (Gelman et al., 2004), that is, about \( p(\text{parameter}|\text{data}) \). The method is called Bayesian because the computation of the posterior distribution requires the application of Thomas Bayes’ famous theorem:

\[
p(A|B) = \frac{p(B|A)p(A)}{p(B)},
\]

where in this case the left hand side is the posterior distribution. Then \( p(B|A) \) is the likelihood of the model, \( p(A) \) are the prior distributions of parameters, and \( p(B) \) is a normalizing constant. The computation of the posterior distribution and associated descriptive measures is often very difficult and computationally expensive, but some software packages exist to assist in the implementation. We have used WinBUGS for this purpose (Lunn, Thomas, Best, & Spiegelhalter, 2000; see also Vandekerckhove et al., 2008).

**Priors**

In order to initiate a Bayesian analysis, we have to define the prior distribution of each parameter. For parameter estimation, one typically chooses prior distributions that reflect a very vague knowledge of the possible parameter values (uninformative priors). For the present analysis, we have chosen priors that are somewhat informative, reflecting what we know about plausible values of the different parameters, and they are restricted to a specific interval to avoid numerical over–or underflow. The intervals are chosen to reflect a reasonably wide range of possible parameter values. The priors were defined exactly as follows:

\[
\begin{align*}
a & \sim N(0.08, 1/400) \cdot I(0.02, 0.18), \\
T_{cr} & \sim U(0.01, 0.60), \\
B & \sim N(0.50, 1/200) \cdot I(0.20, 0.80), \\
\nu & \sim U(0.00, 0.90), \\
\eta & \sim U(0.001, 0.4999), \\
s_b & \sim U(0.001, 0.4999), \\
s_t & \sim U(0.001, 0.4999),
\end{align*}
\]

where the second parameter of the normal distribution \( N(\mu, \sigma^2) \) is always its variance, the indicator function \( I(\cdot) \) truncates the distribution between its arguments, and \( U(L, H) \) is the uniform distribution between \( L \) and \( H \). The relatively large overall number of data means that the influence of the shape of the prior distributions is in fact very small. Indeed, we repeated the analysis several times with different priors and encountered no meaningful differences.

28
Convergence and model fit

We ran 6 independent Markov chains of 10,000 iterations (after a burn-in of 500) and tested their convergence with the $\hat{R}$ criterion (Gelman et al., 2014). Convergence ($\hat{R} < 1.1$) was typically fast for the main parameters of the diffusion model ($a$, $T_{er}$, $B$, and $v$). Visual inspection of autocorrelation plots showed high dependency at small lags, but none after downsampling the chains by a factor of 200.

2.B Results for the Variability Parameters

In Appendix B we describe the posterior distributions of the three diffusion model parameters that quantify trial-to-trial variability. Ratcliff and colleagues introduced these trial-to-trial variabilities in order to account for certain key aspects of RT data. That is, the presence of variability in drift ($\eta$) and variability in starting point ($s_z$) allows the model to account for errors that are slower and faster than correct responses, respectively, whereas variability in $T_{er}$ ($s_t$) allows the model to account for a relatively gentle rise in the leading edge of an RT distribution.

Variability in Drift Rate ($\eta$)

Figures 2.8 and 2.9 show the posterior distributions of the variability in drift rate $\eta$, separately for each block and stimulus type. As expected, parameter $\eta$ is generally higher for words than for nonwords. For the accuracy stressed participants, the variability in drift for word stimuli appears to decrease with practice. For the speeded participants, no structural effect on $\eta$ is present. The spread of the posterior distributions is large, especially for variability in word drift rates for the speeded participants.

Variability in Starting Point ($s_z$)

Figure 2.10 shows the posterior distribution of variability in starting point for each block. The estimates are stable over blocks and about equal for all participants.

Variability in Non-decision Time ($s_t$)

Figure 2.11 shows the posterior distribution of variability in non-decision time for each block. For participants A1, A2 and S2, $s_t$ declines within the first 10 blocks of practice, while, for participant S1, $s_t$ does not structurally change with practice.
2. Decomposition of Practice

Figure 2.8: Posterior distributions of the variability in drift rate per stimulus type and across practice blocks (accuracy stressed participants). Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.
2.B. Results for the Variability Parameters

Figure 2.9: Posterior distributions of the variability in drift rate per stimulus type and across practice blocks (speed stressed participants). Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.
2. Decomposition of Practice

Emphasis on Accuracy

<table>
<thead>
<tr>
<th></th>
<th>participant A1</th>
<th></th>
<th>participant A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s^2_f$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>participant S1</th>
<th></th>
<th>participant S2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$s^2_f$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

practice block

Figure 2.10: Posterior distributions of the variability of bias across practice blocks. Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.
2.B. Results for the Variability Parameters

Figure 2.11: Posterior distributions of the variability of non–decision time across practice blocks. Dark colors represent high density. White lines are cubic smoothed splines through medians of posterior distributions.
This chapter has been published as as:
Task–Related vs. Stimulus–Specific Practice: A Diffusion Model Account
Experimental Psychology, 58, 434–442.

Abstract

When people repeatedly practice the same cognitive task, their response times invariably decrease. Dutilh, Vandekerckhove, Tuerlinckx, and Wagenmakers (2009) argued that the traditional focus on how mean RT decreases with practice offers limited insight; their diffusion model analysis showed that the effect of practice is multifaceted, involving an increase in rate of information processing, a decrease in response caution, adjusted response bias, and, unexpectedly, a strong decrease in nondecision time. In this study, we aim to further disentangle these effects into stimulus–specific and task–related components. The data of a transfer experiment, in which repeatedly presented sets and new sets of stimuli were alternated, show that the practice effects on both speed of information processing and time needed for peripheral processing are partly task–related and partly stimulus–specific. The effects on response caution and response bias appear to be task–related. This diffusion model decomposition provides a perspective on practice that is more detailed and more informative than the traditional analysis of mean response times.

When people repeatedly perform the same task, their performance becomes fast, accurate, and relatively effortless. For example, you are able to read this text quickly, virtually without errors, and, hopefully, without investing too much effort. The ease with which you just read the previous sentences is due to your extensive practice with reading in general; however, there is a chance—admittedly remote— that your ease of reading was enhanced because you were familiar with an article that started with the very same two sentences (i.e., Dutilh et al., 2009). This example highlights the distinction between learning that is related to the task (e.g., the skill of reading in general) and learning that
is specific to the stimulus (e.g., reading particular sentences). This distinction is the focus of this article.

Research on skill acquisition and practice has a long history of theorizing and experimenting and involves many different skills in many different domains (e.g., Crossman, 1959; Pirolli & Anderson, 1985; Logan, 1992). Researchers in the field of cognitive skill acquisition have focused mainly on the practice-induced speed-up in mean response time (MRT). This speed-up is typically largest at the initial stage of training and diminishes as practice progresses. The functional form of the speed-up has been the topic of a fierce academic debate. The primary candidate to account for the shape of the speed-up is the power function (i.e., the famous “power law of practice”, A. Newell & Rosenbloom, 1981, p. 3, Logan, 1990). This power law speed-up is consistent with several theories of skill acquisition, most notably Logan’s instance theory (Logan, 1988, 2002). However, other researchers have argued that evidence in favor of the power law might be an artifact of averaging over many different learning curves, and that the true underlying curve is characterized by an exponential speed-up (Heathcote et al., 2000).

Although some theories take into account the entire distribution of response time (RT), the large majority of experimental studies consider only MRT. In recent work we showed that this selective focus on MRT, a focus that ignores the distributional form of RT and its interdependence with accuracy, offers a deceivingly limited view on the effects of practice (Dutilh et al., 2009). By applying Ratcliff’s diffusion model to the data of a 25-block lexical decision task, we were able to decompose the effect of practice into changes in several underlying psychological processes that together determine observed performance. Our main conclusion was that the effect of practice on performance is multifaceted: not only did practice increase the speed of information processing, as expected, but this increase was accompanied by a decrease in response caution and a decrease in the time needed for nondecisional processes. Since the study did not feature transfer blocks with new stimuli, we were not able to conclude whether the practice effects were caused by an increased familiarity with the stimuli, an increased familiarity with the task, or a combination of both.

In this article, we aim to use the diffusion model again to decompose the effects of practice, but this time we also differentiate task–related from stimulus–specific factors (e.g., Forbach, Stanners, & Hochhaus, 1974; Ahissar & Hochstein, 1993; Logan, 1990). In order to do so, we designed an interleaved transfer experiment in which participants responded to alternating blocks of new and repeated stimuli.

The outline of this paper is as follows. We first introduce Ratcliff’s diffusion model. Next we summarize the results of our previous study and emphasize the importance of unraveling task–related and stimulus–specific effects of practice. Then we describe our current transfer experiment and discuss its results both descriptively and in terms of diffusion model parameters.

3.1 The Diffusion Model

The diffusion model provides a formal, comprehensive, and detailed account of performance on speeded two–choice tasks (e.g., Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002; Voss et al., 2004; for recent reviews see Ratcliff & McKoon, 2008; Wagenmakers, 2009). The model accounts not just for MRT but captures the entire distribution of correct and error RTs as well as percentage correct. One of the major strengths of the diffusion model is that its main parameters can be interpreted in terms of latent psychological processes that drive performance.
Here we describe the diffusion model as it applies to the lexical decision task, where a participant has to decide quickly whether a presented letter string is a word (e.g., *party*) or a nonword (e.g., *drapa*). The core of the model is the Wiener diffusion process that describes how the relative evidence for one of two response alternatives accumulates over time. The meandering lines in Figure 3.1 illustrate the continuous and noisy accumulation of evidence for a word response over a nonword response, when a word is presented. When the amount of diagnostic evidence for one of the response options reaches a predetermined response threshold (i.e., one of the horizontal boundaries in Figure 3.1), the corresponding response is initiated. The dark line in Figure 3.1 shows how the noise inherent in the accumulation process can sometimes cause the process to end up at the wrong (i.e., nonword) response boundary.

![Figure 3.1: The diffusion model as it applies to the lexical decision task. A word stimulus is presented (not shown) and two example sample paths represent the accumulation of evidence which result in one correct response (light line) and one error response (dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper word boundary is reached more often than the incorrect, lower nonword boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a nondecision component that represents the time needed for processes such as stimulus encoding and response execution.](image)

The Diffusion Model Parameters

As in Dutilh et al. (2009), the version of the diffusion model that we apply in this article has seven parameters. These are:

1. Mean drift rate ($v$). Drift rate quantifies the deterministic component in the information—accumulation process. This means that when the absolute value of drift rate is high,
decisions are fast and accurate; thus, $v$ indexes task difficulty or subject ability.

2. Across–trial variability in drift rate ($\eta$). This parameter reflects the fact that drift rate may fluctuate from one trial to the next, according to a normal distribution with mean $v$ and standard deviation $\eta$. The parameter $\eta$ allows the diffusion model to account for data in which error responses are systematically slower than correct responses (Ratcliff, 1978).

3. Boundary separation ($a$). Boundary separation quantifies response caution and modulates the speed–accuracy tradeoff: At the price of an increase in RT, participants can decrease their error rate by widening the boundary separation (e.g., Forstmann et al., 2008).

4. Mean starting point ($z$). Starting point reflects the a priori bias of a participant for one or the other response. This parameter is usually manipulated via payoff or proportion manipulations (Edwards, 1965; Wagenmakers et al., 2008; but see Diederich & Busemeyer, 2005).

5. Across–trial variability in starting point ($s_z$). This parameter reflects the fact that starting point may fluctuate from one trial to the next, according to a uniform distribution with mean $z$ and range $s_z$. The parameter $s_z$ also allows the diffusion model to account for data in which error responses are systematically faster than correct responses.

6. Mean of the nondecision component of processing ($T_{er}$). This parameter encompasses the time spent on common processes, i.e., processes executed irrespective of the decision process. The diffusion model assumes that the observed RT is the sum of the nondecision component and the decision component (Luce, 1986):

$$RT = DT + T_{er},$$

where $DT$ denotes decision time. Therefore, nondecision time $T_{er}$ does not affect response choice and acts solely to shift the entire RT distribution.

7. Across–trial variability in the nondecision component of processing ($s_t$). This parameter reflects the fact that nondecision time may fluctuate from one trial to the next, according to a uniform distribution with mean $T_{er}$ and range $s_t$. The parameter $s_t$ also allows the model to capture RT distributions that show a relatively shallow rise in the leading edge.

Many experiments attest to the validity and specificity of the parameters of the diffusion model. For instance, Voss et al. (2004), Ratcliff and Rouder (1998), and Wagenmakers et al. (2008) show that accuracy instructions increase boundary separation, easier stimuli have higher drift rates, and unequal reward rates or presentation proportions are associated with changes in starting point. Moreover, simulation studies have shown that the parameters of the diffusion model are well identified (e.g., Ratcliff & Tuerlinckx, 2002; Wagenmakers, Van der Maas, & Molenaar, 2003). Finally, Ratcliff (2002) has shown that the model fits real data but fails to fit fake but plausible data. These and other studies justify the psychological interpretation of the diffusion model parameters in terms of underlying cognitive processes and concepts.
A Diffusion Model Account of Practice

In Dutilh et al. (2009), participants completed a 10,000-trial lexical decision task that consisted of 25 blocks administered in five sessions on five consecutive days. Each block consisted of the same set of 200 unique words and 200 unique nonwords. Contrary to the standard interpretation of the practice effect as a unitary phenomenon, a diffusion model decomposition of the data suggested that the practice effect was multifaceted: not only did practice increase the participants’ rate of information processing $v$, but it also caused a decrease in response caution $a$. This combination of effects allowed participants to speed up with practice, while retaining the same level of accuracy. Furthermore, practice also resulted in a change in the $a$ priori preference $z$ for the “word” vs. “nonword” response; over the course of the experiment, the slight $a$ priori preference for “word” responses shifted to a slight $a$ priori preference for “nonword” responses. A final and unexpected effect of practice was to decrease the nondecision time $T_{er}$. This nondecision component decreased by about 100 ms over the course of the entire experiment, a decrease that accounted for about 40% of the total practice-induced change in MRT.

Task–Related vs. Stimulus–Specific Effects

One important question not addressed in Dutilh et al. (2009) is the extent to which the practice-induced changes in the diffusion model parameters are task–related or stimulus–specific. Because the Dutilh et al. (2009) study did not feature any transfer blocks consisting of new stimuli, it is impossible to claim with certainty that, say, the practice-induced decrease in nondecision time $T_{er}$ is exclusively due to an increase in familiarity with the stimuli; instead, an alternative explanation is that the effect on $T_{er}$ is due to an increase in familiarity with the task (i.e., the experimental setup including the button boxes, the computer screen, and the pace at which the trials were presented). This distinction between task–related and stimulus–specific learning is also important in light of existing theories of practice, as we explain below.

Most theories of skill acquisition and learning assume that practice enhances the association between individual stimuli and the appropriate response category (e.g., Logan’s instance theory of automatization, Logan, 1988; its successor, the instance theory of attention and memory, Logan, 2002; and Rickard’s component power laws model, Rickard, 1999). This assumption maps naturally onto the increase in drift rate that we found with practice (e.g., Logan, 2002, p. 391). For our current study, it is important to note that this stimulus–specific explanation of the practice effect on drift rate implies that it does not transfer to new stimuli.

Furthermore, Logan (1992) suggests that “intercept processes”, such as perceptual registration and response execution might contribute to the practice-induced speed-up in MRT. This suggestion is supported by the practice-induced decrease in $T_{er}$ reported in Dutilh et al. (2009). Although this has not been explicitly stated, Logan’s “intercept processes” (i.e., $T_{er}$ in the diffusion model) are most easily conceived of as task–related rather than stimulus–specific. Therefore, we expected the practice effect on $T_{er}$ to transfer to new stimuli without loss.

In order to assess the relative contribution of task–related vs. stimulus–specific learning, we conducted a new experiment in which transfer blocks with unique stimuli were presented in alternation with a single block of repeated stimuli.
3. Task–Related vs. Stimulus–Specific Practice

3.2 Method

Participants
Eight native speakers of Dutch each participated for course credit on four consecutive days in sessions that took approximately one hour.

Materials
We composed 11 lists of 200 unique words and 200 unique nonwords. The words were selected from the Subtlex database (Brysbaert & New, 2009). The average frequency per million of the 22,000 words was 1.23 (SD = 0.81). Each list consisted of 40 four–letter words (mean frequency = 1.10, SD = 0.83), 80 five–letter words (freq = 0.99, SD = 0.80) and 80 six-letter words (freq = 1.54, SD = 0.69). The 22,000 nonwords were created using the WUGGY program (Keuleers & Brysbaert, 2010), which constructed pronounceable nonwords by replacing one letter in an existing Dutch word, vowel by vowels, consonants by consonants. The words that served as the basis for constructing the nonwords were again selected from the Subtlex database but none of them were also present in the word lists. In similar fashion a set of 15 words and 15 nonwords was created to serve as a practice list.

Design
In the experimental procedure, we distinguish 10 stimulus lists that are each presented once (labeled A through J) and one stimulus list that is presented repeatedly (labeled R). The trials were presented in blocks, each of which contained one complete stimulus list that was shuffled before presentation. Blocks from stimuli lists R and A through J were presented as follows: A R B R C R D R E R F R G R H R I R J R, thus alternating blocks containing new stimuli, presented only once, with blocks containing the repeatedly presented stimulus list.

The blocks were distributed over 4 sessions. The first session started with the presentation of the practice list and featured four experimental blocks. The next two sessions each featured five blocks and the final session featured six blocks.

Procedure
Stimuli were presented on a 17-inch CRT screen about 40 cm from the participant, using the Presentation software for Windows (Version 10.3). Letters were presented in lowercase font, 6 mm in height, in white on a black background. Responses were registered using a two-button response device attached to the computers parallel port to achieve maximum timing accuracy. The experimenter was in the same room as the participants for the entire duration of the experiment.

The stimuli remained on the screen until a response was made. Feedback and instruction to participants promoted accurate but fast responding: responses slower than 2000 ms were followed by the message “TE LANGZAAM” (i.e., “too slow”), and responses faster than 200 ms were followed by the message “TE SNEL” (i.e., “too fast”). Correct responses within the 200–2000 ms window triggered no feedback at all, whereas error responses were followed by the message “FOUT” (i.e., “incorrect”). The duration of the feedback was 1200 ms. Each trial started with a blank screen that was presented for 250 ms.
3.3 Results

Preliminary inspection of the data revealed that one participant behaved erratically throughout the experiment, with accuracy oscillating between 83% and 94%, and an anomalous but substantial increase in mean RT with practice. In addition, this participant’s mean RT showed a 250 ms shift upwards between the third and fourth session. Therefore, we omit this participant’s data here. A plot of the data from this anomalous participant can be found in the online appendix available on the first author’s website. This website also contains the raw data of all participants. Below, we first present the descriptive results and next present the results from our diffusion model decomposition.

Outlier Removal

Before analyzing the data, we determined lower and upper bounds to exclude outlier RTs. Visual inspection revealed that 250 ms and 2000 ms were reasonable cut-offs for fast and slow outliers, respectively. Out of all 56,000 responses in the entire experiment, only 33 responses were excluded as slow outliers and 20 were excluded as fast outliers.

Descriptive Results

Figure 3.2 shows the RT and accuracy data averaged over participants. The left and right column of panels represent the responses to word stimuli and nonword stimuli, respectively. The upper four panels show for each practice block the RT distribution summarized by the .1, .3, .5, .7, and .9 quantiles, as averaged over participants. The upper two RT panels contain correct RTs, and the lower two RT panels contain error RTs. In each panel, the white symbols show the data of the repeated stimulus list R, whereas the black symbols show the data of the new lists that were only presented once. Note that the first repeated block is colored black, since at the first presentation, the stimuli in this block were never seen in the experiment before. The 95% confidence intervals are based on the participant × practice block interaction error term in a within-subjects ANOVA. As shown in Loftus and Masson (1994), these confidence intervals can be used to infer the statistical significance of a within-subjects effect, in this case, the effect of practice.

As expected, both mean and spread of the RT distributions decrease with practice. In fact, the entire RT distribution—from the leading edge to the tail—decreases with practice, for both words and nonwords, and for correct and error responses. This general effect is present for the repeated blocks and, to a lesser extent, also for the blocks containing new stimuli, from now on referred to as transfer blocks.

The bottom left and bottom right panels show the average mean accuracy over participants for words and nonwords, respectively. For word stimuli, practice slightly increases accuracy for the repeated block and slightly decreases accuracy for the transfer blocks. For nonword stimuli, practice has little effect on accuracy.

Diffusion Model Analyses

In order to obtain insight in the processes that underly the data patterns described above, we fitted the diffusion model to the data. We chose to fit the model using the $\chi^2$ method,
one of the methods in the freely available DMAT package in Matlab (Vandekerckhove & Tuerlinckx, 2007, 2008).

As in Dutilh et al. (2009), we did not restrict the parameters over participants or blocks, that is, every participant × block cell was treated as an independent condition, in which all seven parameters of the diffusion model were estimated separately. Furthermore, we allowed words and nonwords to have different drift rates $v$ and different variability of drift rate $\eta$. Starting point $z_0$ was modeled as a fraction of boundary separation $a$, i.e., as relative bias $B$. This method of analysis is consistent with its exploratory nature; as yet, little is known about the functional form by which the diffusion model parameters may change with practice—in fact, little is known about the extent to which the diffusion model parameters change at all.

Figure 3.3 shows the most important parameters of the diffusion model—drift rate $v$, boundary separation $a$, response bias $B$ and nondecision time $T_{er}$ along with variability in nondecision time, $s_t$—as they change with practice. In each panel and the black lines represent the parameter estimates for the transfer blocks. The white lines represent the parameter estimates for repeatedly presented blocks. Again, the first repeated block is colored black, since this is its first presentation. The parameter estimates are averaged over participants. As in Figure 3.2, the 95% confidence intervals are based on the within subjects error term in a repeated measures ANOVA conducted on the new and repeated blocks separately. We discuss the parameters in turn below.

We used Bayesian linear regression to determine statistically whether the diffusion model parameters were affected by practice and whether this practice effect depended on the repeated presentation of stimuli (Gelman & Hill, 2007). Specifically, we conducted regression analyses in which each diffusion model parameter served as criterion. The predictors in these regression analyses were practice block number and a binary variable indicating whether the block contained new or repeated stimuli. The resultant posterior distributions for the regression weights reflect the uncertainty in their estimate and allow one to calculate the probability that a regression weight is larger than zero (i.e., the mass of the posterior distribution to the right of zero). When this probability for the interaction term was smaller than .95 we excluded that term and interpreted the regression model without the interaction.

Diffusion model inference: Drift rate

The upper two panels of Figure 3.3 show the effect of practice on drift rate $v$. The left and right panel show drift rates for word and nonword stimuli, respectively. Drift rate clearly increases for repeatedly presented word–stimuli. This increase is strongest over the first five blocks and levels off later in practice. For new word–stimuli, however, drift rate appears to be more or less stable. This interaction between practice and repeated presentation is confirmed by a regression weight ($\hat{\beta}$) of the interaction that is greater than 0 ($Pr(\beta > 0) \approx 1$). For nonword–stimuli, drift rates appear to increase with practice, no matter whether the stimuli were presented only once or repeatedly (the main practice effect’s $\beta$ is greater than 0, $Pr(\beta > 0) \approx 1$). Note, however, that the between–subjects variance is rather large for the repeated nonwords, as expressed by the larger confidence intervals.

\footnote{Our preferred method of analysis—Bayesian parameter estimation—was plagued by numerical problems causing slow sampling and slow convergence.}
3.3. Results

Diffusion model inference: Boundary separation

The middle panel of the left column of Figure 3.3 shows the practice effect on boundary separation $a$. As in the Dutilh et al. (2009) study, we find that the increase in drift rate described above is accompanied by a clear decrease in response caution that persists over the entire experiment ($Pr(\beta > 0) \approx 1$). The combination of increased drift rate and lowered response caution explains how participants became faster while retaining more or less constant accuracy. Note that the setting of response caution appears independent of the type of stimuli, either new or repeated.

Diffusion model inference: Response bias

The middle panel of the right column of Figure 3.3 shows the practice effect on response bias $B$. For repeated and transfer blocks, response bias does not change systematically with practice. On average, participants display a slight a priori preference for the “word” over the “nonword” response.

Diffusion model inference: Nondecision time

The lower left panel of Figure 3.3 shows the practice effect on nondecision time $T_{er}$. For both repeated and new stimuli, nondecision time decreases over the entire experiment. Specifically, the average decrease in nondecision time is about 50 ms, which is about 20% of the total practice induced decrease in RT. Parameter $T_{er}$ decreases more strongly for repeated stimuli than for new stimuli (the interaction’s $\beta$ is greater that 0, $Pr(\beta > 0) \approx 1$).

Diffusion model inference: Variability parameters

The lower right panel of Figure 3.3 shows how variability in nondecision time $s_t$ decreases with practice. We found that this effect is more pronounced for repeated than for unique stimuli (interaction’s $\beta$ is greater than 0, $Pr(\beta > 0) \approx 1$). Figures for the two remaining variability parameters of the diffusion model (i.e., variability in drift rate $\eta$ for words and nonwords and $s_z$) are omitted here but can be found on the first author’s website. Variability in drift rate $\eta$, estimated for words and nonwords separately, does not change systematically with practice. Variability in bias ($s_z$) does appear to decrease systematically with practice ($Pr(\beta < 0) = 0.90$).

Discussion

In this study, we used Ratcliff’s diffusion model to disentangle the effects of practice into their task-related and stimulus-specific components, components that were confounded in the design by Dutilh et al. (2009). Here we conducted a transfer experiment in which blocks of new stimuli alternated with blocks of repeatedly studied stimuli. In this transfer experiment, stimulus-specific learning is indicated when performance on repeated stimuli benefits from practice but performance on new stimuli does not. In contrast, task-related learning is indicated when performance on repeated stimuli benefits from practice just as much as performance on new stimuli.

The descriptive results showed that practice benefits performance on both repeatedly presented and new stimuli. With practice, the entire RT distribution shrinks and the leading edge shifts downward; this occurs for both new and repeated stimuli, but more so for repeated stimuli. Effects on accuracy were small and only present for word stimuli.
In terms of the diffusion model parameters, we found that nonword drift rate benefits from practice independently of the stimuli presented whereas word drift rate benefits for repeated stimuli only. This result suggests that the task of deciding that a letter string is a nonword is a skill that improves with practice. On the other hand, recognizing a letter string as an existing word is a skill that has been developed before the experiment but can still benefit from increased stimulus familiarity. This result implies that the practice-induced increase in drift rate reported in Dutilh et al. (2009) and Ratcliff et al. (2006b) might be partly stimulus-specific and partly task-related.

Next, we found that response caution decreased with practice at an equal rate for both new and repeated stimuli. This general decrease in response caution in combination with increasing drift rates explains why accuracy remains stable with decreasing RT. It also explains why responses on new word stimuli, for which drift rate remained stable, became less accurate with practice.

Furthermore, nondecision time $T_{er}$ decreased with practice, for both repeated and transfer blocks. This implies that the practice effect on nondecision time is partly task-related, supporting Logan’s idea of a speed–up in “intercept processes” such as perceptual registration and response execution. However, the decrease of $T_{er}$ was stronger for repeated stimuli than for uniquely presented stimuli. This suggests that some perceptual processes might benefit from word familiarity. Further study is needed to clarify how this benefit of familiarity can occur without a concomitant increase in accuracy (remember that $T_{er}$ does not affect accuracy).

In sum, the findings reported here replicate those from Dutilh et al. (2009): practice is a multifaceted phenomenon that involves changes in speed of information processing, response caution, and nondecision processing. The experimental design of the current study allowed us to further decompose these effects into task–related and stimulus–specific effects. Our analysis showed that effects on drift rate and nondecision time are partly task–related and partly stimulus–specific, whereas the effects on response caution and response bias are mostly task–related. In general, we conclude that the model–based decomposition of practice brings insights much deeper than those provided by a standard analysis of mean response times.
Figure 3.2: Practice-induced changes in RT distributions (separately for word and nonword stimuli, and for correct and error responses)

and response accuracy. All R blocks (white) contain the same stimuli, whereas the A-J blocks (black) are transfer blocks that contain unique stimulus lists. See text for details.
Figure 3.3: Practice–induced changes in the five most important parameters from the diffusion model. All R blocks (white) contain the same stimuli, whereas the A–J blocks (black) are transfer blocks that contain unique stimulus sets. See text for details.
Chapter 4

Testing Theories of Post–Error Slowing

Abstract

People tend to slow down after they make an error. This phenomenon, generally referred to as post–error slowing, has been hypothesized to reflect perceptual distraction, time wasted on irrelevant processes, a priori bias against the response made in error, increased variability in a priori bias, or an increase in response caution. Although the response caution interpretation has dominated the empirical literature, little research has attempted to test this interpretation in the context of a formal process model. Here we used the drift diffusion model to isolate and identify the psychological processes responsible for post–error slowing. In a very large lexical decision data set we found that post-error slowing was associated with an increase in response caution and—to a lesser extent—a change in response bias. In the present data set, we found no evidence that post–error slowing is caused by perceptual distraction or time wasted on irrelevant processes. These results support a response monitoring account of post–error slowing.

"What does a man do after he makes an error?" This question is just as valid as when it was first articulated by Rabbitt and Rodgers (1977) over 30 years ago. One answer to this question is that, after he has made an erroneous decision, a man slows down on his next decision—an empirical regularity known as post–error slowing (PES; Laming, 1968, 1979b, 1979c; Rabbitt, 1966, 1979; Rabbitt & Rodgers, 1977). However, this answer raises a new and more interesting question, namely, why does a man slow down after he makes an error? Various answers have been proposed and one of the main goals of this article is to implement these answers in a formal model of decision making so as to compare their adequacy in a precise and quantitative fashion.
4. Testing Theories of Post–Error Slowing

The competing explanations for PES, detailed in the next section, are (1) increased response caution; (2) \textit{a priori} bias away from the response that was just made in error; (3) an overall decrease in the across-trial variability of \textit{a priori} bias; (4) distraction of attention; (5) delayed startup due to irrelevant processes (e.g., overcoming disappointment). We propose that these five explanations map on uniquely to parameters in a drift diffusion model for response time (RT) and accuracy (\cite{Ratcliff, 1978; Ratcliff & McKoon, 2008}). As we will explain below, this one-to-one mapping between psychological processes and model parameters allows an informative diffusion model decomposition of PES and a rigorous assessment of the extent to which each explanation (or indeed any combination of them) holds true.

A major practical obstacle that we needed to overcome is that the drift diffusion model requires relatively many observations to produce informative parameter estimates; as a rule of thumb, the model requires at least 10 error RTs in each experimental condition. Because the interest here centers on trials that follow an error, this means that the model requires at least 10 errors that immediately follow an error. With an error rate of 5\% throughout, the minimum number of observations is already 4,000. Thus, a reliable diffusion model decomposition of PES requires a relatively large data set (or a data set with many errors). Here we fit the model to a lexical decision data set featuring 39 participants who each completed 28,074 trials of speeded two–choice decisions (\cite{Keuleers, Brysbaert, and New, 2010}).

In the next sections we briefly discuss the different explanations for PES and formalize these predictions in the context of the drift diffusion model. We then test the different explanations by fitting the model to the lexical decision data from \cite{Keuleers, Brysbaert, and New, 2010}.

4.1 Explanations for Post–Error Slowing

Over the years, several explanations have been proposed to account for PES. The first explanation (i.e., \textit{increased response caution}) is that an error prompts people to accumulate more information before they initiate a decision. The underlying idea is that people can adaptively change their response thresholds—becoming slightly less cautious after a correct response, and more cautious after an error—and thereby self–regulate to an optimal state of homeostasis characterized by fast responses and few errors (e.g., \cite{Botvinick, Braver, Barch, Carter, and Cohen, 2001; Cohen, Botvinick, and Carter, 2001; Brewer & Smith, 1989; Fitts, 1966; Rabbitt & Rodgers, 1977; G. A. Smith & Brewer, 1995; Vickers & Lee, 1998}). This explanation is so appealing that it is often assumed to be correct without further testing. That is, PES is often interpreted as a direct measure of cognitive control. Conclusions about cognitive control are then based on associations between PES and physiological measures such as anterior cingulate activity (\cite{Li, Huang, Constable, and Sinha, 2006; Danielmeier, Eichele, Forstmann, Tittgemeyer, and Ulisperger, 2011}), error–related negativity (ERN) and positivity (Pe, \cite{Hajcak, McDonald, and Simons, 2003}), or cortisol levels (\cite{Tops & Boksem, 2010}). Alternatively, conclusions about cognitive control may be based on a comparison of PES between clinical groups (e.g., \cite{Shiels & Hawk, 2010}).

The second explanation (i.e., \textit{a priori bias}) is that people become negatively biased against the response option that was just executed in error (e.g., \cite{Laming, 1968; Laming, 1978; Rabbitt & Rodgers, 1977}). This implies that errors facilitate response alternations and hinder response repetitions, both with respect to response speed and probability of occurrence.
4.2 A Drift Diffusion Model Decomposition of Response Times

The third explanation (i.e., decreased variability in bias) is that, following an error, people more accurately control the timing of the onset of information accumulation. The idea, first promoted by Laming (1968, 1979a), is that in speeded RT tasks people often start to sample information from the display even before the stimulus is presented. This advance sampling of stimulus-unrelated information induces trial-to-trial variability in a priori bias. This variability may cause fast errors, and therefore a cautious participant starts the information accumulation process at stimulus onset, but not before.

The fourth explanation (i.e., distraction of attention) is that the occurrence of an error is an infrequent, surprising event that distracts participants during the processing of the subsequent stimulus (Notebaert et al., 2009). Thus, the error-induced distraction contaminates the process of evidence accumulation.

The fifth explanation (i.e., delayed startup) is that errors delay the start of evidence accumulation on the next trial—for instance, participants might need time after an error to re-assess their own performance level and to overcome disappointment (Rabbitt & Rodgers, 1977).

In the literature, the first explanation of PES (i.e., increased response caution) has always been the most dominant. Many studies that associate post–error slowing with cognitive control affirm this association simply by citing Rabbitt (1966). However, Rabbitt (1966, p. 272) concluded that his data “do not allow a choice between possible explanations”. Other studies did not test the competing explanations in a rigorous and quantitative manner (but see White, Ratcliff, Vasey, & McKoon, 2010). Here we set out to test the above five explanations in the context of what is arguably the most popular and successful model for response times and accuracy, the drift diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008).

4.2 A Drift Diffusion Model Decomposition of Response Times

In the analysis of speeded two-choice tasks, performance is usually summarized by mean RT and proportion correct. Although concise, this summary ignores important aspects of the data and makes it difficult to draw conclusions about the underlying cognitive processes that drive performance (Wagenmakers, Van der Maas, & Grasman, 2007). A more detailed and more informative analysis takes into account the entire RT distributions for both correct and error responses, in addition to proportion correct. These RT distributions can be analyzed with the help of formal models — here we focus on the drift diffusion model.

The drift diffusion model has been successfully applied to a wide range of experimental tasks including brightness discrimination, letter identification, lexical decision, recognition memory, signal detection and the implicit association test (e.g., Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2003; Ratcliff et al., 2006b; Wagenmakers et al., 2008; Dutilh et al., 2009; Ratcliff, Thapar, & McKoon, 2010; Klauser, Voss, Schmitz, & Teige-Mocigemba, 2007; van Ravenzwaaij, Van der Maas, & Wagenmakers, 2011). In these tasks and others, the model has been used to decompose the behavioral effects of phenomena such as practice (Dutilh et al., 2009; Dutilh, Krypotos, & Wagenmakers, in press; Petrov, Horn, & Ratcliff, in press), aging (Ratcliff, Thapar, & McKoon, 2001; Ratcliff et al., 2006b, 2010), psychological disorders (White, Ratcliff, Vasey, & McKoon, 2009, 2010a; White et al., 2010b), sleep deprivation (Ratcliff & Van Dongen, 2009), intelligence (Ratcliff, Schmiedek, & McKoon, 2008; Schmiedek, Oberauer, Wilhelm, Suss, & Wittmann, 2007; van Ravenzwaaij, Brown, & Wagenmakers, 2011), and so forth.
The success of the drift diffusion model is due to several factors. First, the drift diffusion model takes into account not just mean RT but considers entire RT distributions for correct and error responses; second, the drift diffusion model generally provides an excellent fit to observed data with relatively few parameters free to vary; third, the drift diffusion model accounts for many benchmark phenomena (Brown & Heathcote, 2008; but see Pratte, Rouder, Morey, & Feng, 2010); fourth, the model allows researchers to decompose observed performance into constituent cognitive processes of interest; finally, evidence accumulation in the drift diffusion model has been linked to the dynamics of neural firing rates, showing that diffusion-like processes can be instantiated in the brain (e.g., Gold & Shadlen, 2002, 2007). Additional advantages (and limitations) of a diffusion model analysis are discussed in more detail in Wagenmakers (2009).

Here we briefly introduce the drift diffusion model as it applies to the lexical decision task, a task where participants have to decide quickly whether a presented letter string is a word (e.g., party) or a nonword (e.g., drapa). The core of the model is the Wiener diffusion process that describes how the relative evidence for one of two response alternatives accumulates over time. The meandering lines in Figure 4.1 illustrate the continuous accumulation of noisy evidence following the presentation of a word stimulus. When the amount of diagnostic evidence for one of the response options reaches a predetermined response threshold (i.e., one of the horizontal boundaries in Figure 4.1), the corresponding response is initiated. The dark line in Figure 4.1 shows how the noise inherent in the accumulation process can sometimes cause the process to end up at the wrong (i.e., nonword) response boundary.

The standard version of the drift diffusion model decomposes RTs and proportion correct in seven different parameters:

1. Mean drift rate ($v$). Drift rate quantifies rate of information–accumulation from the stimulus. This means that when the absolute value of drift rate is high, decisions are fast and accurate; thus, $v$ relates to task difficulty or subject ability.

2. Across–trial variability in drift rate ($\eta$). This parameter reflects the fact that drift rate may fluctuate from one trial to the next, according to a normal distribution with mean $v$ and standard deviation $\eta$. The parameter $\eta$ allows the drift diffusion model to account for data in which error responses are systematically slower than correct responses (Ratcliff, 1978).

3. Boundary separation ($a$). Boundary separation quantifies response caution and modulates the speed–accuracy tradeoff: At the price of an increase in RT, participants can decrease their error rate by widening the boundary separation (e.g., Forstmann et al., 2008).

4. Mean starting point ($z$). Starting point reflects the a priori bias of a participant for one or the other response. This parameter is usually manipulated via payoff or proportion manipulations (Edwards, 1965; Wagenmakers et al., 2008; but see Diederich & Busemeyer, 2003). Here, we report $z$ as a proportion of boundary separation $a$, referred to as bias $B$.

5. Across–trial variability in starting point ($s_z$). This parameter reflects the fact that starting point may fluctuate from one trial to the next, according to a uniform distribution with mean $z$ and range $s_z$. The parameter $s_z$ also allows the drift diffusion model to account for data in which error responses are systematically faster than correct responses (Laming, 1968; Ratcliff & Rouder, 1998). Analogous to the transformation of $z$ to $B$, $s_z$ is often transformed to $s_B$. 

50
4.2. A Drift Diffusion Model Decomposition of Response Times

Figure 4.1: The drift diffusion model as it applies to the lexical decision task. A word stimulus is presented (not shown) and two example sample paths represent the accumulation of evidence which result in one correct response (light line) and one error response (dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper word boundary is reached more often than the incorrect, lower nonword boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a non–decision component that represents the time needed for processes such as stimulus encoding and response execution.

6. Mean of the non–decision component of processing ($T_{er}$). This parameter encompasses the time spent on common processes, i.e., processes executed irrespective of the decision process. The drift diffusion model assumes that the observed RT is the sum of the non–decision component and the decision component (Luce, 1986):

$$RT = DT + T_{er},$$

where $DT$ denotes decision time. Therefore, non–decision time $T_{er}$ does not affect response choice and acts solely to shift the entire RT distribution.

7. Across–trial variability in the non–decision component of processing ($s_{t}$). This parameter reflects the fact that non–decision time may fluctuate from one trial to the next, according to a uniform distribution with mean $T_{er}$ and range $s_{t}$. The parameter $s_{t}$ also allows the model to capture RT distributions that show a relatively shallow rise in the leading edge (Ratcliff & Tuerlinckx, 2002).

As noted above, one of the strengths of the drift diffusion model is that it allows us to decompose observed performance into several latent psychological processes. Such a decomposition relies on the validity of the mapping between model parameters and the postulated psychological processes. Fortunately, many experiments attest to the
specificity and reliability of the model parameters. For instance, Voss et al. (2004), Ratcliff and Rouder (1998), and Wagenmakers et al. (2008) showed that accuracy instructions increase boundary separation, easier stimuli have higher drift rates, and unequal reward rates or presentation proportions are associated with changes in starting point. Moreover, simulation studies have shown that the parameters of the diffusion model can be estimated reliably (e.g., Ratcliff & Tuerlinckx, 2002; Wagenmakers, Van der Maas, & Molenaar, 2005). Finally, Ratcliff (2002) has shown that the model fits real data but fails to fit fake but plausible data.

4.3 From Process to Parameter: A Drift Diffusion Model Perspective on Post–Error Slowing

Many recent applications of the drift diffusion model have been exploratory in nature; for instance, researchers have used the drift diffusion model to study the psychological processes that change with practice (Dutilh et al., 2009; Dutilh, Krypotos, & Wagenmakers, in press), sleep–deprivation (Ratcliff & Van Dongen, 2009), hypoglycemia (Geddes et al., in press), and dysphoria (White et al., 2009, 2010a), but this work was seldom guided by strong prior expectations and theories. This is different in the case of PES, perhaps because explanations for PES have originated in part from a framework of sequential information processing (e.g., Laming, 1979a). Therefore, the competing explanations for PES—in terms of the cognitive processes that change after an error—can be mapped selectively to different parameters in the drift diffusion model, as is shown in Figure 4.2.

Figure 4.2: Cognitive process explanations for PES map on uniquely to different parameters from the drift diffusion model. See text for details.

Thus, the cognitive process explanation of increased response caution maps onto an increase in boundary separation \( \alpha \); the explanation of \textit{a priori} bias corresponds to a shift in bias \( B \) away from the boundary that was just reached in error; the explanation of decreased variability in bias translates to a decrease in across-trial variability \( s_x \); the explanation of distraction of attention entails a decrease in mean drift rate \( v \); and, finally, the explanation of delayed startup is associated with an increase in mean non–decision time \( T_{er} \). The unique link between process and parameter means that competing explanations for PES can be rigorously tested in any particular paradigm, as long as the drift
4.4 Method

The present data set was originally collected to validate a new measure for word frequency (i.e., SUBTLEX–NL; Keuleers, Brysbaert, & New, 2010). Each of 39 participants contributed 28,074 lexical decisions for a grand total of 1,094,886 decisions. Half of the stimuli were uniquely presented words and the other half were uniquely presented nonwords. The word stimuli were selected from the CELEX database (R. H. Baayen, Piepenbrock, & Van Rijn, 1993) and the nonword stimuli were created with the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010).

The experiment was presented in blocks of 500 trials with a self-paced break after every 100 trials. Each trial started with a 500 ms fixation period. The stimulus was then presented until the participant responded, up to a maximum of 2000 ms. A new trial started 500 ms after the response. Participants received feedback about their accuracy after each block of 500 trials. Importantly, participants did not receive trial-by-trial feedback concerning errors. This means that any post-error effects are not contaminated by the possibly distracting presence of error feedback. A more detailed description of the experimental methods is presented in Keuleers, Brysbaert, and New (2010) and Keuleers, Diependaele, and Brysbaert (2010).

The enormous amount of lexical decision trials in this data set contains a commensurate amount of errors; across all participants, 118,566 trials (i.e., 10.80%) were made in error. This abundance of errors allowed us to examine the explanations for PES across various conditions. Specifically, we were able to compare post-error effects separately for nonword stimuli and for word stimuli of varying word frequencies. That is, we used word frequency (based on SUBTLEX) to divide all words into six equally large bins, the five cut points being 0.11, 0.48, 1.33, 3.73, and 14.16 occurrences per million.

4.5 Results and Discussion

Below we first discuss the effects of errors on observed performance, that is, RT and proportion correct. Next, we fit the drift diffusion model to the data and discuss the effects of errors on the latent psychological processes hypothesized to explain PES.¹

Post–Error Effects on Observed Data

The different hypotheses about PES entail effects on RT, effects on proportion correct, or a combination of the two. It is therefore informative to show—both for post-error trials and post-correct trials, and for different stimulus categories—entire distributions of RT for correct and error responses, together with proportion correct. A convenient tool to paint this multivariate picture is the quantile probability plot (e.g., Ratcliff, 2002). Figure 4.3 shows a quantile probability plot for the data from Keuleers, Brysbaert, and New (2010), based on averaging RT quantiles and proportions across individual participants.

Figure 4.3 features two important factors in the design of this study, that is, post-error trials vs. post-correct trials (i.e., triangles vs. circles) and word frequency of the

¹The analyses reported here concern the difference between post-correct trials and post-error trials. Results based on the difference between pre-error and post-error trials yielded quantitatively and qualitatively similar results. These results can be found on the first author’s website.
Figure 4.3: Post–error and word frequency effects on RT distributions and accuracy: Each colon of dots on the right half of the figure reflects the distribution of correct RTs in a condition and its position on the x–axis defines the accuracy in this condition. Each correct RT distribution on the right half has its incorrect RT counterpart on the left half of the figure at one minus the accuracy on the x–axis. Post–error trials are slower and somewhat more accurate than post–correct trials. This pattern holds for all word stimuli but it is more pronounced for low frequency words (“freq 1”) than for high frequency words (“freq 6”). In addition, the effect is more pronounced in the tail of the RT distribution.

The current stimulus (including nonwords, grey scales). The plot is read as follows. Each column of points summarizes a single RT distribution by five quantiles (i.e., the .1, .3, .5, .7, and .9 quantiles—the .1 quantile, for instance, is the RT value for which 10% of the RT distribution is faster; the .5 quantile is the median RT). Each column in the right half of the figure describes a correct RT distribution for a particular condition; its position on the x–axis shows the corresponding proportion correct (e.g., \(x = 0.61\) for the post–correct, low–frequency “freq 1” words). This correct RT distribution has an associated distribution of incorrect RTs, shown in the left half of the figure (e.g., \(x = 1 – 0.61 = 0.39\) for the post–correct, low–frequency “freq 1” words).

Figure 4.3 shows that word frequency benefits performance: high frequency words are associated with low error rates and fast RT quantiles. More important for the present study, RT quantiles are slower after an error (triangles) than after a correct response (circles). The slowdown is smallest in the leading edge of the distribution (for correct
responses, on average 3 ms at the .1 quantile) and biggest at the tail (on average 38 ms at the .9 quantile). These PES effects are more pronounced for low frequency words (frequency groups one and two) than for high frequency words. Figure 4.4 zooms into the PES effect by presenting the data (and the model fit discussed later) as a delta plot (De Jong, Liang, & Lauber, 1994; Pratte et al., 2010; Speckman, Rouder, Morey, & Pratte, 2008). In a delta plot, the factor of interest—in this case, the PES effect—is shown as a function of response speed. Here, Figure 4.4 shows the average PES effect (i.e., the PES effect across all experimental conditions, quantile–averaged across participants). The delta plot indicates that the PES effect is negligible for the very fast responses and becomes more prominent when response times are slow.

In the following, we will assess differences between conditions by quantifying the evidence in favor of or against the null hypothesis using a default Bayesian t–test (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009; Wetzels et al., in press). The resulting Bayes factor $BF_{10}$ quantifies how much more (or less) likely the data are under the alternative hypothesis than under the null hypothesis. For instance, a $BF_{10}$ of 2 indicates that the data are twice as likely under the alternative hypothesis than under the null hypothesis, whereas a $BF_{10}$ of 1/2 indicates that the data are twice as likely under the null hypothesis than under the alternative hypothesis.

For the low frequency words, accuracy is slightly higher following an error than following a correct response ($BF_{10} = 51.8$). For higher word frequencies, no change in accuracy was present (all Bayes factors $BF_{10} < 1/2.61$). The small decrease in post–error accuracy for nonwords (about 7%) is supported by a Bayes factor of $10^{-0.0}$.

Although the post–error effects in Figure 4.3 are qualitatively consistent across different levels of word frequency, it is not unambiguously clear how these effects should be interpreted in terms of underlying psychological processes. The simultaneous increase in RT and accuracy seems to support an explanation in terms of increased response caution. However, the observed results could also be produced by a combination of increased attention (i.e., drift rate $v$) and a delayed startup of processing (i.e., non–decision time $T_{er}$). And even if one would ignore this alternative interpretation, it is by no means certain that the observed data support a single psychological mechanisms for PES. In order to address this issue and provide a comprehensive account of the data in terms of the underlying, possibly interacting psychological processes that cause PES, we now turn to a diffusion model decomposition.

### Post–Error Effects on Latent Processes

We fit the model to the individual data using the MATLAB package “DMAT” (Vandekerckhove & Tuerlinckx, 2007, 2008) that allows the user to estimate the model parameters using maximum likelihood. As noted above, the size of the present data set allowed us to examine several experimental conditions or factors. The primary factor was the correctness of the previous trial, and secondary factors were stimulus type (i.e., word vs. nonword) on the current trial, word frequency of the word stimuli on the current trial, and stimulus type on the previous trial.

For the secondary factors we used the BIC (Bayesian information criterion; Schwarz, 1978; Kaftery, 1995) to eliminate excess parameters and select the most parsimonious model that still gives an acceptable fit to the data. This BIC–best model was then used to quantify the impact of the primary factor of interest, that is, the factor post–error vs. post–correct was allowed to affect all of the diffusion model parameters.
4. Testing Theories of Post–Error Slowing

Table 4.1: BIC results for the model selected to analyze the PES effect. The middle panel indicates how alternative models were created by restricting the selected model in several ways. The rightmost column shows the BIC difference (averaged across participants), in favor of the selected model against each of the alternative models. Positive BIC values indicate that the selected model is better. For the post–correct trials, the omission of any of the effects results in a clearly worse fit. For the post–error trials, the BIC results suggest that the word frequency effect on \( T_{er} \) and the stimulus type effect on \( \eta \) are not needed. See text for details.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Omitted Effect</th>
<th>Average BIC difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>post–correct</td>
<td>word frequency on ( T_{er} )</td>
<td>197.98</td>
</tr>
<tr>
<td></td>
<td>stimulus type on ( \eta )</td>
<td>182.82</td>
</tr>
<tr>
<td></td>
<td>previous stimulus on ( B &amp; s_B )</td>
<td>227.14</td>
</tr>
<tr>
<td></td>
<td>word frequency on ( v )</td>
<td>1702.75</td>
</tr>
<tr>
<td>post–error</td>
<td>word frequency on ( T_{er} )</td>
<td>-21.76</td>
</tr>
<tr>
<td></td>
<td>stimulus type on ( \eta )</td>
<td>-8.52</td>
</tr>
<tr>
<td></td>
<td>previous stimulus on ( B &amp; s_B )</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>word frequency on ( v )</td>
<td>211.88</td>
</tr>
</tbody>
</table>

In this BIC–best model the different factors affected the model’s parameters as follows: 

*Stimulus type of the current trial* was allowed to affect drift rate \( v \) and its variability \( \eta \), and non–decision time \( T_{er} \). 

*Word frequency* was allowed to affect drift rate \( v \) and \( T_{er} \). 

*Stimulus type of the previous trial* was allowed to affect bias \( B \) and its variability \( s_B \). 

In support of the model selected to analyze the PES effect, Table 4.1 shows the BIC difference (averaged across participants) between the selected model described above and four alternative models. These alternatives implement restrictions on the selected model to test the necessity of including (1) the effect of word frequency on \( T_{er} \); (2) the stimulus type effect on \( \eta \); (3) the previous stimulus’ effect on \( B \) and its variability \( s_B \); and (4), the word frequency effect on \( v \). Table 4.1 shows that for the post–correct trials, the alternative models all perform worse than the selected model. This indicates that the parameters excluded in each of the alternative models are essential to account for the data. For the post–error trials, the BIC recommends two alternative models over the selected model. However, the use of different models for the post–correct and post–error conditions hinders a direct comparison between them, and therefore we opted to analyze the data with the single model outlined above.

Figure 4.4, discussed earlier, compares the data against the model predictions. The solid dots represent the empirical data (i.e., the PES effect in all experimental conditions, quantile–averaged across participants), and the lines with open dots represent the predictions of the best-fitting model parameters. Overall, the fit is good, except perhaps for the .9 quantile; this may be due to the fact that this quantile is the most difficult to estimate reliably.

Figure 4.5 and Figure 4.7 show the estimates for the diffusion model parameters, averaged over participants. The associated Figures 4.6 and 4.8 present the differences in the model parameters for post–correct vs. post–error trials. The most obvious effect in Figure 4.5 and 4.6 is the increase in boundary separation after an error, shown in the upper left panel. This increase in boundary separation indicates that on average, participants become more cautious after committing an error. The Rouder et al. (2009) default Bayesian t–test indicates that the data are about 180, 000 times more likely under
4.5. Results and Discussion

Figure 4.4: Delta plots of PES effect against response speed in post–correct trials, separately for all word frequencies and nonwords. Solid circles represent empirical data (error bars indicate standard errors) and lines with open circles represent predictions from the model. For both empirical data and model predictions, the effects are obtained by quantile-averaging the results across participants.

the alternative hypothesis of unequal boundary separation than under the null hypothesis of equal boundary separation; this is considered extreme evidence in favor of an effect.

The bottom left panels of Figures 4.5 and 4.6 show the post–error effect on bias. After an error, participants shifted their a priori preference toward the “word” response, both when the erroneous response was “word” ($BF_{10} = 11.2$) and when it was “nonword” ($BF_{10} > 8.51 \times 10^7$). In combination with the overall error-induced increase in boundary separation, this means that following an error, people became somewhat more careful to respond “word”, but even more careful to respond “nonword”. The reason for this asymmetry is presently unclear and more empirical work is needed to ascertain whether
The four main parameters of the diffusion model shown separately for post–correct and post–error trials. Bias $B$ was estimated separately for post–word and post–nonword conditions. Drift rate $v$ and non–decision time $T_{er}$ were estimated separately for nonwords and for different categories of word frequency. The most prominent post–error effect is an increase in boundary separation. Error bars represent standard errors of the mean.

The bottom left panel of Figure 4.5 also shows a response repetition effect: participants had a bias towards the response that was executed on the previous trial, regardless of whether that response had been correct ($BF_{10} > 1.35 \times 10^9$ for the comparison of bias between post–word responses and post–nonword responses) or incorrect ($BF_{10} > 2.34 \times 10^3$).

The right two panels show the post–error effects on drift rate and non–decision time. Neither drift rate (for all frequencies, $BF_{10} < 1/3.06$) nor non–decision time (for all frequencies, $BF_{10} < 1/3.00$) were affected by whether or not the response on the previous trial was incorrect. Drift rate did increase with word frequency, indicating that high frequency words were easier to classify than low frequency words (see also [Ratcliff, Gomez, & McKoon, 2004; Wagenmakers et al., 2008]). Non–decision time was also affected by word frequency, indicating that processes such as stimulus encoding and response execution took less time for high frequency words than they did for low frequency words. This finding is conceptually consistent with that of Dutilh, Krypotos, and Wagenmakers. 

58
4.6 Concluding Comments

What does a man do after he makes an error? Data from a 1,094,886–trial lexical decision task showed that people slow down after an error, and a diffusion model decomposition showed that this slowdown can be attributed almost exclusively to an increase in response caution. This result confirms the traditional explanation of PES in terms of self-regulation and cognitive control (e.g., Botvinick et al., 2001; Cohen et al., 2000; Brewer & Smith, 1989; Fitts, 1966; Hajcak et al., 2003; Li et al., 2006; Rabbitt & Rodgers, 1977; press) who found that practice for specific lexical items reduced non–decision time.

Figure 4.6 shows the estimates for the variability parameters of the diffusion model, averaged over participants. The associated Figure 4.7 presents the differences in the model parameters for post–correct vs. post–error trials. The figures suggest that none of the variability parameters are responsible for PES. However, we did find that the variability in drift \( \eta \) was larger for words than it was for nonwords, replicating the result from an earlier lexical decision study (Dutilh et al., 2009).

In sum, the diffusion model decomposition supports an explanation of PES in terms of increased response caution.

Figure 4.6: For each of the four main diffusion model parameters, box plots represent the distribution of the PES effect over participants. Comparison with the dashed horizontal line at zero suggests PES effects express themselves only on boundary separation and a priori bias (after an erroneous nonword response). Boxes contain 50% of the values, whiskers enclose 80% of the values.
4. Testing Theories of Post–Error Slowing

Figure 4.7: The three variability parameters of the diffusion model shown separately for post–correct and post–error trials. All variability parameters are slightly higher for post–error trials than for post–correct trials, but the error bars show that these effects are unreliable. Error bars represent standard errors of the mean.

Figure 4.8: For each of the three variability parameters of the diffusion model, box plots represent the distribution of the PES effect over participants. Comparison with the dashed horizontal line at zero suggests that PES effects on the variability parameters are not reliable. Boxes contain 50% of the values, whiskers enclose 80% of the values.

Shiels & Hawk, 2010; G. A. Smith & Brewer, 1995; Tops & Boksem, 2010; Verguts, Notebaert, Kunde, & Wühr, 2011; Vickers & Lee, 1998): that is, people adaptively change their response thresholds to a possibly nonstationary environment—by becoming more daring after each correct response, and by becoming more cautious after each error, people reach an optimal state of homeostasis that is characterized by fast responses and few errors.

Although this explanation of PES has strong face validity it is entirely possible that other explanations could also be correct in particular cases. Only by applying a formal process model can we evaluate the competing accounts of PES quantitatively. Our results are partially consistent with those of White et al. (2010b), who applied the drift diffusion model to data from a recognition memory task and found that participants with high–trait anxiety responded more carefully after making an error (i.e., increased boundary separation $a$ following an error). However, the data from White et al. (2010b) did not show a response caution effect for participants with low–trait anxiety; in addition, the behavioral data did not show a PES effect, and, moreover, the diffusion model decomposition revealed that for both anxiety groups, errors were followed by an unexpected decrease in non–decision time and a decrease in discriminability (i.e., drift rate differ-
ence between targets and lures). Therefore, we feel the current study presents a more compelling case in favor of the increased response caution explanation of PES.

The present study shows that the drift diffusion model can be used not only to theorize about the causes of PES, but also to decompose the behavioral after–effects of an error into its constituent psychological processes. Such a decomposition is considerably more informative than the standard analysis of mean RT and accuracy, and we believe that future on PES can benefit from taking a similar approach.
A Diffusion Model Account of Age Differences in Post–Error Slowing

Abstract

People generally slow down after they make an error, a phenomenon that is more pronounced for elderly participants than it is for young participants. Here we examine the origin of this age–related difference in post–error slowing (PES) by applying the diffusion model to data from young and elderly participants performing a random dot motion task. Results show that the PES effects on response time and accuracy were qualitatively different for young and elderly participants. A diffusion model analysis revealed that following an error, elderly participants became more cautious, processed information less effectively, and spent more time on irrelevant processes (e.g., overcoming disappointment). For young participants, however, PES was captured predominantly by a modest increase in time wasted on irrelevant processes. These findings indicate that for elderly participants, PES originates from the interplay of three different psychological processes.

5.1 Introduction

In most cognitive tasks, elderly people respond more slowly than young people (Salthouse, 1985). Early research has lead to the pessimistic claim that elderly people slow down because they suffer from a general decrease in the efficiency with which neurons transfer information (Brinley, 1965; Cerella, 1985). Today however, many researchers agree that elderly people slow down, at least in part, because they choose to be more cautious than young people. This means that elderly participants often choose to collect more evidence before they are willing to commit to a decision, a strategy that may result in a substantial
loss of speed in return for a small gain in accuracy (Salthouse, 1979; Starns & Ratcliff, 2010; Strayer, Wickens, & Braune, 1987; Ratcliff et al., 2010, 2006b). This relatively coarse cognitive control is thought to be reflected in more pronounced post–error slowing (PES). Post–error slowing is the phenomenon that participants, after committing an error, tend to slow down on the next trial. The common explanation of PES states that participants constantly monitor their performance and that an error signals the need for more cognitive control in order to keep performance at an acceptable and relatively constant level of accuracy. Specifically, participants are thought to interpret an error as a sign that response criteria need to be increased; this increase ensures that the decision on the next trial is based on more information, reducing the probability of a second consecutive error. Thus, the fact that elderly participants show a relatively pronounced PES effect may be due to an excessive increase in response criteria following an error. This explanation is conceptually consistent with the hypothesis that elderly are more cautious because they are highly motivated to avoid errors. However, other explanations of post–error slowing have been proposed. For instance, PES could also be the result of a distraction of attention (Notebaert et al., 2009) or delayed startup of information processing due to time spent on irrelevant processes (e.g., overcoming disappointment; Rabbitt & Rodgers, 1977).

In this study, we aim to assess the psychological processes responsible for the difference in PES between old and young participants. To do so, we need to overcome two obstacles that complicate the analysis of PES. First, speed and accuracy are in trade–off and only simultaneous analysis of both variables can reveal the psychological processes that underly behavior. Therefore, in this study, we apply the Ratcliff diffusion model (Ratcliff, 1978; Ratcliff & McKeren, 2002) to our data. The diffusion model allows one to take into account response time (RT) and accuracy simultaneously and to decompose the effects on both variables in psychologically relevant underlying constructs. A second challenge is pointed out by Dutilh, Van Ravenzwaaij, and Wagenmakers (Manuscript submitted for publication), who showed that the standard method of quantifying PES (i.e., calculating the difference $RT_{\text{post-error}} - RT_{\text{post-corrected}}$) is confounded with global changes in performance over the course of an experiment; these changes can create spurious PES or mask real PES. therefore, we apply the solution that is described by Dutilh, Van Ravenzwaaij and Wagenmakers (Manuscript submitted for publication) and quantify PES by a local “robust” measure, that is, by the difference in RT prior to and following each individual error.

The outline of this article is as follows. First, we shortly discuss the literature on aging effects in RT tasks and the role of post–error slowing. Second, we introduce the diffusion model. The third section describes how PES analyses can be confounded by global changes in behavior. Next, we present the current study. We present the results both in terms of RT and accuracy and in terms of the diffusion model parameters. While presenting the results, we illustrate the importance of using the method described by Dutilh, Van Ravenzwaaij, and Wagenmakers (Manuscript submitted for publication) to overcome confounds of global changes in performance.

5.2 Response Speed and Age

For many years, the difference in response speed between young and elderly participants has been taken as support for a generalized–slowing hypothesis of aging. This hypothe-
sis states that people slow down with age because all processes in the brain deteriorate (Brinley, 1965; Salthouse, 1996). This unitary view of age–related slowing was popularized by the Brinley plot, a graph that displays, for different experimental conditions, mean RT for both age groups against each other. Brinley plots often result in a straight line, the slope of which is then interpreted as a measure of the relative processing speed of the two groups.

Recently, Ratcliff, Thapar, Spieler, and McKoon have nuanced this unitary view of age differences in RT (e.g., Thapar, Ratcliff, & McKoon, 2003; Ratcliff, Thapar, & McKoon, 2004; Ratcliff, Spieler, & McKoon, 2000; Ratcliff, Thapar, & McKoon, 2006a; Ratcliff et al., 2006b, 2010). In several studies, these researchers used the diffusion model to analyze the effect of aging on performance. As we show below, the diffusion model allows one to decompose effects on RT and accuracy into unobserved psychological processes. Diffusion model decompositions showed that elderly participants were almost always slower than young participants in the non–decisional components of RT (e.g., stimulus encoding and motor time). The diffusion model decompositions also showed that in many tasks, elderly participants responded more cautiously than young participants (see also Strayer et al., 1987). Importantly, only in a subset of tasks (i.e., perceptual decision making tasks) did the diffusion model decompositions suggest that elderly participants were less effective than younger participants in extracting information from the stimulus. These findings suggest that a general slowing hypothesis is problematic. Instead, detailed diffusion model analyses suggest that age–related slowing is be a multifaceted phenomenon that is dominated by processes unrelated to the effectiveness with which information is extracted and accumulated.1

Elderly participants not only respond more slowly than young participants on average, but they also exhibit a more pronounced PES effect (Rabbitt, 1979; Rabbitt & Vyas, 1980; G. A. Smith & Brewer, 1995; Band & Kok, 2000). It is tempting to conclude that, following an error, elderly participants increase their response thresholds more than young participants (G. A. Smith & Brewer, 1995). However, there are several alternative explanations of how an error influences performance on the next trial, such as: (1) the error distracts attention away from the task (Notebaert et al., 2004); (2) the error shifts people’s a priori bias away from the response that was executed in error (Laming, 1968; Rabbitt & Rodgers, 1977); (3) the error delays the startup of information accumulation because time is wasted on irrelevant processes such as overcoming disappointment (Rabbitt & Rodgers, 1977); or (4) the error causes people to time the onset of the stimulus more precisely, preventing the sampling of irrelevant information (Laming, 1968, 1979a).

Dutilh, Vandekerckhove, Forstmann, and Wagenmakers (in press) have shown that the diffusion model is able to discriminate between the alternative explanations of PES listed above (see also White et al., 2010c). In this study, we use the diffusion model to assess whether the PES effect in young and elderly participants originates from changes in the same or different psychological processes.

### 5.3 Accounting for Speed and Accuracy Simultaneously

There is a trade–off relation between RT and accuracy, and this trade–off compromises the analysis of task performance when RT and accuracy are considered in isolation (e.g.,

---

1In related work, Ratcliff and colleagues have shown that the diffusion model can generate linear Brinley plots when either drift rate or boundary separation differs between elderly and young participants (e.g., Ratcliff et al., 2004, Figure 8; see also Van Ravenzwaaij, Brown, & Wagenmakers, 2011 but see Myerson, Adams, Hale, & Jenkins, 2003.)
In particular, it can be difficult to assess performance differences between young and elderly participants when the latter are known to respond more cautiously, accepting a large loss in RT for small gains in accuracy. In order to quantify performance differences in the face of the speed-accuracy trade-off, one requires a formal model of information processing. In this study, we apply the diffusion model. The diffusion model decomposes effects on RT and accuracy into effects on underlying psychologically relevant processes. The diffusion model naturally accounts for the speed-accuracy trade-off, a key feature for the analysis of age-related slowing.

The Diffusion Model

In the diffusion model for speeded two-choice tasks (Ratcliff, 1978), stimulus processing is modeled as the noisy accumulation of evidence over time. When the accumulated evidence reaches a predefined evidence boundary, a response is initiated (Figure 5.1). The four main components of the diffusion model are (1) the speed of information processing, quantified by drift rate \( v \). The value of drift rate depends on the difficulty of the stimulus and the ability of the participant; (2) response caution, quantified by boundary separation \( a \); (3) a priori bias, quantified by starting point \( z \); and (4) nondecision time, quantified by \( T_{er} \). The model assumes that at stimulus onset, a participant starts sampling information with a signal-to-noise ratio governed by drift rate \( v \). Values of \( v \) near zero produce long RTs and high error rates. Boundary separation \( a \) determines the speed-accuracy trade-off; lowering \( a \) leads to shorter RTs at the cost of a higher error rate. Together, these parameters generate a distribution of decision times (\( DT \)). The observed RT, however, also consists of stimulus-nonspecific components such as response preparation and motor execution, which together comprise non-decision time \( T_{er} \). The model assumes that \( T_{er} \) simply shifts the distribution of \( DT \), such that \( RT = DT + T_{er} \) (Luce, 1986). The model specification is completed by including parameters that specify across-trial variabilities in drift rate, starting point, and non-decision time (Ratcliff & Tuerlinckx, 2002). These variability allow the model to account for empirical phenomena such as situations in which errors are systematically faster or slower than correct responses. For a more elaborate introduction to the diffusion model and a review of applications, see Ratcliff and McKoon (2008) and Wagenmakers (2009).

The preferred method to analyze RT data is to use the full diffusion model as described above. However, to fit the diffusion model, one needs at least enough data to estimate the distribution of correct RTs and, more problematic, a distribution of at least, say, 10 error RTs. Because the crucial condition in this study consists of trials that follow an error, we would need at least 10 errors that immediately follow errors. For this reason, it is not possible to fit the full diffusion to data of individual participants in this study.

Fortunately, we can still apply the EZ-diffusion model (Wagenmakers et al., 2007) to the data. A crucial advantage of this simplified version of the diffusion model is that one can calculate the model parameters in closed form from the percentage correct, the RT variance, and the RT mean. In contrast to the full model, this model uses only the parameters drift rate \( v \), boundary separation \( a \) and non-decision time \( T_{er} \). Thus, the model lacks the variability parameters and bias parameter \( z \). Nonetheless, Ravenzwaaij and Oberauer (2009) have show that the EZ-diffusion model is capable of capturing experimental effects in its parameters and that the model is especially useful when studying individual participants. For other applications of the EZ-diffusion model see for example Schmiedek et al. (2007), Schmiedek, Lövdén, and Lindenberger (2009), Kamienkowski, Pashler, Dehaene, and Sigman (2011), and van Ravenzwaaij, Dutilh, and Wagenmakers (2012).
5.4 Quantification of Post–Error Slowing

The EZ–diffusion model allows us to test three different explanations of PES for this data set: (1) an error distracts attention from the stimulus (PES effect in $v$); (2) an error leads to an increase in response caution (PES effect in $a$); and (3) an error delays the startup of information processing due to time wasted on irrelevant processes (PES effect in $T_{er}$). Note that these explanations are not mutually exclusive, and behavior can be effected by more than one of these processes simultaneously.

5.4 Quantification of Post–Error Slowing

The most obvious and popular way to quantify post–error slowing is to compare trials that follow errors to trials that follow correct responses. However, Dutilh et al. (2011, see also Laming, 1979b, p. 205) showed that this traditional analysis of PES is vulnerable to a confound of global changes in behavior over the course of an experiment. The confound can be best understood by considering the following hypothetical scenario. A participant starts an experimental session highly motivated. The participant is very focused and responds quickly and accurately. Over the course of the experimental session, however, fatigue starts to kick and consequently responses become slower and less accurate. Now, consider that this participant does not slow down after errors, that is, there is no true PES effect whatsoever. When calculating the difference between RT on post–error trials versus post–correct trials, most post–error observations will originate from the last part of the session when the participant was least motivated and most errors occurred. In this part of the session, the participant responded slowly overall, resulting in a relatively slow mean RT for post–error trials. In contrast, most post–correct trials originate from

Figure 5.1: Graphical illustration of the diffusion model. A “Right” stimulus is presented (in this study: random dots motion to the right, see method section). The two example sample paths represent the accumulation of evidence which result in one correct response (“Right”, light line) and one error response (“Left”, dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper boundary is reached more often than the incorrect, lower boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a non–decision component that represents the time needed for processes such as stimulus encoding and response execution.
the first part of the session, where accuracy was high and RT low, yielding a relatively low mean RT on post–correct trials. Thus, the comparison of post–error vs. post–correct trials will indicate a PES effect, although real post–error slowing is absent. Dutilh, Van Ravenzwaaij, and Wagenmakers (Manuscript submitted for publication) show that, in addition to spuriously detecting post–error slowing, the traditional analysis can also underestimate or mask real PES.

The solution offered by Dutilh, Van Ravenzwaaij, and Wagenmakers (Manuscript submitted for publication) is simple: compare post–error trials to those post-correct trials that are also pre–error trials. This extra condition assures that trials that comprise both sides of the comparison originate from the same locations in the data set: surrounding errors. For convenience we will refer to the conditions in this robust comparison as post–error vs. pre–error.

5.5 Method

Participants

We tested 15 young people (student age; mean=21.80, SD=2.27) and 19 elderly people (60–80 years old; mean = 68.27, SD=5.89) on two occasions within a time span of two weeks. Elderly participants were recruited from the Seniorlab database.2 The young participants were rewarded by course credits. The elderly people participated for a monetary reward.

Individualized Deadlines

In the experiment we alternated blocks that required participants to respond accurately with blocks that required participants to respond quickly. In the accuracy blocks, participants were told to respond accurately but not waste time. In the speed blocks, participants were asked to respond quickly and risk committing errors. To encourage fast responding in the speed blocks, we used a deadline. When participants took longer than the deadline to respond they received the warning message “too slow” on the computer screen.

As mentioned above, recent studies suggest that elderly participants have longer non–decision times than younger participants (Thapar et al., 2003; Ratcliff et al., 2006b). Therefore, it is unreasonable to confront participants of both ages with the same deadline. In the diffusion model, non–decision time is related to total RT as follows:

\[ \text{RT} = \text{decision time} + T_{er} \]

This equation shows that with the same deadline on RT, older participants—who have relatively long non–decision times \( T_{er} \)—have available less decision time than young participants. Because we wanted young and old participants to experience a comparable level of speed stress, the deadline was adjusted for each participant individually and computed as follows:

\[ RT_{\text{deadline}}^i = \text{decision time} + T_{er}^i \]

, in which \( RT_{\text{deadline}}^i \) is the individualized deadline for participant \( i \). This personal deadline was computed by adding participant \( i \)'s personal estimate of \( T_{er} \) to a fixed decision

---

2Seniorlab is an organization that mediates between elderly people and researchers. http://www.seniorlab.nl.
5.5. Method

time that was the same for all participants. Each participant’s personal estimate of \( T_{er} \)
was calculated at the start of the experiment. This procedure permits similar decision
times for participants with varying non–decision times.

Materials

Participants completed a random dots motion task (Britten et al. 1992). The random
dots motion stimulus consists of a circular display of dots. The dots appear, disappear,
and are replaced in such a way that the entire circle of dots appears to move. The apparent
motion that the participant perceives can be best described as the flickering of a turning
disco ball in a spotlight. This illusion is created as follows. At each frame (50 ms), 120 dots
are displayed. Every next frame, an experimentally defined proportion \( P_{move} \) of the dots
from the former frame are shifted a certain distance \( l_{move} \) to the target side (e.g., right, if
the correct response is right). The remaining portion of the pixels is randomly replaced
in the circle (independent of their previous positions). \( P_{move} \) was set to 50\%. \( l_{move} \) was
set to one pixel. We choose a dot size that is bigger than normal (diameter of 4 pixels),
to account for the fact that some of the elderly might have small visual impairments. The
circular aperture had a 13 cm diameter. Presentation (Version 09.24.07) for Windows
was used to present the stimuli and register the responses. Participants responded to the
stimuli by pressing one of two response buttons to indicate the direction of the apparent
movement (left or right).

Procedure

In two sessions within two weeks, participants were tested individually in a room in which
the experimenter was present during the entire experiment. For each participant, the two
sessions took one hour each.

The first session started with two 50–trial training blocks. One training block had speed
instructions. In this block, a “too slow” feedback was displayed when a deadline of
650 ms was passed. The other training block had accuracy instructions and an “incorrect”
feedback message was given on incorrect responses. Also, after 1500 ms, a “no response”
warning was shown.

After participants were trained on the stimuli and the deadline in the speed blocks,
another 300–trial speed block was administered. The data of the last 250 trials of this
block were used for the personal estimate of \( T_{er} \). This personal estimate of \( T_{er} \) was then
used calculate the personal deadline, as described above. Based on a pilot study with the
same task, we choose \( RT_{decision} = 200ms \), since this value plus an average \( T_{er} \) yielded a
reasonable deadline. The personal deadline that was determined with this procedure was
used in the speed blocks throughout both sessions of the experiment.

After the practice blocks and the block to calculate the non–decision time estimate, the
main experiment started. Blocks with speed instructions were alternated with blocks with
accuracy instructions. Each block contained 100 trials. After each block, participants
had a self–paced break. Testing was limited to two hours and hence the amount of
blocks administered depended on the participant’s pace. Young participants completed
on average 15.6 speed blocks and 15.4 accuracy blocks. Elderly participants completed
on average 13.0 speed blocks and 13.5 accuracy blocks.
5.6 Results

Deadlines

As expected, the non–decision times estimated at the beginning of the experimental session were somewhat higher for elderly (368 ms, SD=42) than for young participants (350 ms, SD=28). Although this difference is modest on the group level, individualized non–decision time estimates ranged from 265 ms to 435 ms. This 130 ms range in non–decision time estimates suggests that the deadline adjustment procedure is useful to equate speed pressure over participants.

Below, we will first report the descriptive results on RT and accuracy and then turn to the analyses with the EZ–diffusion model.

Descriptive Results

The current data set illustrates the potential impact of analyzing PES either with the traditional method or with the robust method. In this section we first report the results of both analyses, beginning with the robust analysis.

Post–error effect on RT and accuracy: Robust analysis

Figure 5.2 shows the results of the robust analysis, based on a comparison between post–error and pre-error trials. The left panels show results for young participants, right panels show results for elderly participants. The upper panels show the post–error versus pre–error difference for each RT quantile, plotted against the RT quantiles in pre–error trials. This delta plot ([De Jong et al., 1994; Speckman et al., 2008; Pratte et al., 2010]) shows PES–induced changes across the entire RT distribution.

In both the speed and the accuracy condition, young participants responded between 20 and 40 ms slower after an error, a slowdown that is roughly constant across the RT distribution. For elderly participants, in contrast, the PES effect increased across the RT distribution: in the fast, 10% quantile elderly participants were about 30 ms slower after an error, but this difference increased to around 90 ms in the slow, 90% quantile. Thus, PES expressed itself differently for young and elderly participants; for young participants, the RT distribution is shifted by a constant amount, whereas for elderly participants errors shift and skew the RT distribution. This qualitative difference between young and elderly participants is only evident from the distribution of RT and would have gone unnoticed had we only considered mean RT.

In both the speed and the accuracy condition, for both age groups, there appears to be no PES effect on accuracy. Clearly, however, accuracy for both groups is higher in the accuracy condition than in the speed condition.

A comparison of overall performance for elderly participants and young participants shows that elderly participants are less accurate than young participants overall. Also, for both the speed and the accuracy condition, the pre–error RT distribution for elderly participants is shifted and stretched out relative to that for the young participants. This indicates that elderly are slower than young participants over the entire distribution, a difference that increases in the tail of the distribution.

Post–error effect on RT and accuracy: Traditional analysis

Figure 5.3 shows results based on the traditional analysis of PES, that is, where the y–axis represents the traditional quantile RT difference between post–error and post–correct...
Figure 5.2: Robust analysis of post–error slowing. The figure shows how RT and accuracy differ between post–error and pre–error trials for both age groups, separately for speed and accuracy blocks. All RT quantiles are slower post–error than pre–error; for elderly participants, this difference increases over the 5 RT quantiles (at 10, 30, 50, 70, and 90%). The quantiles on the x–axis and differences on the y–axis were calculated per participant before averaging. Error bars contain two standard errors around across–participant means. Lower panels show that there is no post–error effect on accuracy.

trials (without the correction discussed above). At first glance, the figure looks similar to Figure 5.2. However, in Figure 5.3 the PES effect is smaller in both conditions and for both age groups. Also, Figure 5.3 suggests that elderly participants have a PES effect on accuracy. Because we deem these results an artifact of the traditional analysis procedure, we will not discuss them further here.

Now that we have illustrated the importance of contrasting the right conditions, we turn to the diffusion model analysis of the effects found with the robust PES method.

Modeling Results

Figure 5.2 shows that the PES effect is qualitatively different for young and elderly participants, suggesting that different psychological processes underly the effect for both groups. To quantify this intuition, we applied the EZ–diffusion model and plotted the model parameters in Figure 5.4.

In order to gauge the difference between parameter values shown in Figure 5.4, we quantify the evidence in favor of or against the null hypothesis using a default Bayesian t–test (Rouder et al. 2009; Wetzels et al. 2009, in press). The resulting Bayes factor $BF_{10}$
quantifies how much more (or less) likely the data are under the alternative hypothesis than under the null hypothesis. For instance, a $BF_{10}$ of 2 indicates that the data are twice as likely under the alternative than under the null hypothesis. An important advantage of calculating the Bayes factor is that it allows us to quantify evidence in favor of the null as well as in favor of the alternative hypothesis. Analyses were performed on data collapsed over irrelevant conditions.

The two leftmost panels of Figure 5.4 show that drift rate is higher for young participants than for elderly participants ($BF_{10} = 104.5$), indicating that the efficiency of information processing was higher for young than for elderly participants. The two center panels of Figure 5.4 show that boundary separation is higher for elderly participants than for young participants ($BF_{10} = 1207.0$), indicating that elderly participants were more cautious than young participants. The two rightmost panels of Figure 5.4 show that non–decision time is higher for elderly participants than for young participants ($BF_{10} = 1631.4$), indicating that elderly participants needed more time to encode stimuli and execute the motor response than young participants.
5.6. Results

Inspection of the two center panels of Figure 5.4 also confirms that boundary separation was higher in the accuracy blocks than in the speed blocks for both young ($BF_{10} = 350.1$) and elderly ($BF_{10} = 358.7$) participants.

Note that the error bars in Figure 5.4 were computed separately for each condition. A more accurate assessment of the PES effects on the model parameters can be obtained by plotting these effects within–subjects, as is done in Figure 5.5.

Figure 5.4: The post–error effect on the three parameters of the EZ–diffusion model separately for the two age groups and the two experimental conditions. Error bars enclose two standard errors. See text for details.

Figure 5.5 shows, for both young and elderly participants, the difference in the parameter estimates between pre–error and post–error trials. The post–error effects on neither of the parameters differed between speed and accuracy conditions (all Bayes factors $BF_{10} < .5$); therefore, we collapsed the results across the speed and accuracy conditions before plotting the figures and conducting the analyses.

The advantage of Figure 5.5 over Figure 5.4 is that it allows an assessment of the distribution of post–error effects over participants. Inspecting the leftmost box plots in each panel from Figure 5.5, the data show that young participants have a non–decision time that is systematically higher after an error, $BF_{10} = 143.5$. Neither drift rate ($BF_{10} = 0.16$), nor boundary separation ($BF_{10} = 0.28$) differed before and after errors.

The rightmost box plots in each panel from Figure 5.5 show the results for elderly participants. Here, three effects are present. First, elderly participants have lower drift rates after errors ($BF_{10} = 637.8$), indicating a decrease in the speed of information processing. Second, boundary separation increases after errors ($BF_{10} = 13.7$), indicating an increase in response caution. Finally, as for young participants, non–decision times increase after errors ($BF_{10} = 28.0$), indicating that elderly participants might waste some
time on irrelevant processes after committing an error.

The above results are statistically compelling: the Bayes factors clearly indicate that, for elderly participants, post–error effects are reflected in drift rate, boundary separation, and non–decision time; for young participants, the Bayes factors clearly indicate that post–error effects are reflected in non–decision time. Compared to p values, Bayes factors are more conservative and less eager to support the alternative hypothesis (e.g., Berger & Sellke, 1987; Wetzels et al., in press). For example, data for which the p value equals .05 yield at most a Bayes factor of 2.46, and data for which the p value equals .01 yield at most a Bayes factor of 7.99 (Sellke, Bayarri, & Berger, 2001). In addition, the default Bayesian t test is relatively conservative. Taken together, this means that the evidence for the above effects is relatively strong.

Moreover, our conclusions are conceptually consistent with the observed data: following an error, the RT distributions for elderly participants skew out, and this is accommodated by a decrease in drift rate. At the same time, boundary separation has to increase in order to keep accuracy constant.

A note of caution: our statistical analyses do not warrant the claim that, for young participants, post–error effects are reflected only in non–decision time. This claim requires that the Bayes factor supports the null hypothesis of no error–induced effects on drift rate and boundary separation. For the present data set, the evidence in favor of both null hypotheses is suggestive but not conclusive. This ambiguity is also apparent from the fact that the error–induced effects on drift rate and boundary separation are not statistically different between young and elderly participants (i.e., $BF_{10} = 1.24$ and $BF_{10} = 0.46$, respectively). Hence, it would be premature to conclude that the PES effects on the model parameters are different between the young and the old participants (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Nevertheless, the data

---

3 These Bayes factors are upper bounds obtained by considering the data and choosing the parameter prior that provides maximum support for the alternative hypothesis. See Sellke et al. (2001) for details.
do warrant the conclusion that for elderly participants, PES manifests itself in drift rate, boundary separation, and non-decision time, whereas for young participants, PES manifests itself in non-decision time.

**Model Fit**

Figure 5.6 shows the empirical quantiles for all experimental conditions plotted against the diffusion model predictions based on the parameter values displayed in Figure 5.4. The fit is very satisfactory in general, although in the accuracy blocks for elderly participant there is a slight misfit in the highest (90%) quantile for both pre-error and post-error trials.

![Model Fit](image)

Figure 5.6: Model fit. Empirical RT quantiles (10, 30, 50, 70, and 90%) are plotted against RT quantiles predicted by the EZ-diffusion model.

### 5.7 Discussion

This study focused on performance differences between young and elderly participants in a perceptual decision making task. Our results confirmed that elderly participants are more cautious than young participants. Also, our results confirmed that non-decision times are larger for elderly participants and that, in perceptual tasks, the speed of information processing tends to be lower for elderly participants (Ratcliff et al., 2006a, 2006b). More
specifically however, the goal of this study was to study age–related differences in post–
error performance. Our PES effects are consistent with those reported by G. A. Smith
and Brewer (1995): elderly participants slow down more strongly after errors than young
participants. A delta plot of the data revealed that the PES effect for young participants
was qualitatively different from that for elderly participants; for the young, an error only
shifts the RT distribution, whereas for the elderly, an error shifts and skews the RT
distribution.

To obtain a more detailed understanding of the psychological processes that drive the
post–error effects, we applied the diffusion model to the observed data. The diffusion
model decomposition showed that for both young and elderly participants, non–decision
time increases after an error. This change in the shift of the RT distribution can have
several reasons, but two hypotheses seem particularly plausible in this study. First, it
is possible that participants needed time to evaluate the error, to overcome disappoint-
ment, or to internally berate themselves. Second, it is possible that on the trial following
an error participants pressed the response button less firmly because of error–induced
hesitation or lack of confidence. The diffusion model cannot distinguish between these
hypotheses because the non–decision time encompasses both the time prior to the onset
of the information accumulation process and the time after the information accumulation
process has terminated (see Figure 5.1). Nevertheless, future work could test the hesita-
tion hypothesis empirically by using a membrane key or force button – if the hesitation
hypothesis is true, a more sensitive measurement of response execution may reduce or
eliminate the post–error effect on non–decision time.

The diffusion model decomposition also revealed that, for the elderly participants,
the occurrence of an error reduced drift rate. A plausible explanation for this reduction
in the efficiency of information processing is that the evaluation of the error message is
still ongoing when the next stimulus has to be processed; as a result, fewer cognitive
resources are available for the primary task. Finally, the diffusion model decomposition
showed that elderly participants tended to become more cautious after committing an
error. This finding is in line with the common explanation of post–error slowing as a
strategic adjustment of response thresholds and it is consistent with conclusions drawn
by G. A. Smith and Brewer (1995). By increasing response thresholds, participants are
able to increase the probability of a correct response at the cost of a decrease in response
speed.

Altogether, the results from this perceptual task challenge the common explanation
of PES in terms of an adjustment in boundary separation. For elderly participants,
PES appears to be a multifaceted phenomenon that is associated with changes in three
psychological processes; for young participants, PES is reflected by an increase in non–
decision time. However, it may well be that the nature of PES is task–specific. For
instance, Dutilh, Vandekerckhove, et al. (in press) applied the diffusion model to assess
and decompose PES in a large lexical decision experiment with college–aged participants.
Dutilh, Vandekerckhove, et al. (in press) concluded that PES was largely due to an
increase in response caution after errors. The difference between the results of the present
study and those reported in Dutilh, Vandekerckhove, et al. (in press) may be due to the
details of the task. In the lexical decision task, participants are usually aware of their
errors; moreover, participants may realize that if they had responded more carefully they
could have avoided most errors. In the moving dots task, in contrast, participants may
only have limited error awareness; moreover, they may not feel that responding more
carefully reduces the probability of making an error.

The conclusions in this study were made possible by a combination of two rela-
tively novel methods in the analysis of PES. First, the application of a robust measure
for PES (i.e., $RT_{\text{post-error}} - RT_{\text{pre-error}}$) – this measure is not confounded by global changes in motivation or response caution in the way that the traditional measure (i.e., $RT_{\text{post-error}} - RT_{\text{post-correct}}$) is. Second, the application of the diffusion model allowed us to study not just changes in RT and accuracy, but draw conclusions based on the underlying psychological processes that change following an error.
Chapter 6

How to Measure Post–Error Slowing: A Confound and a Simple Solution

This chapter has been submitted for publication as:
Dutilh, G., Ravenzwaaij, D., Nieuwenhuis, S.,
van der Maas, H.L.J., Forstmann, B.U., and Wagenmakers, E.–J.
How to Measure Post–Error Slowing: A Confound and a Simple Solution

Abstract

In many response time tasks, people slow down after they make an error. This phenomenon of post–error slowing (PES) is thought to reflect an increase in response caution, that is, a heightening of response thresholds in order to increase the probability of a correct response at the expense of response speed. In many empirical studies, PES is quantified as the difference in response time (RT) between post–error trials and post–correct trials. Here we demonstrate that this standard measurement method is prone to contamination by global fluctuations in performance over the course of an experiment. Diffusion model simulations show how global fluctuations in performance can cause either spurious detection of PES or masking of PES. Both biases are highly undesirable and can be eliminated by a simple solution: quantify PES as the difference in RT between post–error trials and the associated pre–error trials. Experimental data are used as an empirical illustration.

6.1 Introduction

People tend to slow down after they commit an error, a phenomenon known as post–error slowing (PES). Ever since the classic article “What does a man do after he makes an error?” (Rabbitt, 1966), the PES phenomenon has received considerable attention in the response time (RT) literature and several explanations have been proposed to explain its existence (e.g., Rabbitt & Rodgers, 1977; Laming, 1968, 1979b; Notebaert et al., 2003; see Dutilh, Vandekerckhove, et al., in press, for an empirical comparison). The most popular account of PES states that it reflects an error–induced increase in response
caution that allows a participant to maintain a relatively constant level of accuracy (e.g., Botvinick et al., 2001; G. A. Smith & Brewer, 1995).

Specifically, this account holds that participants continually monitor their performance and interpret errors as a sign that the chosen response threshold was too liberal. Consequently, participants heighten their threshold following an error in order to increase the probability of a correct response on the next trial. The heightened threshold leads to fewer errors but also causes slower responding (i.e., the PES phenomenon).

At the same time, participants interpret correct responses as a sign that the chosen response threshold was too conservative, and therefore they are assumed to lower their threshold following each correct response. Thus, participants become more cautious after an error and slightly more daring after a correct response; in this way the system self-regulates to a state of homeostasis characterized by fast responses and few errors. Figure 6.1, based on fictive but representative data, illustrates the typical pattern of modest post–correct speeding and pronounced post–error slowing (e.g., Brewer & Smith, 1989; G. A. Smith & Brewer, 1995).

![Figure 6.1: Typical fluctuations in mean RT surrounding an error E. Fictive data are representative of results reported in Smith and Brewer (1995) and Laming (1979). Participants tend to speed up until they commit an error, after which they slow down considerably. Subsequent to the post–error trial E + 1, participants start to speed up again.](image)

This response–monitoring interpretation of PES suggests that the amplitude of PES can be used as a direct measure of cognitive control. Although the response monitoring/cognitive control interpretation might not be appropriate in all cases (e.g., Notebaert et al., 2003; Dutilh, Forstmann, Vandekerckhove, & Wagenmakers, Manuscript submitted for publication), in many studies it is assumed to be correct from the outset. Consequently, the magnitude of PES is often treated as an important dependent variable that is correlated with neurophysiological variables such as anterior cingulate activity (Li et al., 2006; Danielmeier et al., 2011), error–related negativity (ERN) and positivity (Pe; Hajcak et al., 2003), and cortisol levels (e.g., Tops & Boksem, 2010).

In this article we discuss how PES can best be measured. First we explain how, although straightforward and intuitive, the traditional method to quantify PES can create coarseness of the fluctuations of RT around errors as a negative indicator of cognitive control. These authors argued that elderly have coarser control over their speed accuracy trade–off, indicated by larger fluctuations in RT surrounding an error.

---

1Note that Rabbit (1979), Brewer and Smith (1989), and G. A. Smith and Brewer (1995) interpret the coarseness of the fluctuations of RT around errors as a negative indicator of cognitive control. These authors argued that elderly have coarser control over their speed accuracy trade–off, indicated by larger fluctuations in RT surrounding an error.
6.2 The Measurement of Post–Error Slowing

There are several methods to quantify PES. The most insightful method plots the fluctuations in mean RT surrounding an error (e.g., Brewer & Smith, 1989; G. A. Smith & Brewer, 1995; see Figure 6.1 for an example). The resulting graph shows mean RT for error trial $E$, mean RT for subsequent trials $E+1$, $E+2$, etc., and mean RT for preceding trials $E-1$, $E-2$, etc. The form of the graph depends slightly on what trials are included in the calculations. For example, one may choose to include pre–error trials that are also post–error trials, one may include errors that are simultaneously pre–error or post–error trials, and so forth. Although the graphical method is very informative, researchers often prefer a method that quantifies the magnitude of PES in a single number.

The traditional and most intuitive method to quantify PES in a single number is to calculate the difference in mean RT (MRT) between trials post–error and trials post–correct. This difference, $PES_{traditional} = MRT_{post-error} - MRT_{post-correct}$, is often calculated per condition per participant and is used as a behavioral variable for further analysis. The magnitude of $PES_{traditional}$ may depend slightly on whether or not error trials are included in the calculation of $MRT_{post-error}$ and $MRT_{post-correct}$ (e.g., Hajcak & Simons, 2008).

6.3 A Confound

The traditional method of quantifying PES, $PES_{traditional} = MRT_{post-error} - MRT_{post-correct}$, has strong face validity. However, the method is vulnerable to a confound that was already hinted at by Laming (1979b, p. 205) when he suggested . . . the possibility that errors and the increased RT on trials which follow them are jointly due to a local deterioration in performance. Suppose, for example, that the subject suffers short periods of relative inattention to the CR [choice response] task . . . During these periods RTs are longer and errors more frequent than normal.

Such a local deterioration of performance leads to two possible complications when PES is calculated using the traditional method. The first complication is that global changes in motivation may lead to spurious post–error slowing. The second complication is that global changes in response caution may lead to spurious post–error speeding. Both situations are illustrated in Figure 6.2.

First, consider the hypothetical scenario where a participant starts a one–hour, 1000–trial experimental session with high motivation. As a result, the participant’s responses are fast and accurate. However, as the session proceeds, fatigue starts to kick in and motivation drops. This decrease in motivation is illustrated in the upper left panel of Figure 6.2. With low motivation, the participant’s responses become increasingly slow.

\footnotetext[2]{A related method compares (correct) post–error trials to all correct trials (both post–error and post–correct). This method is also vulnerable to the confound of global fluctuations in performance that we discuss in this paper.}
Figure 6.2: Schematic representation of two confounds. The three panels in the left column illustrate that, when a participant’s motivation decreases during an experimental session, slow RTs co–occur with low accuracy. In this case, calculation of $\overline{PES}_{\text{traditional}}$ may result in spurious or inflated estimates of PES. The three panels in the right column illustrate that, when a participant’s caution decreases during an experimental session, slow RTs co–occur with high accuracy. In this case, calculation of $\overline{PES}_{\text{traditional}}$ may result in spurious post–error speeded, or deflated estimates of PES.

(middle left panel) and inaccurate (bottom left panel). Now suppose that this participant does not slow down after errors, that is, real PES is completely absent. Now, we quantify PES in this participant’s data with the standard method $\overline{PES}_{\text{traditional}} = MRT_{\text{post–error}} - MRT_{\text{post–correct}}$. Notice that most post–error RTs will originate from the second half of the session, because there are more errors in that half. Likewise, most post–correct RTs will originate from the first half of the session, because there are more correct responses in that half. Because of the decrease in motivation, responses in the first half of the experiment are quicker than responses in the second half of the experiment. Therefore, post–correct trials will on average be faster than post–error trials, despite the fact that there is no real PES. Thus, the traditional comparison of post–error RTs with post-correct RTs can yield an artificial PES effect (or even mask post–error speeding).

Second, consider the hypothetical scenario where a participant starts an experimental session very keen on being accurate. The participant’s responses are then highly accurate but slow. As the session continues, the participant may get increasingly bored
6.4 Simulation Studies

and careless. The associated decrease in response caution is illustrated in the upper right panel of Figure 6.2. With decreasing caution, the participant’s responses become quicker (middle right panel) at the cost of accuracy (bottom right panel). Analogous to the previous case, suppose that for this participant, real PES is completely absent. Once again we analyze PES in this participant’s data set in the standard fashion, that is, $\overline{PES}_{\text{traditional}} = MRT_{\text{post-error}} - MRT_{\text{post-correct}}$. Note that most error RTs and therefore most post–error RTs will again originate from the second half of the experiment, where the participant was rather careless. As before, most post–correct RTs originate from the first half of the experiment, where the participant was relatively careful. In contrast to the first scenario, however, responses are faster in the second half of the experiment than in the first half. Therefore, post–correct trials are on average slower than post–error trials, despite the fact that there is no real PES. Thus, the traditional comparison of post–error RTs with post–correct RTs can yield an artificial post–error speeding effect (or mask PES when it is present).

The two scenarios described above show that global changes in performance may result in a systematic bias in the estimation of PES, at least when it is quantified as $\overline{PES}_{\text{traditional}}$. Below we present two simulation studies that support this conclusion by quantifying the impact of the two confounds using a process model of RT.

6.4 Simulation Studies

In two simulation studies we used the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Wagenmakers, 2009) to make the scenarios described above more concrete. The diffusion model produces both RTs and percentage correct. Most importantly for this study, the model can describe the specific influences of motivation and response caution on response time data.

The Diffusion Model

In the diffusion model for speeded two–choice tasks (Ratcliff, 1978), stimulus processing is modeled as the noisy accumulation of evidence over time. A response is initiated when the accumulated evidence reaches a predefined boundary (Figure 6.3). The main components of the diffusion model are (1) speed of information processing, quantified by drift rate $v$. Low absolute values of $v$ produce relatively long RTs and high error rates; (2) response caution, quantified by boundary separation $a$. Low values of $a$ lead to relatively short RTs and high error rates; and (3) a priori bias, quantified by starting point $z$. Together, these parameters generate a distribution of decision times $DT$. The observed RT, however, also consists of stimulus–nonspecific components such as response preparation and motor execution, which together make up non–decision time $T_{er}$. The model assumes that $T_{er}$ simply shifts the distribution of $DT$, such that $RT = DT + T_{er}$ (Luce, 1986). The model specification is completed by including parameters that specify across–trial variabilities in drift rate, starting point, and non–decision time (Ratcliff & Tuerlinckx, 2002). These variability parameters allow the model to account for empirical phenomena such as the finding that errors can be systematically faster or systematically slower than correct responses.

After the diffusion model has been fit to data, the parameters can be interpreted in terms of psychological processes (e.g., ability, caution, bias), allowing a decomposition of performance into its constituent cognitive elements. The diffusion model has been applied to a range of different tasks such as lexical decision (e.g., Dutilh et al., 2009) and
perceptual discrimination (e.g., Ratcliff & Rouder, 1998) with the goal to increase our knowledge of aging (e.g., Ratcliff et al., 2006b, 2010), practice (e.g., Dutilh et al., 2009; Dutilh, Krypotos, & Wagenmakers, in press), and clinical disorders (e.g., White et al., 2009, 2010a, 2010b).

![Graphical illustration of the diffusion model](image)

Figure 6.3: Graphical illustration of the diffusion model. Consider a trial in which a “right” stimulus is presented (in the case of the empirical example we describe below, this would correspond to a random dot motion stimulus moving to the right). The two example sample paths represent the accumulation of evidence which result in one correct response (“right”, light line) and one error response (“left”, dark line). Repeated application of the diffusion process yields histograms of both correct responses (upper histogram) and incorrect responses (lower histogram). As is evident from the histograms, the correct, upper “right” boundary is reached more often than the incorrect, lower “left” boundary. The total RT consists of the sum of a decision component, modeled by the noisy accumulation of evidence, and a nondecision component that represents the time needed for processes such as stimulus encoding and response execution.

In the simulation studies, we use the diffusion model to simulate plausible RT data that reflect the two hypothetical scenarios described in the previous section. For the first scenario, in which motivation changes systematically across trials, we simulated data in which all parameters of the diffusion model were constant, but drift rate $v$ changed. For the second scenario, in which response caution changes systematically across trials, we simulated data in which all parameters of the diffusion model were constant, but boundary separation $a$ changed.

### Simulation I: Spurious Post–Error Slowing

In this section we reproduce the first scenario described above: spurious PES as a consequence of global changes in motivation. In order to minimize the effects of sampling error, we simulated an experiment of 100,000 consecutive trials. Data were generated from the diffusion model using a default set of plausible parameters (Matzke & Wagenmakers, 2009): $a = 0.12, z = a/2, T_{er} = .300, s_z = .2a$ (i.e., across–trial variability in starting point), $\eta = 0$ (i.e., across–trial variability in drift rate), and $s_t = 0$ (i.e., across–

---

3For each panel shown in Figure 6.3, we simulated one experiment.
We used drift rate as a proxy for motivation. Note that in this simulation the PES effect was completely absent.

We explored the effect of four types of fluctuation in motivation (i.e., drift rate): drift rate constant, drift rate decreasing as a step function, drift rate decreasing linearly, and drift rate fluctuating as a sine function. The simulation results are shown in Figure 6.4 with one panel for each type of fluctuation (as denoted in figure headers). Each panel shows mean error RT, and mean correct RT for the three trials preceding an error and the three trials following an error. The asterisk represents mean correct RT for post–correct trials. The vertical grey bars quantify the traditional measure for PES, that is,  

\[ PES_{\text{traditional}} = MRT_{\text{post-error}} - MRT_{\text{post-correct}}. \]

When drift rate fluctuates systematically across trials (bottom–left, as a step function; top–right, linearly; bottom-right, as a sine function), the traditional measure leads to spurious detection of PES (the vertical grey bars). In contrast, the robust measure supports the correct conclusion: there is no PES. Error bars indicate one standard error of the mean.

In the top–left panel of Figure 6.4, drift rate is a constant \( v = .22 \) across trials. Consequently, mean RT is also constant and does not depend on the position relative to the error \( E \). The measure \( \bar{PES}_{\text{traditional}} \) appropriately indicates a negligible PES effect.

Now consider the upper right panel of Figure 6.4. The only difference with the upper left panel is that drift rate \( v \) is now linearly decreasing from \( v_{\text{high}} = .38 \) on the first trial to \( v_{\text{low}} = .06 \) on the last trial. Again, mean RT around errors is practically constant, suggesting (correctly) that there is no real PES. However, the traditional measure \( \bar{PES}_{\text{traditional}} \) now yields a spurious PES effect of 53 ms. The bottom left panel
shows the results of a similar simulation where drift rate is high and constant in the
first half of the experiment (i.e., \( v_{\text{high}} = .38 \)) and low and constant in the second half
of the experiment (i.e., \( v_{\text{low}} = .06 \)). The bottom right panel displays the results of a
simulation where drift rate was fluctuating according to a sine wave (with a period of
100 trials; \( \max(v) = v_{\text{high}} = .38; \min(v) = v_{\text{low}} = .06 \)). Just as in the simulation with
a linear change in drift rate, mean RT around errors is practically constant, suggesting
(correctly) that there is no real PES. However, the traditional measure \( \hat{\text{PES}}_{\text{traditional}} \)
yields spurious PES effects. Figure 6.4 also shows the results of the robust measure for
PES that we discuss later. This method appropriately indicates that there is no PES.

In sum, these simulations demonstrate that a systematic change in motivation biases
\( \hat{\text{PES}}_{\text{traditional}} \), possibly leading to spurious PES. Note that the direction of change is
irrelevant, since both an increase and a decrease in motivation yield similar local differ-
ences.

Simulation II: Masked Post–Error Slowing

In this section we reproduce the second scenario described earlier: masked PES as a con-
sequence of global changes in caution. In order to minimize the effects of sampling error,
we again simulated an experiment of 100,000 consecutive trials. Data were generated
from the diffusion model using \( v = 0.22, z = a/2, T_{er} = .300, s_z = .2a \) (i.e., across–trial
variability in starting point), \( \eta = 0 \) (i.e., across–trial variability in drift rate), and \( s_t = 0 \)
(i.e., across–trial variability in nondecision time). We used boundary separation as a
proxy for caution. In this simulation, we generated a PES effect by increasing boundary
separation after an error. Specifically, all trials following a correct response used a lower
boundary separation of \( a(t) - .008 \) (where \( a(t) \) indicates the local level of boundary sep-
aration that may be subject to systematic variation as indicated below), and all trials
following an error used a higher boundary separation of \( a(t) + .008 \).

We explored the effect of four types of fluctuation in caution (i.e., boundary separa-
tion): boundary separation constant, boundary separation decreasing as a step func-
tion, boundary separation decreasing linearly, and boundary separation fluctuating as a
sine function. The simulation results are shown in Figure 6.5 with one panel for each
type of fluctuation (as denoted in figure headers). As before, each panel shows mean
error RT, and mean correct RT for the three trials preceding an error and the three
trials following an error. The asterisk represents mean correct RT for post–correct tri-
als. The vertical grey bars again quantify the traditional measure for PES, that is,
\[
\hat{\text{PES}}_{\text{traditional}} = \text{MRT}_{\text{post–error}} - \text{MRT}_{\text{post–correct}}.
\]

In the top–left panel of Figure 6.5, boundary separation is a constant \( a(t) = .18 \)
across trials. Mean RT depends on the position relative to the error \( E \), and the measure
\( \hat{\text{PES}}_{\text{traditional}} \) appropriately indicates a PES effect of 34 ms.

Now consider the upper right panel of Figure 6.5. The only difference with the upper
left panel is that boundary separation \( a(t) \) is now linearly decreasing from \( a_{\text{high}} = .22 \)
on the first trial to \( a_{\text{low}} = .14 \) on the last trial. Again, mean RT around errors is not constant, suggesting (correctly) that there is PES. However, the traditional measure
\( \hat{\text{PES}}_{\text{traditional}} \) reports a deflated PES effect of only 6 ms. The bottom left panel shows
the results of a similar simulation where boundary separation is high and constant in
the first half of the experiment (i.e., \( a_{\text{high}} = .22 \)) and low and constant in the second
half of the experiment (i.e., \( a_{\text{low}} = .14 \)). The bottom right panel displays the results of
a simulation where boundary separation was fluctuating according to a sine wave (with
a period of 100 trials; \( \max(a) = a_{\text{high}} = .22; \min(a) = a_{\text{low}} = .14 \)). Just as in the
6.5 A Simple Solution

The above confounds arise because post–correct and post–error trials (i.e., the trials used to calculate $\overline{PES}_{traditional}$) are not evenly distributed across the time series. The confounds can be eliminated when we compare post–error trials to post–correct trials that originate from the same locations in the time series. One natural option is to use post–correct trials that are pre–error trials at the same time. So, instead of comparing the mean RTs of all post–error trials to those of all post–correct trials, we conduct pairwise simulation with a linear change in boundary separation, mean RT around errors is not constant, suggesting (correctly) that there is PES. However, the traditional measure $PES_{traditional}$ yields very small or even negative PES effects (i.e., post–error speeding). Figure 6.5 also shows that the robust measure for PES that we discuss later appropriately indicates that there is indeed PES.

In sum, these simulations demonstrate that a systematic change in caution biases $PES_{traditional}$, possibly leading to an underestimated or masked PES. Note that the direction of change is again irrelevant, since both an increase and a decrease in response caution yield similar local differences.

Figure 6.5: When boundary separation fluctuates systematically across trials (bottom–left, as a step function; top–right, linearly; bottom-right, as a sine function), the traditional measure fails to detect positive PES (the vertical grey bars). In contrast, the robust measure supports the correct conclusion: there is PES. Error bars indicate one standard error of the mean.
6. How to Measure Post–Error Slowing

comparisons around each error (see Nelson, Boucher, Logan, Palmeri, & Schall, 2010 for a similar method in a different context). In other words, we take the average of $RT(E+1) - RT(E-1)$ for all errors $E$. This comparison will now be referred to as the robust measure $\hat{PES}_{\text{robust}}$ as opposed to the traditional measure $\hat{PES}_{\text{traditional}}$.

Figures 6.4 and 6.5 allow a comparison of $\hat{PES}_{\text{robust}}$ and $\hat{PES}_{\text{traditional}}$. As anticipated, $\hat{PES}_{\text{robust}}$ is not much affected by global fluctuation in motivation and caution, but $\hat{PES}_{\text{traditional}}$ is. For the simulations without PES but with relatively pronounced fluctuations in motivation, $\hat{PES}_{\text{robust}}$ correctly estimates PES to be very small; for the simulations with PES but with relatively pronounced fluctuations in caution, $\hat{PES}_{\text{robust}}$ correctly estimates PES to be in between 29 and 48 ms. This behavior is the mirror opposite of $\hat{PES}_{\text{traditional}}$, a measure that consistently arrives at the incorrect conclusion.

6.6 Empirical Illustration: The Confound is Real

The simulations above showed that $\hat{PES}_{\text{traditional}}$ can detect spurious PES and mask real PES. We now provide an empirical illustration of these two situations. For this illustration, we selected data from two individual participants from a larger study that will be published elsewhere.

Method

Elderly participants were presented with the random dot motion task (Britten et al., 1992), a task that is popular in cognitive neuroscience and research on monkeys. The random dot stimulus consists of a circular display of dots. The dots appear, disappear and are replaced in such a way that the entire circle of dots appears to move either left or right. The apparent motion that the participant perceives can be best described as the flickering of a turning disco ball in a spotlight. This illusion is created as follows. At each frame (50 ms), 120 dots are displayed. Every next frame, an experimentally defined proportion $P_{\text{move}}$ of the dots from the former frame are shifted a certain distance $l_{\text{move}}$ to the target side (e.g., right, if the correct response is right). The remaining portion of the pixels is randomly replaced in the circle (independent of their previous positions). $P_{\text{move}}$ was set to 50%. $l_{\text{move}}$ was set to one pixel. Participants were instructed to indicate the direction of the apparent movement (left or right) by pressing one of two response buttons. For the data presented here, participants were asked to respond accurately, but not waste too much time. Two participants were selected for the current empirical illustration.

Results

Data for the two participants are shown in Figure 6.6. The interpretation of these panels follows that of Figures 6.4 and 6.5: each panel shows the mean RT of correct trials that either precede ($E-1, E-2, E-3$) or follow ($E+1, E+2, E+3$) an error. The asterisk again represents the mean post–correct RT, and the vertical grey bars indicates the traditional measure of PES, that is, $\hat{PES}_{\text{traditional}} = \frac{\text{MRT}_{\text{post-error}} - \text{MRT}_{\text{post-correct}}}{2}$.

The participant whose data ($n = 1400$ trials) are displayed in the upper panel of Figure 6.6 shows a pattern similar to the synthetic patterns shown in Figure 6.3. The

4Recall that the PES effect was simulated as an increase in boundary situation. The resulting effect on mean RT is moderated by the fluctuations in baseline boundary separation across trials.
6.7 Concluding Comments

In the last two decades, cognitive control has become a major research field in psychology. PES is assumed to be an indicator of cognitive control and as such it is often used as an important behavioral variable. Recently, however, some studies have questioned whether PES really reflects cognitive control (e.g., Notebaert et al., 2009; Dutilh, Vandekerckhove, et al. in press). In this study, however, we focused on a more elementary issue regarding the application of PES as a dependent variable: the reliability of the quantification of PES.

5The two participants have unequal number of trials because the testing protocol was to collect as many data in a fixed time period.
The traditional method to quantify PES compares mean RT after errors with mean RT after correct responses. We argued that this analysis is vulnerable to two confounds that both result from fluctuations of a participant’s performance over the course of an experiment. These confounds are global changes in motivation and global changes in response caution. We showed how these confounds may lead to both spurious and masked PES effects, both in simulations and in two empirical data sets. The solution we offered is both simple and adequate: in the comparison of post–error mean RT and post–correct mean RT, only those post–correct trials should be included that are also pre–error trials. This additional condition of our robust method ensures that post–error and post–correct trials originate from the same locations in the data set. This property of the robust method makes it immune to global fluctuations in performance.

The robust method, that compares pre–error post–correct trials to post–error trials, allows for pairwise comparison of RTs. As such, the robust method strongly increases the power to detect PES. A potential complication of this method is that both pre–error and post–error trials can be either correct or incorrect trials. Therefore, sometimes one is contrasting two correct trials, sometimes two error trials, and sometimes one correct and one error trial. When errors are systematically faster or slower than correct responses, such a mixed set of comparisons increases the variability of the PES measure. The ultimate solution to this problem is to apply a model that accounts for both RT and accuracy. The most popular model to do so is Ratcliff’s diffusion model (described in Section “The Diffusion Model”). The model offers a method to describe the interactions between RT distributions and accuracy as the product of underlying psychological components. For the quantification of PES, this decomposition is particularly valuable, because a researcher often has very clear ideas about the psychological processes that underly an increase in RT after errors. If the psychological process that underlies PES is assumed to be response caution (which often is the case in studies on cognitive control, as we discussed above), the direct measure of this phenomenon is the boundary separation parameter of the diffusion model, rather than mean RT. Dutilh, Vandekerckhove, et al. (in press) have applied the diffusion model to show that, for a different data set, boundary separation is indeed the important parameter that changes after an error.

The field of research that uses PES as a behavioral variable has expanded strongly over the past decade. The PES phenomenon itself however has not gained interest proportionally. Consequently, little attention has been given to the theoretical background of PES, let alone related methodological issues. There seems to be no particular reason why researchers have stuck to the traditional method of quantification. The robust method we propose here is easy to apply, insensitive to changes in performance over the course of an experiment, and hence allows a more informative and veridical measurement of post–error slowing.
Abstract

Most models of response time in elementary cognitive tasks implicitly assume that the speed–accuracy trade–off is continuous: when payoffs or instructions gradually increase the level of speed stress, people are assumed to gradually sacrifice response accuracy in exchange for gradual increases in response speed. This trade–off presumably operates over the entire range from accurate but slow responding to fast but chance-level responding (i.e., guessing). In this article, we challenge the assumption of continuity and propose a phase transition model for response times and accuracy. Analogous to the fast guess model \(Ollman\,1966\), our model postulates two modes of processing: A guess mode and a stimulus controlled mode. From catastrophe theory we derive two important predictions that allow us to test our model against the fast guess model and against the popular class of sequential sampling models. The first prediction—hysteresis in the transitions between guessing and stimulus controlled behavior—was confirmed in an experiment that gradually changed the reward for speed versus accuracy. The second prediction—bimodal response time distributions—was confirmed in an experiment that required participants to respond in a way that is intermediate between guessing and accurate responding.
7. Phase Transitions in SAT

7.1 Introduction

One of the key phenomena in response time research is the speed–accuracy trade–off, by which a decision maker can speed up at the expense of accuracy and become more accurate at the expense of speed ([Schouten & Bekker, 1967; Wickelgren, 1977; Bogacz; Wagenmakers, Forstmann, & Nieuwenhuys, 2010]). The interdependence of response time (RT) and accuracy implies that people can be accurate and slow in one situation, yet fast and inaccurate in another, although their efficiency in information processing did not change. The speed–accuracy trade–off therefore frustrates a straightforward interpretation of RT in terms of cognitive processing time, and forces researchers to jointly consider RT and accuracy.

Most models that account for the speed–accuracy trade–off, including most sequential sampling models, implicitly assume that the speed–accuracy trade–off is a continuous function. This assumption implies that a participant who is responding accurately on a certain task can gradually increase speed at the cost of gradual decreases in accuracy, until speed reaches ceiling and accuracy is at chance level (i.e., fast guessing). Here we challenge this assumption and hypothesize that, with increasing pressure to respond quickly, relatively accurate behavior suddenly collapses into guessing behavior, without going through all the intermediate stages between accurate responding and guessing.

To account for this discontinuous shift in performance, we introduce a phase transition model for the speed–accuracy trade–off. The model postulates that guessing and stimulus controlled responding are irreconcilable modes of processing. This means that, when experimental settings continuously change and force people to switch from one mode of processing to the other, this switch will be abrupt. When participants are, for example, forced to speed up over trials (and become less careful), at first they will be able to persist in fairly accurate responding. However, with a gradual increase in speed stress performance will at some point break down completely and participants abruptly resort to fast guessing. Our model predicts a similar abrupt switch when the experimental conditions gradually encourage participants to stop guessing and be more careful (and respond more slowly).

Our phase transition model finds its roots in Ollman’s fast guess model ([Ollman, 1966]). However, our model offers a more dynamic account of the speed–accuracy trade–off and allows for a connection to sequential sampling models of RT such as Ratcliff’s diffusion model ([Ratcliff, 1978]). The phase transition model has the form of a cusp model from catastrophe theory. Catastrophe theory is a mathematical theory that applies to dynamical systems in which continuous changes of environmental variables lead to sudden changes of observed behavior (e.g., [Zeeman, 1976]). From this model, we derive two signature predictions of the phase transition model: hysteresis and bimodality. We test these two predictions in two experiments.

The outline of this paper is as follows. In the first section, we discuss sequential sampling models and varied state models of the speed–accuracy trade–off. In the second section, we introduce the phase transition model. In the third section, we explain the two experiments that test the predictions of our model. Next the experimental data are described and discussed by means of quantitative models (i.e., a hidden Markov model and our cusp model).
7.2 The Speed–Accuracy Trade–off

In many speeded choice tasks, participants are instructed to respond “as fast and accurately as possible”. These instructions leave it to the participant to assess the relative importance of speed versus accuracy. This implies that both RTs and proportion of errors depend largely on the participant’s judgment. Therefore, the separate analysis of either mean RT or accuracy can be deceiving. Only through an understanding of how RT and accuracy trade off can observed behavior be translated into conclusions in terms of psychologically interesting constructs.

The exact nature of the trade–off between speed and accuracy has been studied for almost a century (Henmon, 1911). Over the years, many studies have been devoted to the speed–accuracy trade–off (from now on referred to as “SAT”). One approach of studying the SAT is to explore experimentally the entire range from chance performance (i.e., guessing) to asymptotic accuracy (e.g., R. Pachella & Pew, 1968; Swensson, 1968; Wickelgren, 1977; Yellott, 1971). In most of these studies, the trade–off is assumed to be under experimental control through the use of response deadlines, response signals, differential pay–off, and various other methods (e.g., Schouten & Bekker, 1967; R. Pachella & Pew, 1968; Meyer, Irwin, Osman, & Kounios, 1988; Verhelst, Verstralen, & Jansen, 1997). The objective of these studies was to formulate a function that describes how participants move along the hypothetical SAT-curve.

The behavior at the extremes of the hypothetical SAT curve is uncontroversial: When rewards only emphasize accuracy, participants respond accurately but slowly; when rewards only emphasize speed, participants respond fast but at chance accuracy. The controversy, however, is about how participants can shift from highly accurate to very fast performance. Figure 7.1 summarizes several conflicting points of view. Lines A1 and A2 in the left panel represent continuous accounts of this shift. Function A1 represents the predictions of the SAT by most RT models, including sequential sampling models (Wickelgren, 1977; Figure 1). This functional form is also found in some empirical studies (e.g., B. Dosher, 1979, Ratcliff, 2006). Another form of the SAT function that is reported in some empirical studies is represented by sigmoid function A2 (e.g., Schouten & Bekker, 1967).

A very different account is given by models that assume that behavior originates from different states. In particular, the fast guess model (Ollman, 1970), predicts the stepwise (discontinuous) trade–off depicted by line B1. In this model, behavior originates from one of two states. The phase transition model that we present in this article also predicts a discontinuous SAT and is represented by line B2.

7.3 Current Models of the Speed–Accuracy Trade–Off

Sequential Sampling Models

Sequential sampling models form the dominant class of models to account for both RT distributions and accuracy. Its members include, for instance, Ratcliff’s diffusion model (Ratcliff, 1978), the leaky competing accumulator model (Usher & McClelland, 2001), the linear ballistic accumulator model (Brown & Heathcote, 2002, Brown & Heathcote, 2008), and Poisson counter models (e.g., P. L. Smith & Van Zandt, 2000). Sequential sampling models generally produce a good fit to behavioral data, and they allow researchers to decompose effects on RT and accuracy into effects on underlying psychological constructs (e.g., Ratcliff & Rouder, 1998; Dutilh et al., 2009). Motivated in part by the availability of easy–to–use fitting routines, sequential sampling models are applied increasingly often,
Figure 7.1: The speed accuracy trade–off. A: Continuous trade–off of speed for accuracy, as predicted by most models of RT (A1), and as found in some empirical studies (A2). B: Two discontinuous trade–off functions. B1 is the functional form predicted by the fast guess model. B2 is the functional form predicted by the phase transition model.

Both in experimental psychology (e.g., Wagenmakers, 2009) and in the neurosciences (e.g., Bogacz et al., 2010; Forstmann et al., 2008).

All sequential sampling models postulate a decision making system that samples stimulus information over time. Often, but not always, this information is assumed to be noisy. The accumulated information reflects the evidence for each of the possible response options (see the meandering line in the left panel of Figure 7.2). When the evidence for one response option reaches a pre–set response criterion (boundary A), the response is initiated. The setting of the response criterion in such a model governs the SAT. When the response criterion is high, the system requires strong evidence before a response is initiated, and this results in responses that are accurate but slow. When the response criterion is low, the system requires only weak evidence to decide, and this results in responses that are fast but inaccurate. The right panel of Figure 7.2 shows how the distribution of RT changes when the response criteria (boundaries) are set so as to generate percentages correct of 95% (boundary A), 75% (boundary A’), and 55% (boundary A”). Decreasing boundary separation not only leads to lower accuracy, but also to faster responses and smaller spread of the RT distribution. Note that all intermediate values of speed and accuracy are accessible, that is, they can be achieved by an intermediate setting of response criteria. In this sense, sequential sampling models predict a continuous SAT.

It should be acknowledged that sequential sampling models assume only implicitly that the transition from very fast to highly accurate behavior is smooth and continuous. Although the stability of intermediate trade–offs seems a crucial assumption of sequential sampling models, one could imagine a mechanism that governs discrete changes in boundary setting that would result in discrete changes in behavior. However, as we will discuss in the concluding remarks, we think that this a rather unnatural interpretation of the boundary principle and that the resulting RT model would not be very plausible.

The Fast Guess Model

A different account of the SAT is given by varied state models, which assume that behavior originates from various separate states. The most extensively studied varied state model is Ollman’s simple fast guess model (Ollman, 1960, 1970). The simple fast guess model assumes that the behavior in choice RT tasks is governed by two distinct processes, namely
7.3. Current Models of the Speed–Accuracy Trade–Off

Figure 7.2: Illustration of a generic diffusion–style sequential sampling model. The lower left panel shows a possible trajectory of how sampled information is accumulated until a response boundary is reached. The RT distribution associated with this process is drawn above. The three hypothetical boundary settings (A, A', and A'') in the left panel are associated with three different RT distributions (right panel) and proportions of correct responses (given a positive drift).

a guess mode (sometimes called the pre–programmed mode) and a stimulus controlled mode.

The guess mode (GM) corresponds to the way information is processed in simple detection tasks, i.e., no discrimination between stimuli occurs. Responses in this mode are fast and accuracy is at chance level. In the stimulus controlled mode (SCM), discrimination between stimuli does occur and hence responses are slower and accuracy approaches 100%. Consequently, intermediate values of RT and accuracy can only be achieved by mixing responses from the two modes. Based on this fast guess model, Yellott (1971) proposed an easy procedure to correct mean RT estimates for fast guessing. Link (1982) discussed this correction for two state models in general.

The predictions of the fast guess model were supported by the studies of Swensson and Edwards (1971), Swensson (1972), and Yellott (1971). In these studies, the pressure on speed relative to accuracy was changed in small steps over the entire domain of the SAT. The results showed that, with stimuli that are difficult to discriminate, participants achieved intermediate values of accuracy by mixing fast guesses and stimulus controlled responses. Townsend and Ashby (1983), Luce (1986), and Yantis, Meyer, and Smith (1991) review several tests of the simple fast guess model and conclude that, when the experimental stimuli are easy to discriminate, the results were ambiguous, either supporting a continuous SAT or a discrete SAT.
From the perspective of the fast guess model, optimal performance involves choosing the response strategy (i.e., guessing or stimulus controlled responding) that is most profitable for the current payoff settings, and to apply this strategy on every trial that features the same payoff settings. Hence, in the fast guess model, changing demands on speed versus accuracy yields a discrete and stepwise SAT, at least when people consistently apply one and the same strategy for a specific payoff setting.

The consistent response strategy described above is optimal according to the fast guess model. However, as suggested by the results of Swensson (1972), participants might use a suboptimal “probability matching” strategy and meet increasing demands for speed by increasing the proportion of fast guesses. This mixing of response modes would yield a continuous SAT when trials are averaged.

7.4 The Phase Transition Model

Introduction

The fast guess model predicts a very simple trade-off function. As argued above, when the relative payoff for speed versus accuracy is changed from emphasizing only accuracy through emphasizing only speed, the model predicts that optimal behavior requires a stepwise speed accuracy trade-off (line B1 in Figure 7.1). A more advanced model of the dynamics can be formulated by use of the concept of phase transitions developed in research on nonlinear dynamical systems.

Phase transitions occur in all kinds of systems, ranging from those that are physical, chemical and biological (e.g., Poston & Stewart, 1978b) to those that are social and psychological (e.g., Jansen & Van der Maas, 2011; Latané & Nowak, 1993; Schöner, Haken & Kelso, 1986; Stewart & Perego, 1983; Zeeman, 1975). These phase transitions have been studied from various theoretical perspectives, including catastrophe theory, synergistics, and non-equilibrium thermodynamics. These perspectives are mathematically similar. Here we focus on the catastrophe perspective because it is especially useful in cases where it is not feasible to derive an exact mathematical model of the transition process under investigation (Wagenmakers, Van der Maas, & Molenaar, 2005).

Catastrophe Theory

We limit our review of catastrophe theory to a small number of concepts that are required for the present purposes (for more details, see Arnold, Afrajmovich, Il’yashenko & Shilnikov, 1999; Castrigiano & Hayes, 1993; Gilmore, 1981; Poston & Stewart, 1978a; Thom, 1975). Catastrophe theory, a branch of bifurcation theory, is a mathematical theory about dynamical systems that are governed by the gradient of a potential function. Such systems optimize some quantity, like energy or profit. Consequently, the behavior in these systems attains those values, called equilibrium states, that lead to a zero gradient (defined as the first derivative of the potential function). Catastrophe theory describes and classifies changes in equilibrium behavior. These changes come about when smooth changes of the system’s parameters lead to the sudden appearance or disappearance of stable states. The simplest catastrophe, in which such discontinuities occur, is the cusp catastrophe.

To get insight in the dynamics of the cusp catastrophe, Figure plots the equilibria of the cusp catastrophe. To illustrate the function of the axes in Figure, we use the famous example of the phase transition between the liquid and solid states of water (Poston & Stewart, 1978). For relatively high values of pressure ($\beta$), such as the pressure
at sea level, smooth changes in temperature ($\alpha$) lead to sudden jumps between the solid and liquid phase of water ($Z$). In cusp terminology, $\beta$ is the splitting axis, and $\alpha$ is the normal axis. $Z$ represents a function of the behavioral variables, in this example the state of the water.

![Diagram of the cusp catastrophe model]

Figure 7.3: The equilibrium of the cusp catastrophe describes behavior as a function of control variables $\alpha$ and $\beta$. Jumps in behavior take place when the setting of control variables leaves the area within the bifurcation lines, in which the system exhibits bimodality. The ball in the valleys represents the state that the system adopts when possible states appear and disappear as a function of $\alpha$.

The cusp has eight distinguishing properties, which are known as catastrophe flags. These catastrophe flags, formally derived by Gilmore (1981), can be used to test the model empirically and to detect phase transitions (Van der Maas & Molenaar, 1992). The most important flags are illustrated in Figure 7.3. Sudden jump is a sudden large change in behavior. In the water example, this describes the sudden temperature–induced change from water to ice and vice versa that takes place when pressure $\beta$ is large. Bimodality means that two stable modes of behavior exist for a range of values of the independent or control variables. Inaccessibility means that the behavior in between these stable modes is unstable and repelling; no stable state in between water and ice exists. However, when pressure $\beta$ is low, such an intermediate, syrupy state exists. Hysteresis is a delay in the sudden jump when control variables are changed up and down. In terms of the freezing water example, hysteresis refers to the fact that, when the pressure is high and in perturbation free conditions, water freezes at $-4^\circ$C and ice melts at $0^\circ$C. Divergence refers to a strong dependency on the initial conditions with respect to the mode of behavior that will be selected. Anomalous variance refers to a strong increase in variability in behavior near the sudden jump. The last two flags specify the effect of external perturbations of the system: Critical slowing down refers to delayed recovery of equilibrium behavior.
and divergence of linear response refers to large oscillations induced by perturbations. A more elaborate explanation of catastrophe theory and the catastrophe flags can be found in, e.g., Gilmore (1981) and Ploeger, Van der Maas, and Hartelman (2002).

Critique of Catastrophe Theory

Catastrophe theory carried with it the prospect that all kinds of systems could be formally described, as long as they exhibit phase transitions. This prospect caused a burst of applications in the 1970’s, until Sussmann and Zahler (1978) articulated strong reservations about applications of catastrophe theory in the social and behavioral sciences. The main concerns that were raised are the following. First, in contrast to its formal and deterministic nature, catastrophe theory was often applied to qualitative concepts and stochastic variables. Second, the choice of control variables was often arbitrary and not well motivated. Today, catastrophe theory has regained importance in many fields of science (e.g., K. M. Newell, Liu, & Mayer-Kress, 2000; Wales, 2001; Tamaki, Torii, & Maeda, 2003). As in other modern applications of catastrophe theory, our application to the SAT is scientifically informative and falsifiable. Both our control variables, payoffs for RT and accuracy, and our dependent variables, RT and accuracy, are well–defined and naturally observable quantities. Furthermore, we test both the quantitative and qualitative predictions of the cusp model.

The Phase Transition Model of SAT

Now that we have discussed the general properties of the cusp catastrophe, we formulate a cusp model for the SAT. Below, we will first show how the SAT can be mapped onto the cusp catastrophe. Second, we will describe the dynamics implied by this mapping. Then, having formulated the SAT phase transition model, we will derive testable predictions.

Optimality

The use of potential functions in catastrophe theory is based on the assumption that the system under investigation optimizes (minimizes or maximizes) some quantity. In RT experiments this assumption is uncontroversial since participants are asked to respond as quickly and accurately as possible, or to maximize profits. The experimental settings thus define the function that is optimized.

Definition of the behavioral variable

The main behavioral variables in RT experiments are response time and/or accuracy. Since RT and accuracy strongly covary in the SAT, we can take RT, accuracy, or some function of both as the behavioral variable Z. For graphical representations of the behavior we will use RT here; RT is a continuous measure and therefore has better measurement properties than does accuracy. In the statistical analyses with hidden Markov models, we will fit both RT and accuracy simultaneously.

In accordance with the fast guess model, we hypothesize that the stable states of these behavioral variables are the guess mode (GM)—where responses are fast and accuracy is at chance level—and the stimulus controlled mode (SCM)—where responses are slow but accuracy is high. The availability of these modes or states will depend on the values of

---

1For a detailed response to Sussmann and Zahler (1978) see the online appendix available on the first author’s website.
the control variables. Note that the phase transition model, however, does not describe the process by which responses are generated in both modes.

Choice of control variables
The control variables in a cusp catastrophe are the normal factor ($\alpha$), which is associated with the hysteresis effect, and the splitting factor ($\beta$), which is associated with divergence. It seems obvious to relate the normal factor $\alpha$ to the traditional manipulations of the SAT (e.g., deadlines, instructions, and payoffs), because they force the participant to select one of the modes. Analogous to other catastrophe models in psychology ([van der Maas & Molenaar, 1992; Zeeman, 1976; Latane & Nowak, 1994]), we propose to relate the splitting factor $\beta$ to motivation or involvement. Involvement in RT tasks can be quantified for instance by the reward that can be earned in the experiment. In a cusp catastrophe, the size of the jumps and the magnitude of the hysteresis effect depend on the value of the splitting factor. Therefore, we hypothesize that only when rewards are significant, strong discontinuities in behavior occur. Below, we formulate this hypotheses more precisely in terms of payoffs.

Dynamics
Payoffs are factors that weigh the speed and accuracy of the response in the computation of a reward, denoted $R_t$, that a subject receives on every trial $t$. We distinguish two types of payoffs, the payoff for speed $P_{RT}^t$ and the payoff for accuracy $P_{Acc}^t$. When the payoff for speed is near zero and the payoff for accuracy is large, we expect that participants select the stimulus controlled mode. When $P_{RT}^t$ is large and $P_{Acc}^t$ is near zero, we expect that participants select the guess mode. Finally, when both factors are large and thus the involvement is high, the typical speed–accuracy conflict is expected to arise. In this conflict situation, we expect sudden mode switches. We can express these predictions in the cusp catastrophe, in which the normal and splitting axis are functions of the payoff factors. The payoff factor for speed induces the guess mode and the payoff factor for accuracy induces the stimulus controlled mode (see Figure 7.4). A rotation of 45° of these factors leads to clear definitions of the normal and splitting axis of the cusp. They are simple functions of the difference and the sum of the payoff factors, respectively.

$$\alpha = a(P_{Acc} - P_{RT})$$
$$\beta = b(P_{Acc} + P_{RT})$$

where $a$ and $b$ are scaling parameters.

Given these choices of the behavioral and control variables, we can derive clear predictions about behaviors corresponding to the catastrophe flags. When involvement $P_{Acc} + P_{RT}$ is high, Sudden jumps between the modes should occur when payoff factors are varied. Bimodality and inaccessibility are expected within the bifurcation lines. Hysteresis should occur when $P_{Acc} - P_{RT}$ varies at a high constant value of $P_{Acc} + P_{RT}$. Jumps to the stimulus controlled mode occur when $P_{Acc}^t$ is (much) higher than $P_{RT}^t$, whereas jumps to the guess mode are expected when $P_{RT}^t$ is (much) higher than $P_{Acc}^t$ (hysteresis). If both payoff factors are zero, behavior is in the vicinity of the neutral point $(0,0,0)$ and neither mode is selected. Here participants may either guess slowly or not respond at all. This seems reasonable since they can not win or loose anything (i.e., there is no involvement whatsoever). Divergence is expected when we increase involvement $P_{Acc} + P_{RT}$ but hold $P_{Acc} - P_{RT}$ at zero. The participant may then choose
the stimulus controlled mode or the guess mode, but the participant can not maintain an intermediate position (inaccessibility). *Anomalous variance* implies that the variance of Z (i.e., RT and accuracy) in each of the modes increases strongly near the jump. If the subject is perturbed (by incorrect feedback, for instance) behavior should show large oscillations (*divergence of linear response*), which take a long time to fade away (*critical slowing down*).

![Graphical representation of the phase transition model.](image)

Figure 7.4: Graphical representation of the phase transition model. At high values of $P_{RT} + P_{Acc}$, the system consists of two irreconcilable modes, the stimulus controlled mode and the guess mode. At intermediate values of $P_{RT} - P_{Acc}$, both states are possible. Which state is adopted at a certain moment depends on whether the system was in either the guessing state or the stimulus controlled state before that moment.

**Empirical Predictions**

The phase transition model we propose here can be considered a generalization of the fast guess model. As in the fast guess model, the phase transition model assumes two stable states, a guessing state and an accurate state. However, in the fast guess model, the switches between these states take place at the same point in either direction, i.e., when $P_{Acc} = P_{RT}$. The phase transition model, on the other hand, predicts that when the stakes $P_{Acc} + P_{RT}$ are high enough, hysteresis occurs in the switching between the two states, that is, the switches between these states do not take place at the same point in either direction. In catastrophe theory the size of the hysteresis effect depends on noise in or perturbations of the system (Gilmore, 1981). Without noise the hysteresis effect will be strong (so called Delay convention), with noise the hysteresis effect diminishes until switches between modes take place at the same point (the Maxwell convention), as in the fast guess model.
This hysteresis prediction implies that when the subject has to speed up, accurate responding might persist well beyond the point where the fast guess model, with its stepwise trade-off, predicts a collapse of accurate responding. This phenomenon relates to a phenomenon that many chess players experience when playing a game with less and less time on the chess clock. Although difficult to test, in many sports and activities performance seems to be relatively unaffected, albeit within a certain range, by time pressure.  

Standard manipulations of the SAT are not able to reveal the presence of hysteresis. Therefore, to discriminate between the phase transition model and the fast guess model, we designed an experiment where changing payoffs push a participant from accurate behavior (SCM) to guessing (GM) and vice versa, thereby switching in either direction between the two hypothesized states. A possible limitation of this procedure is that participants may perceive changes in payoffs with a delay, potentially resulting in an artificial hysteresis effect. To exclude this possibility, we included an extra experiment (1c), based on a “modified method of limits” (Hock, Kelso, & Schöner, 1993).

The difference between the phase transition model and most sequential sampling models is arguably more structural. Whereas sequential sampling models assume that all responses originate from a single unitary process, the phase transition model predicts that two distinct modes of processing underly the behavior. When the two modes of processing that underly the behavior are clearly separated, bimodality of the behavioral variables is expected. Standard manipulations of the SAT (“respond as fast and accurately as possible”) cannot reveal this bimodality, since participants usually respond at above 90% correct, where the phase transition model does not predict bimodality. Bimodality is expected at intermediate values of accuracy (e.g., 75% correct). Therefore, in order to test the phase transition model against pure sequential sampling models, we designed an experiment in which participants were pressed to respond at 75% correct, as fast as possible.

Below, we describe two experiments designed to detect hysteresis and bimodality. We first present the hysteresis experiment and a follow-up extension that corroborates the initial results. Second, we present the bimodality experiment.

7.5 Experiment 1: Hysteresis

The hysteresis experiment was conducted in two stages. In Experiment 1a, three participants (A-C) were tested. Experiment 1b was conducted as an autonomous replication of Experiment 1a. In this replication, eight more participants (D-K) were tested with a slightly improved experimental design. The modified method method of limits experiment is discussed in the next section.

Method Experiment 1a

In the hysteresis experiments, we aimed to force the participants to perform over the entire range of the SAT. Therefore, we applied a reward function that incorporates both speed and accuracy, enabling us to change the emphasis on speed versus accuracy in small steps. By smoothly changing the reward for speed versus accuracy back and forth, one could argue that trading off accuracy for speed is not so much a free strategic choice, but a competence in itself. It requires a competence to operate close to the transition point, at a high speed, very near a collapse to complete inaccurate responding. This competence may be an age-related factor in the explanation of trade-off differences in age groups (e.g., Botvinick et al., 2001; Rabbitt, 1979).
from a situation with complete emphasis on speed to a situation with complete emphasis on accuracy it is possible to reveal hysteresis in the transitions between guessing and stimulus controlled responding.

Participants

Three students from the University of Amsterdam participated for course credit. Students were native speakers of Dutch and did not participate in any of the other experiments.

Materials and procedure

We used a lexical decision task, in which participants were required to discriminate word from non–word stimuli by pressing either the “z” button (for “word”) or the “/” button (for “non–word”). We choose to use a lexical decision task, because response times from accurate responses are expectedly easy to distinguish from fast guesses. The stimuli were sampled randomly without replacement from a list of 120 words and 120 non–words. When all stimuli were used, sampling started from the whole list again. Macromedia’s Authorware for Macintosh was used to both present the stimuli (black on white) and register the response (at a resolution of 60 Hz, i.e., precision of 16.6 ms). The response stimulus interval varied randomly between 1000 and 3000 ms, to prevent anticipatory responses. To further discourage this behavior, responses faster than 80 ms resulted in a “too early” warning message.

On every trial, the participant was rewarded according to

\[ R_t = R_t^{RT} + R_t^{Acc}, \]  

(7.1)

where \( R_t \) is the total reward earned on the current trial, and \( R_t^{RT} \) and \( R_t^{Acc} \) are the rewards for speed and accuracy, respectively. First, \( R_t^{RT} \) is given by

\[ R_t^{RT} = P_t^{RT} \times \left( \frac{RT_{SCM} - RT_t}{RT_{SCM} - RT_{GM}} \right), \]  

(7.2)

where \( P_t^{RT} \) (\( P^{RT} \in [0, 24] \))\(^3\) is the payoff weight for speed at trial \( t \), \( RT_{SCM} \) is the mean RT in the stimulus controlled mode, \( RT_{GM} \) is the mean RT in the guess mode, and \( RT_t \) is the response time at trial \( t \). Equation (7.2) ensures that responding at guessing speed (\( RT_t = RT_{GM} \)) yields an expected reward of \( R_t^{RT} = P_t^{RT} \).\(^4\) At the same time, responding slow enough to yield perfect accuracy (i.e., \( RT_t = RT_{SCM} \)) yields no reward for speed at all (\( R_t^{RT} = 0 \)).

Second, \( R_t^{Acc} \) is given by

\[ R_t^{Acc} = P_t^{Acc} \times Acc_t, \]  

(7.3)

where \( P_t^{Acc} \) is the current payoff weight for accuracy and \( Acc_t \) is -1 when the response on trial \( t \) is an error and 1 when the response is correct. This rule rewards responding

\[^3\]The value of 24 of the payoff weights was chosen arbitrarily to scale the reward rule.

\[^4\]E(\( R_t \mid GM \)) = \( P_t^{RT} \left( \frac{RT_{SCM} - RT_{GM}}{RT_{SCM} - RT_{GM}} \right) + \left( .5 \times -P^{Acc} + .5 \times P^{Acc} \right) = P_t^{RT} \).
accurately by the amount of $P_{Acc}^t$. At the same time, guessing yields an expected reward of zero, since half of the time, Acc will be $-1$, and half of the time Acc will be $+1$.

The reward for speed and accuracy and the total reward were presented to the participant on a feedback screen (Figure 7.5). The values of the rewards were represented by horizontal bars, that, when positive, were green and rightwards, and that, when negative, were red and leftwards.

Figure 7.5: On the feedback screen, the reward for speed, accuracy and the summed total were represented with three horizontal bars underneath the stimulus. These bars were green when positive (grey in figure, note the reward for accuracy) and red when negative (black in figure, note the negative reward for speed). The horizontal bar above the stimulus was displayed only in Experiment 1b and 1c. This bar contained an orange portion on the left side (grey in figure) that represented the current value of $P_{RT}^t$, and a blue portion on the right side (black in figure) that represented the current value of $P_{Acc}^t$.

The values for $RT^{SCM}$ and $RT^{GM}$ were determined by each participant’s performance in two training blocks. In one block, $P_{RT}^t$ was set to zero (for estimating $RT^{SCM}$). In the other block, $P_{Acc}^t$ was set to zero (for estimating $RT^{GM}$).

To test for hysteresis, $P_{Acc}^t + P_{RT}^t$ was kept constant at the arbitrary value of 24, whereas the difference $P_{Acc}^t - P_{RT}^t$ was varied step by step between $-24$ and $+24$. We used a simple adaptive algorithm to adjust the payoff factors. A session always started with $P_{RT}^t = 24$ and $P_{Acc}^t = 0$. At these payoff settings, the participant was supposed to engage in fast guessing, because only speed is rewarded. Once it was established that the participant indeed engaged in fast guessing (mean RT over the last 5 trials smaller than $RT^{GM} + 50ms$), we increased $P_{Acc}^t$ and decreased $P_{RT}^t$, in steps of 1 on each new trial. When the participant had given 5 consecutive correct responses, the direction of the steps reversed, i.e., $P_{RT}^t$ was increased and $P_{Acc}^t$ was decreased, until the participant again met the aforementioned criteria for guessing. By following this simple algorithm, the direction of change of the payoff factors reversed as soon as the participant was stable, either in the guess mode or the stimulus controlled mode. The algorithm ensures that participants passed the bifurcation set (i.e., the area where sudden jumps can take place) as often as possible, generating the essential data for a test of the hysteresis hypothesis.

\[
E(R_t|SCM) = P_{RT}^t \left( \frac{RT^{SCM} - RT^{SCM}}{RT^{SCM} - RT^{GM}} \right) + P_{Acc}^t = P_{Acc}^t
\]
The lexical decision task was administered in blocks of 100 to 200 trials, depending on how many trials a participant took to engage in the required mode of responding. Participant A was tested on three occasions within three weeks and the other participants were tested once for an hour. The participants were trained on a slightly different task for about 20 minutes. In this training task, the payoff factors were fixed within blocks of trials. Each block ended when the total reward exceeded 200, which took between 10 and 25 trials. In each block, the payoff setting was different, so the participants learned to optimize their performance at each payoff setting. This training was very important, since naive participants tend to show sub–optimal behavior when \( P_{\text{Acc}}^t \) is small and \( P_{\text{RT}}^t \) is large. Specifically, participants tend to display stimulus controlled behavior, whereas, at these payoff settings, fast guessing is much more rewarding.

**Results Experiment 1a**

Below, we will refer to a set of trials in which the participants switched from guessing to accurate responding and vice versa as a series (in which \( P_{\text{Acc}} \) first increases and then decreases\(^6\)). The first series for each participant served as training on the main task, leaving 372, 436, and 299 trials for A, B, and C, in which they completed 12, 14, and 10 series respectively.

![Participant A Payoff for Accuracy](image)

![Participant B Payoff for Accuracy](image)

![Participant C Payoff for Accuracy](image)

Figure 7.6: Experiment 1a: Mean RT increases with \( P_{\text{Acc}} \). When \( P_{\text{Acc}} \) decreases (and participants are speeding up), participants are slower at intermediate values of \( P_{\text{Acc}} \) than when \( P_{\text{Acc}} \) increases (slowing down). The BIC’s allow for comparison between a linear model in which RT is regressed on both \( P_{\text{Acc}} \) and direction of change (2–factor), and the model with only \( P_{\text{Acc}} \) as predictor (1–factor). Lower BIC’s indicate the better model. Also, the posterior probabilities (Schwarz weights) for both models are reported.

**Descriptive results**

For each participant, mean RT was calculated at each value of \( P_{\text{Acc}} \), for both directions of change in \( P_{\text{Acc}} \) separately, as shown in Figure 7.6. As expected, we found that high values of \( P_{\text{Acc}} \) provoked slow and accurate responding, whereas low values of \( P_{\text{Acc}} \) provoked fast

\(^6\)We only write the change of \( P_{\text{Acc}} \) for brevity. Note that the change in \( P_{\text{RT}} \) is implied, since: \( P_{\text{RT}}^t = 24 - P_{\text{Acc}}^t \).
7.5. Experiment 1: Hysteresis

Figure 7.7: Experiment 1a: Mean accuracy increases with $P^{Acc}$. With sparse data, visual inspection of the binary variable accuracy does not allow one to draw clear conclusions about the presence or absence of hysteresis.

and inaccurate behavior. Note, however, that mean RTs at intermediate values of $P^{Acc}$ are higher when the participants are directed away from slow, accurate responding (i.e., decreasing $P^{Acc}$), than when they are directed away from fast and inaccurate responding (i.e., increasing $P^{Acc}$). In support of this finding, model selection based on the Bayesian Information Criterion (BIC) prefers a linear model that regresses RT on both $P^{Acc}$ and direction of change over a linear model that regresses RT on $P^{Acc}$ only (see BIC values in Figure 7.6). BICs quantify the relative performance of the models by striking a balance between goodness–of–fit and parsimony (Schwarz, 1978). Along with the BICs, the $Pr$ values show the accompanying Schwartz weights, i.e., the posterior probabilities for both models (given equal prior probabilities for the models, Raftery, 1995; Wagenmakers & Farrell, 2004). The finding that RT depends on both $P^{Acc}$ and the direction of change is in line with the hysteresis hypothesis. Due to the binary nature of the response variable, the results for accuracy (Figure 7.7) are less clear. Our statistical analysis, reported later, is based on multivariate hidden Markov models that take mean RTs and proportion correct into account simultaneously. This analysis strongly supports the hysteresis hypothesis. However, we first describe Experiment 1b, which differs from Experiment 1a only by minor methodological improvements.

Method Experiment 1b

Experiment 1b was conducted in order to replicate the results of Experiment 1a. We only describe the method of Experiment 1b insofar as it differs from Experiment 1a.

Participants

Eight first year psychology students participated for course credit. None of these students participated in either Experiment 1a or Experiment 2.
Materials and procedure

We used two different tasks: one lexical decision task, as in Experiment 1a, and one perceptual task. In this perceptual task, participants were asked to judge, by pressing the appropriate button, whether a horizontal line crossing a vertical reference line extended more to the right or to the left. When the line extended more to the right, the distance to the right was about 1 millimeter larger than the distance to the left and vice versa. Participants were either presented the lexical decision task (participants D–G) or the visual perception task (H–K).

In this experiment, we used Presentation® (Version 9.90, www.neurobs.com) to present the stimulus and register the responses. Two response buttons attached to the parallel port were used to maximize timing accuracy. The response stimulus interval varied randomly between 1000 and 3500 ms, to prevent anticipatory responses. To further discourage this behavior, responses before stimulus onset resulted in a warning message.

For both tasks in Experiment 1b, the payoff structure was equivalent to the one used in Experiment 1a. The only difference was that in Experiment 1b RTSCM and RTGM were updated during the experiment. When it was established that a participant was performing accurately, evidenced by 5 consecutive correct responses, these 5 responses were used to update RTSCM. Likewise, when it was established that the participant was guessing, evidenced by 5 consecutive responses faster than RTGM + 50 ms, these 5 responses were used to update RTGM.

An improvement to Experiment 1a was that we added a bar in the top portion of the screen (permanently visible for the participant), that visually displayed the current value of PAcct and PAcct respectively. The portion of this bar that was blue/orange represented the amount of pressure on accuracy/speed (see Figure 7.5). This bar ensured that the participant was aware of the current payoff setting at any time during the experiment.

Participants were tested in two sessions; the first one lasted two hours and the second one lasted one hour. Participants were allowed a short break every 20 minutes.

Results Experiment 1b

As in Experiment 1a, we discarded the first series for each participant. Participants D, E, F, and G (lexical decision task) contributed respectively 1021, 569, 1137, and 449 trials, in which they completed 39, 19, 32 and 10 series. Participants H, I, J and K (visual task) contributed respectively 589, 920, 860 and 481 trials, in which they completed 16, 20, 26 and 13 series.

Descriptive results

Again, for each participant, the mean of RT was calculated at each value of PAcc, for both directions of change in PAcc separately, as shown in Figure 7.6 (participants D–G) and Figure 7.8 (participants H–K). Again, as expected, high values of PAcc invoke stimulus controlled behavior whereas low values of PAcc invoke guessing behavior. More notable, in both the lexical decision version and the perceptual version, we again found hysteresis effects on RT, indicated by higher mean RT at intermediate values of PAcc when decreasing PAcc than when increasing PAcc. Again, visual inspection of response accuracy (Figure 7.7 and 7.8) does not allow one to draw strong conclusions, and this may be due to the aforementioned binary character of the response variable. However, naturally, accuracy is low when PAcc is low and high when PAcc is high.
In sum, visual inspection of changes in mean RT and consideration of the difference in model fit between the 1–factor and 2–factor model both suggest that hysteresis is present for about half of the participants—for the same payoff settings, these participants were faster coming out of the guess mode that they were coming out of the stimulus controlled mode.

The visual inspection of the descriptive results has two important drawbacks. First, averaging over series might have masked sudden shifts in behavior. Second, the univariate descriptives above do not reflect the multivariate character of RT data. In order to address these issues and provide a more formal test of the hysteresis hypothesis we now turn to an analysis using hidden Markov models.

Figure 7.8: Experiment 1b (lexical decision): Mean RT increases with $P_{Acc}$. For participants F and G, responses are slower at intermediate values of $P_{Acc}$ when $P_{Acc}$ decreases (and participants are speeding up) than when $P_{Acc}$ increases (and participants are slowing down). The BIC’s allow for comparison between a linear model in which RT is regressed on both $P_{Acc}$ and direction of change (2–factor), and the model with only $P_{Acc}$ as predictor (1–factor). Lower BIC’s indicate the better model. Also, the posterior probabilities (Schwarz weights) for both models are reported.

**Hidden Markov Analyses**

The SAT phase transition model assumes that, as participants switch from one processing mode to the other, RT and accuracy undergo sudden jumps. This implies that RT and accuracy can be considered a multivariate time series that follows a two state mixture
Figure 7.9: Experiment 1b (lexical decision): Mean accuracy increases with $P_{Acc}$. With sparse data, visual inspection of the binary variable accuracy does not allow one to draw clear conclusions about the presence or absence of hysteresis.

distribution. Such data can be analyzed using hidden Markov models (HMMs, e.g., Böckenholdt, 2005; Vermunt, Langeheine, & Böckenholdt, 1999; Visser, Raijmakers, & Van der Maas, 2009; Wickens, 1982); HMMs allow one to learn about a number of latent states that cannot be observed directly. Additional parameters describe the transition dynamics of these unobserved states and their connection to the observed behavioral variables.

Thus, a hidden Markov model consists of two main parts: the measurement model and the transition dynamics (Visser et al., 2009). The measurement model defines latent states in terms of the observed variables. In our case, the measurement model defines the states (guess mode and stimulus controlled mode) in terms of RT and accuracy. The transition dynamics are defined by the transition probabilities, i.e., the probabilities of switching from one state (in our case GM) to the other (SCM) and vice versa.

**HMM for the phase transition model**

The phase transition model posits two modes of behavior that can be captured by a two state hidden Markov model (see Figure 7.12). The two states differ in mean and variance of RT and in proportion correct. In our analysis, one state (GM) has a mean accuracy of 0.5 and relatively short RTs and one state (SCM) has a high mean accuracy and longer and more variable RTs.

Furthermore, the phase transition model predicts that the probability to switch from GM to SCM ($\pi_{GS}$ in Figure 7.12) increases when $P_{Acc}$ increases and that the probability
Figure 7.10: Experiment 1b (perceptual task): Mean RT increases with $P^{\text{Acc}}$. When $P^{\text{Acc}}$ decreases (speeding), participant I is slower at intermediate values of $P^{\text{Acc}}$ than when $P^{\text{Acc}}$ increases (slowing down). The BIC’s allow for comparison between a linear model in which RT is regressed on both $P^{\text{Acc}}$ and direction of change (2–factor), and the model with only $P^{\text{Acc}}$ as predictor (1–factor). Lower BIC’s indicate the better model. Also, the posterior probabilities (Schwarz weights) for both models are reported.

to switch from SCM to GM ($\pi_{SG}$ in Figure 7.12) increases when $P^{\text{Acc}}$ is lowered. We incorporate this prediction by regressing the transition probabilities on $P^{\text{Acc}}$ via a logit link.

Finally, the hysteresis hypothesis posits that the switching probabilities depend on the direction of change of $P^{\text{Acc}}$. This asymmetry is expressed in the difference between the intercepts of the regression function that describes the probability to switch from GM to SCM and the intercepts of the regression function that describes the probability to switch from SCM to GM (see Figure 7.13).

**HMMs for competitor models**

The model described above represents the phase transition model. We tested this model against three competitor models that contrast with various predictions of the phase transition model.

**Model (1)** The first competitor model is a model that assumes that there is only one (latent) state and that the behavior in this state is completely independent from the
c covariate $P_{\text{Acc}}$. This model serves as a reference model, and contrasts the phase transition model’s prediction of the existence of two qualitatively different modes of behavior.

**Model (2)** The second competitor model also comprises a single state, but here, the payoff for speed and accuracy $P_{\text{Acc}}$ is included as a linear predictor of both RT and accuracy. This model can be seen as a representation of the predictions of sequential sampling models in which there is a continuous trade-off between speed and accuracy. This continuous trade-off contrasts the phase transition model’s prediction of a discrete trade-off.

**Model (3)** The third competitor model comprises two states and $P_{\text{Acc}}$ is modeled to affect the transition probabilities. This model represents the predictions of the fast guess model and thus comprises symmetric transition dynamics between the two states. This symmetry contrasts the phase transition model’s prediction of hysteresis that implies asymmetric transition dynamics.

The three competitor models described above are from now on referred to as model 1, 2, and 3, and the phase transition model is referred to as model 4. To fit the models, we used the R-package `depmixS4` ([Visser & Speekenbrink, 2010](#)). This package allows one to fit HMMs to time series of multiple variables with different distributions (using maximum likelihood). In our case, these differently distributed variables are RT and accuracy.
7.5. Experiment 1: Hysteresis

Figure 7.12: Graphical representation of a two-state hidden Markov model. Both states (circles) are defined by two observed measures (squares). The transition probabilities are represented by $\pi_{GS}$ (from GM to SCM) and $\pi_{SG}$ (from SCM to GM), respectively. These $\pi_{GS}$ and $\pi_{SG}$ are regressed on $P_{Acc}$ via a logit transformation, as is captured in the regression functions below. In the hysteresis model, $\beta_{1}^{SG} = -\beta_{1}^{GS}$. The difference between intercepts $\beta_{0}^{GS}$ and $\beta_{0}^{SG}$ quantifies the hysteresis effect.

\[
\ln \left( \frac{\pi_{GS}}{1 - \pi_{GS}} \right) = \beta_{0}^{GS} + \beta_{1}^{GS} \times P_{Acc}
\]
\[
\ln \left( \frac{\pi_{SG}}{1 - \pi_{SG}} \right) = \beta_{0}^{SG} + \beta_{1}^{SG} \times P_{Acc}
\]

Figure 7.13: The logit link function is used to regress the probability of switching from SCM to GM ($\pi_{SG}$) and from GM to SCM ($\pi_{GS}$) on the payoff for accuracy $P_{Acc}$ (Figure 7.12). Note that we plot $Pr$ (stay in SCM) here, which is equivalent to $1 - Pr$ (switch from SCM to GM).

We chose to model RT by a log normal distribution, so we could estimate a mean and variance of log RT in each state. Accuracy was modeled as binomial, so, for each state, the binomial parameter was to be estimated. Furthermore, the depmixS4
package allows us to put various constraints on the parameters of the models and to include a covariate on the transition probabilities. In the formalization of the phase transition model, we will include $P^{Acc}$ as a covariate on the transition probabilities.

**Hidden Markov results Experiment 1a and 1b**

Table 7.1 shows the BICs and accompanying Schwarz weights (column $Pr$) for the different models fitted to each participant’s data. For 9 out of 11 participants, the best fitting model is model 4, the model that represents our hypothesis of a two state system exhibiting hysteresis. For 8 of those, the odds are strongly in favor of the hysteresis model. For participant G, the odds only weakly favor the hysteresis model. For participants E and H, the model that represents the fast guess model fits best, although the odds are not convincing.
Table 7.1: For each participant, the model that postulates two states and hysteresis outperforms the competitor models, as evidenced by BIC values. Columns Pr show the associated posterior probabilities (assuming the models are equally likely a priori). 1s and 2s stand for one state and two states respectively. Experiment code 1bL indicates lexical decision version of 1b, 1bV indicates visual version. Numbers in boldface are the lowest BIC values per subject.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Participant</th>
<th>Model 1: 1s</th>
<th>Model 2: 1s w/covariate</th>
<th>Model 3: 2s fast guess</th>
<th>Model 4: 2s hysteresis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>A</td>
<td>970.12</td>
<td>749.51</td>
<td>490.15</td>
<td><strong>442.41</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1a</td>
<td>B</td>
<td>995.28</td>
<td>708.14</td>
<td>389.33</td>
<td><strong>358.10</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1a</td>
<td>C</td>
<td>692.19</td>
<td>526.68</td>
<td>415.66</td>
<td><strong>376.39</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1bL</td>
<td>D</td>
<td>2738.43</td>
<td>1929.94</td>
<td>1063.29</td>
<td><strong>982.65</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1bL</td>
<td>E</td>
<td>1263.54</td>
<td>810.07</td>
<td>701.81</td>
<td><strong>702.44</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr = 0.422</strong></td>
</tr>
<tr>
<td>1bL</td>
<td>F</td>
<td>2468.40</td>
<td>1507.48</td>
<td>1066.25</td>
<td><strong>1054.70</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr = 0.997</strong></td>
</tr>
<tr>
<td>1bL</td>
<td>G</td>
<td>972.91</td>
<td>754.51</td>
<td>656.46</td>
<td><strong>656.18</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr = 0.535</strong></td>
</tr>
<tr>
<td>1bV</td>
<td>H</td>
<td>1574.63</td>
<td>970.32</td>
<td><strong>931.90</strong></td>
<td>933.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr = 0.332</strong></td>
</tr>
<tr>
<td>1bV</td>
<td>I</td>
<td>2468.01</td>
<td>1897.47</td>
<td>1692.40</td>
<td><strong>1650.41</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1bV</td>
<td>J</td>
<td>2042.91</td>
<td>1314.84</td>
<td>1117.39</td>
<td><strong>1093.78</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr &gt; .999</strong></td>
</tr>
<tr>
<td>1bV</td>
<td>K</td>
<td>1269.36</td>
<td>817.67</td>
<td>760.30</td>
<td><strong>754.50</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Pr = 0.948</strong></td>
</tr>
</tbody>
</table>
### Table 7.2: Response parameters (Mean of RT, SD of RT and proportion correct) estimated for the hysteresis model (Model 4).

State 1 is fast with accuracy at chance level. State 2 is slower and accuracy is relatively high. Spread of RT is larger in the accurate state. The rightmost columns display the parameters of the link functions, estimated for the hysteresis model (for participants E and H, for the fast guess model with $\beta_{1SG} = \beta_{1GS}$). For all participants, the intercept is smaller for the function linking $P^{Acc}$ to the probabilities to switch to SCM (a1) than the intercept for linking $P^{Acc}$ to the probability to switch to GM (a2). $\beta_{1}$ was constrained to be equal in the regression functions on both $Y_{SG}$ and $Y_{GS}$ (see Figure 7.12).

<table>
<thead>
<tr>
<th>Participant</th>
<th>RT state 1</th>
<th>SD RT state 1</th>
<th>PC state 1</th>
<th>RT state 2</th>
<th>SD RT state 2</th>
<th>PC state 2</th>
<th>$\beta_{1}$</th>
<th>$\beta_{1GS}$</th>
<th>$\beta_{1SG}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>250.09</td>
<td>45.33</td>
<td>0.50</td>
<td>617.66</td>
<td>148.68</td>
<td>0.91</td>
<td>6.48</td>
<td>1.39</td>
<td>4.35</td>
</tr>
<tr>
<td>B</td>
<td>242.03</td>
<td>28.23</td>
<td>0.50</td>
<td>542.81</td>
<td>112.92</td>
<td>0.90</td>
<td>4.95</td>
<td>1.43</td>
<td>3.66</td>
</tr>
<tr>
<td>C</td>
<td>262.65</td>
<td>55.06</td>
<td>0.50</td>
<td>570.85</td>
<td>119.29</td>
<td>0.89</td>
<td>4.55</td>
<td>1.02</td>
<td>3.96</td>
</tr>
<tr>
<td>D</td>
<td>192.81</td>
<td>32.84</td>
<td>0.50</td>
<td>500.76</td>
<td>93.33</td>
<td>0.91</td>
<td>5.14</td>
<td>-0.15</td>
<td>2.05</td>
</tr>
<tr>
<td>E</td>
<td>258.65</td>
<td>51.73</td>
<td>0.50</td>
<td>508.42</td>
<td>98.95</td>
<td>0.90</td>
<td>7.15</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>258.54</td>
<td>51.33</td>
<td>0.50</td>
<td>504.38</td>
<td>96.52</td>
<td>0.89</td>
<td>9.30</td>
<td>0.17</td>
<td>1.60</td>
</tr>
<tr>
<td>G</td>
<td>253.37</td>
<td>33.78</td>
<td>0.50</td>
<td>486.00</td>
<td>118.87</td>
<td>0.75</td>
<td>3.74</td>
<td>0.29</td>
<td>1.23</td>
</tr>
<tr>
<td>H</td>
<td>203.14</td>
<td>27.27</td>
<td>0.50</td>
<td>541.60</td>
<td>168.31</td>
<td>0.78</td>
<td>11.42</td>
<td>7.44</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>215.67</td>
<td>37.48</td>
<td>0.50</td>
<td>521.09</td>
<td>154.82</td>
<td>0.75</td>
<td>3.03</td>
<td>-0.20</td>
<td>1.64</td>
</tr>
<tr>
<td>J</td>
<td>205.32</td>
<td>30.32</td>
<td>0.50</td>
<td>455.21</td>
<td>103.17</td>
<td>0.76</td>
<td>5.46</td>
<td>1.22</td>
<td>3.08</td>
</tr>
<tr>
<td>K</td>
<td>196.48</td>
<td>25.77</td>
<td>0.50</td>
<td>476.44</td>
<td>137.56</td>
<td>0.81</td>
<td>3.61</td>
<td>-0.06</td>
<td>1.21</td>
</tr>
</tbody>
</table>
For each participant, Table 7.2 shows the response parameters and the transition parameters for the hysteresis model 4 (model 3 for participants E and H). The response parameters are mean RT and accuracy for both the guess and the stimulus controlled mode. The response parameters show that for each participant, the modes are clearly separated in terms of mean RT. It is also clear that the spread of RT is larger in the stimulus controlled mode than in the guess mode. Accuracy in the stimulus controlled mode is relatively high, but not perfect, suggesting that participants were able to trade off accuracy for speed within the stimulus controlled mode, at least to some extent.

The three rightmost columns of Table 7.2 show the parameters of the function that links $P^{Acc}$ to the transition probabilities. The values of $\beta_1$ confirm that for all participants, increasing $P^{Acc}$ leads to an increased probability to switch toward the stimulus controlled mode. The difference between intercepts $\beta^{GS}_0$ and $\beta^{SG}_0$ quantifies the hysteresis effect (for all participants but E and H).

### Cusp model results

As a final analysis, we used the R–package CUSP (Grasman, Van der Maas, & Wagenmakers, 2009) to fit the data of each participant to Cobbs (1981) stochastic cusp equation. In this model, the normal and splitting variable $\alpha$ and $\beta$ are modeled as linear functions of the experimental variable $P^{Acc}$. So, for each axis, an intercept parameter (i.e., $\alpha_0$ and $\beta_0$) and a slope parameter (i.e., $\alpha_{P^{Acc}}$ and $\beta_{P^{Acc}}$) are estimated. The behavioral variable $Z$ is modeled as a linear function of $\log(RT)$ (with parameters $Z_0$ and $Z_{P^{Acc}}$). All parameters of the fitted cusp model can be found in Table 7.3. For all participants, the cusp model fitted better than a linear model according to BIC model selection, which indicates that the cusp model gives a proper description of the data. For participants B, D, E, F, and K, $\beta_{P^{Acc}}$ could be constrained to zero, which indicates that for these participants the experimental variable $P^{Acc}$ only related to the normal axis, which is predicted by the phase transition model.

Figure 7.14 shows the best fitting model for each participant. The plotted symbols show how the participant’s behavior at different settings of $P^{Acc}$ maps onto the $\alpha$–$\beta$ plain. The shaded area is the bifurcation set, i.e., the area where two stable behaviors exist. The phase transition model predicts that a substantial part of the behavior falls in this bifurcation set, which is the case for most of the participants.

### 7.6 Interim Conclusion

Experiments 1a and 1b were designed to detect hysteresis. When we changed payoff factors gradually, hysteresis was observed in the behavior of about half of our participants. Hidden Markov analyses confirmed that two modes of responding exist and that the phase transitions between these states displayed hysteresis for the majority of the participants. The fit of a stochastic cusp model provided converging evidence that the SAT can be described as a cusp catastrophe. Still, there are two important reasons to doubt that the SAT can be conceptualized generally as a cusp catastrophe. First, the hysteresis effect could be artificially enlarged by the speed of change of the payoff factors. Although we tried to make the change in $P^{Acc}$ gradual, it could still be the case that hysteresis occurs as an effect of a participant’s delayed awareness of change in $P^{Acc}$. To test this

---

7In the fitting routine, the proportion correct of one mode (the guess mode) was fixed at .5. Preliminary analyses showed that this restriction improved any 2-state model’s BIC performance. Therefore, in Table 7.2, the accuracy parameter in the guess mode is .5 for each participant.
Table 7.3: Parameter estimates (intercept and slope) of the linear functions that relate $P_{Acc}$ to the $\alpha$ and $\beta$ axes of the cusp (columns one through four) for participants A to K. The sixth and seventh column display the parameters of the linear function that relates the cusp surface ($Z$) to $\log(RT)$. In all cases the cusp model gave the best explanation of the data. The phase transition model predicts positive coefficients $\alpha_{P_{Acc}}$ and zero values for $\beta_{P_{Acc}}$, since $P_{Acc}$ in our model is associated with the normal axis $\alpha$. For participants B, D, E, F, and K, $\beta_{P_{Acc}}$ could be constrained to zero.

<table>
<thead>
<tr>
<th>Participant</th>
<th>$a_0$</th>
<th>$a_{P_{Acc}}$</th>
<th>$b_0$</th>
<th>$b_{P_{Acc}}$</th>
<th>$Z_0$</th>
<th>$Z_{P_{Acc}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1.42</td>
<td>0.18</td>
<td>2.27</td>
<td>-0.16</td>
<td>-14.76</td>
<td>2.43</td>
</tr>
<tr>
<td>B</td>
<td>-1.17</td>
<td>0.18</td>
<td>1.84</td>
<td>0.00</td>
<td>-20.11</td>
<td>3.42</td>
</tr>
<tr>
<td>C</td>
<td>-1.07</td>
<td>0.19</td>
<td>1.48</td>
<td>-0.08</td>
<td>-16.61</td>
<td>2.78</td>
</tr>
<tr>
<td>D</td>
<td>-2.27</td>
<td>0.22</td>
<td>2.02</td>
<td>0.00</td>
<td>-16.54</td>
<td>2.88</td>
</tr>
<tr>
<td>E</td>
<td>-2.14</td>
<td>0.22</td>
<td>0.84</td>
<td>0.00</td>
<td>-17.90</td>
<td>3.05</td>
</tr>
<tr>
<td>F</td>
<td>-2.26</td>
<td>0.21</td>
<td>0.94</td>
<td>0.00</td>
<td>-18.34</td>
<td>3.13</td>
</tr>
<tr>
<td>G</td>
<td>-1.71</td>
<td>0.13</td>
<td>1.71</td>
<td>-0.08</td>
<td>-18.22</td>
<td>3.05</td>
</tr>
<tr>
<td>H</td>
<td>-3.49</td>
<td>0.40</td>
<td>1.27</td>
<td>-0.40</td>
<td>-11.36</td>
<td>1.85</td>
</tr>
<tr>
<td>I</td>
<td>-1.47</td>
<td>0.12</td>
<td>1.68</td>
<td>-0.06</td>
<td>-14.14</td>
<td>2.40</td>
</tr>
<tr>
<td>J</td>
<td>-2.42</td>
<td>0.28</td>
<td>1.25</td>
<td>-0.11</td>
<td>-15.40</td>
<td>2.67</td>
</tr>
<tr>
<td>K</td>
<td>-2.54</td>
<td>0.21</td>
<td>1.18</td>
<td>0.00</td>
<td>-16.09</td>
<td>2.79</td>
</tr>
</tbody>
</table>

Figure 7.14: The best fitting cusp model for each participant. The plotted symbols show how the participant’s behavior at all different settings of $P_{Acc}$ maps onto the $\alpha$–$\beta$ plane. The phase transition model predicts that a significant part of the behavior lies in the shaded area, the bifurcation set.

possible artifact, we applied a so-called modified method of limits in Experiment 1c. Second, although hidden Markov analyses showed that a two state mixture describes the
data best, the hysteresis experiments described above do not prove that the behavior is
governed by two distinct modes. (The bimodality in the data could, indeed, be caused
by the experimental manipulations.) To test whether behavior is indeed bimodal, which
is a necessary criterion for the existence of phase transitions, we designed the “75%-task”
in Experiment 3.

7.7 Experiment 1c: Modified Method of Limits

Method Experiment 1c
To test whether the hysteresis effect found in Experiment 1a and 1b is an artifact caused by
delayed awareness of $P_{Acc}$ we applied a modified method of limits. This method is based on
a procedure used by Hock et al. (1993) to study bistability in the perception of apparent
motion. We apply the method by gradually changing $P_{Acc}$ but eventually pausing at
predetermined settings of $P_{Acc}^t$ and $P_{RT}^t$. To understand this procedure, consider a payoff
setting at which optimal behavior requires that a participant switches from his current
mode to another mode. Now, the participant refuses to switch at this payoff setting.
This reluctance to switch can have one of two reasons: first, the participant was not yet
fully aware of the current payoff setting. Second, the current mode of processing is very
stable (hysteresis as predicted by the phase transition model). If the first reason holds,
then waiting for a couple of trials at the same value of $P_{Acc}$ would result in a switch,
because the participant will become aware of the current payoff settings. If the second
reason holds, the participant would stay in the current mode of processing, even after
waiting some trials at that value of $P_{Acc}$. The latter finding would be strong evidence
for hysteresis.

Participants
The method applied in this experiment is demanding in that it requires relatively many
trials. Therefore, we tested only two participants. We chose to test two participants whose
data clearly displayed hysteresis in Experiment 1b. This allows us to determine whether
the hysteresis effect observed in Experiment 1b stands the litmus test of a “modified
method of limits”. We tested participant F from the lexical decision version of Experiment
1b, and participant I from the perceptual version of Experiment 1b.

Materials and procedure
In this experiment, both participants performed the same task they did in Experiment
1b. The difference with Experiments 1a and 1b is the way $P_{Acc}$ changed over trials.
Based on the results of Experiment 1a and 1b, six critical levels of $P_{Acc}$ were chosen
($P_{Acc}^{crit}$), at which most of the jumps between modes appeared to take place. These are
the levels at which it is interesting to examine what happens when $P_{Acc}$ does not change
for a couple of trials. The levels of $P_{Acc}^{crit}$ we chose to examine were 8, 9, 10, 11, 12, and
13.

The experiment consisted of sets of trials. Every set of trials started with a value of
$P_{Acc}^t$ that was clearly favoring either guessing or stimulus controlled responding. Then,
$P_{Acc}^t$ was changed, step by step, toward a predetermined $P_{Acc}^{crit}$ (e.g., $P_{Acc}^{crit} = 10$). When
the sequence of $P_{Acc}$ arrived at $P_{Acc}^{crit}$, it remained at this value for five more trials. These
are the five waiting trials where a delayed switch might or might not happen. After these
five trials, a new set was started at a $P_{Acc}^t$ value that either clearly favored guessing or
7. Phase Transitions in SAT

Figure 7.15: Illustration of the way $P_{Acc}$ changed in the modified methods Experiment 1c. Each dot represents a trial. After a sequence of a few trials with increasing or decreasing $P_{Acc}$, the value of $P_{Acc}$ was kept constant for 6 trials at a $P_{tAcc}$ value of $P_{crit}$.

stimulus controlled responding. As is illustrated in Figure 7.15, $P_{Acc}$ was increased from a low value, that favored fast guessing, upwards to $P_{crit}$ for two sets. Then, $P_{Acc}$ was decreased from a high value, that favored stimulus controlled behavior, downwards to $P_{crit}$ for two sets. Over the entire experiment, the direction of change was alternated in this two-by-two manner. Each set’s value of $P_{crit}$ was chosen semi–randomly from the selected values 8 to 13.

Results Experiment 1c

The above procedure resulted in $6 \times 4 \times 2 = 48$ sequences containing a total of 814 trials per participant. In this procedure, however, the only trials of interest are the waiting trials. We chose to use only the last four waiting trials from each sequence. We did so, because it is reasonable to assume that after two trials with the same $P_{Acc}$ value ($P_{tAcc} = P_{crit}$), the participant is aware of this setting. Thus, the total amount of trials used for calculating the descriptives below are $48 \times 4$ waiting trials.

Figure 7.16 shows the data of both participants. The data of participant F clearly show hysteresis, i.e., both RT and accuracy are higher for decreasing $P_{Acc}$ than for increasing $P_{Acc}$. Again, we compared the fit of a linear model regressing waiting trials’ RT on both $P_{Acc}$ and direction of change ($BIC = 2428.72$, Schwarz weight $Pr = 1.00$), with a model that regressed RT on $P_{Acc}$ only ($BIC = 2443.69$, $Pr = 0.00$). We conducted the same analysis to accuracy data (be it a logistic regression), in which the two–factor model again performed better ($AIC = 212.29$, $Pr = 1.00$) than the one factor–model ($AIC = 498.10$, $Pr = 0.00$). These analyses suggest that both RT and accuracy depend on direction of change, which supports the hysteresis hypothesis. The data of participant I are less clear, yet for all but one critical value of $P_{crit}$, mean RT is higher for decreasing $P_{Acc}$ than for increasing $P_{Acc}$, which is in line with our hypothesis of hysteresis. Also for this participant, the two–factor model ($BIC = 2443.91$, $Pr = 0.79$) outperformed the one–factor model ($BIC = 2446.55$, $Pr = 0.21$) in predicting RT. In the prediction of accuracy also, the two–factor model ($AIC = 267.71$, $Pr = 1.00$) outperformed the one–factor model ($AIC = 498.10$, $Pr = 0.00$). Thus, also the results of participant I support the hysteresis hypothesis.
Figure 7.16: Experiment 1c: Mean RT and accuracy of the “waiting” trials. Participant F’s RT and accuracy are higher when $P_{\text{Acc}}$ was decreased (speeding), than when it was increased (slowing). Participant I’s data are less clear, but RT is usually higher on slowing trials.

Discussion Experiment 1

Hysteresis was present in the data of Experiment 1a and was replicated with the slightly improved method of Experiment 1b. For all participants in Experiments 1a and 1b, the hidden Markov analyses showed that the hysteresis model outperformed the competitor models. Furthermore, hysteresis was still present for one participant when put to the strict test of the modified method of limits, carried out in Experiment 1c. These results favor the phase transition model over the fast guess model that predicts that the jumps from GM to SCM and and the jumps from SCM to GM should take place at the same setting of $P_{\text{Acc}}^t$. However, as argued before, to contrast our model against the pure sequential sampling models, evidence for hysteresis is not enough. When we would also find bimodality in the SAT, this would provide complementary evidence to favor the phase transition model over pure sequential sampling models.

7.8 Experiment 2: Bimodality

The phase transition model predicts bimodality in behavior when the pressure on both speed and accuracy is high (i.e., the area between the two dotted bifurcation lines in Figure [7.14]). Bimodality occurs because the intermediate behavior (e.g., responding at 75% correct) is inaccessible. When a participant nonetheless wants to meet the experimenter’s
demands to respond at 75% correct, this can only be achieved by mixing responses from
the two stable modes, that is, responding accurately on some of the trials and fast guess-
ing on the others. This mixing of strategies yields bimodal distributions of the behavioral
variables. In contrast, according to most sequential sampling models, participants can
simply adjust the bounds of the decision process to reach an accuracy of 75% (see Fig-
ure 7.2), which would lead to a unimodal distribution of the behavioral variables.

We set out to test these diverging predictions in the second experiment by making
participants respond at 75% correct. To evoke 75% correct performance, deadline or
response signal procedures could be applied. However, RTs in these tasks are under
control of the deadline manipulation or response signal and only accuracy is left as a
dependent variable. Accuracy, unfortunately, is a discrete variable and discrete variables
can not be bimodally distributed. For that reason, deadline or response signal procedures
could not be used to study bimodality.

Thus, in the experiment below, we choose not to manipulate speed but instead ma-
nipulate accuracy, allowing RT to be used as the dependent variable. We refer to the
task as “the 75% task”. In the 75% task, participants are instructed to respond at an
accuracy level of 75% correct, and to do so as fast as possible. For comparison, we also
administered a 50% task and a 100% task, in which participants were asked to respond as
fast as possible at 50% correct (guessing) and at 100% correct. Whereas the predictions of
the phase transition model and sequential sampling differ for the 75% task, they agree for
the 50% and 100% tasks. For both tasks, the models predict unimodal RT distributions:
fast and chance–level RTs for the 50% task, slow and almost error–free RTs for the 100%
task.

It is interesting to note, that we could only find a 75% instruction in experiments
of Lappin and Disch (1972). Yet, from a statistical point of view, and assuming that
the speed–accuracy trade-off is continuous, such an instruction would give estimates of
mean RT with much lower standard errors than we get at the more typical 95% correct
target. As Wickelgren (1977) points out, it is precisely at high levels of accuracy where
the variation in RT is very large for very small differences in error percentage.

Method Experiment 2

Participants

Thirteen students at the University of Amsterdam participated for a small monetary
reward.

Materials and procedure

We used the same lexical decision task as used in Experiment 1a. Stimuli were selected
and presented in the same way as in Experiment 1a. Also, the screen refresh rate, response
stimulus intervals, and response button assignment were also identical to those used in
Experiment 1a.

Trials were presented in blocks and sets. A set consisted of a random number of trials
(between 15 and 25), and a block consisted of a number of sets (5 to 10, depending of
the number of trials in each set, such that a block had never more than 150 trials). At
the end of each set, participants received feedback about their performance (i.e., their
penalty score, the ranking of this score on a personal high score list, and a general high
score list). After each set, the penalty score ($PS$) was computed as follows:

$$PS = \frac{|\%\text{correct}_{\text{set}} - \%\text{correct}_{\text{target}}|}{25} + \frac{(RT_{\text{set}} - 100)}{700}$$

Participants were instructed to minimize this penalty score $PS$. Note that, because the participants did not know in advance the amount of trials in each set, their best strategy was to try to maintain a mean accuracy close to 75% correct throughout a set.

Note that the $PS$ is heavily dependent on how close the participant is to the accuracy target (75%) and that speed is of secondary importance. Nevertheless, a small deviation from the optimal percentage correct could be compensated by faster responses. For instance, a 5% deviation from the goal can be compensated with an increase in speed of 140ms. Speed was also included in the $PS$ to prevent that participants used a stimulus controlled strategy and intentionally erred on every fourth trial.

Design

The experiment featured four conditions. In the first condition, the stated target was to obtain 50% correct. In the second condition, the target was to respond at 100% correct. In the third and fourth condition, participants were instructed to respond at 75% correct. The latter two conditions—denoted 75%PT and 75%SS—differed with respect to the instructions. In the 75%PT (phase transition) condition, the instructions were given in terms of the phase transition and fast guess model. The participants were told that optimal performance implied an alternation of guessing and accurate responding. In the 75%SS (sequential sampling) condition, on the other hand, the instruction of the task was given in terms of sequential sampling models. The participants were told that they should respond at such a high speed that accuracy, on average, reaches 75%. In a pilot study the 75% condition was introduced to participants without any specific instruction, but it turned out that most participants then persisted in adapting the highly inefficient strategy of slow responding with intentional errors in one of four cases. RTs associated with this inefficient strategy are equal to, or slower than, those in the 100% condition. It is important to note that, if the sequential sampling models’ prediction of a continuous SAT is correct, instructions in terms of sequential sampling models would yield lower penalty scores than instructions in terms of the phase transition model.

A complete experimental session consisted of a series of blocks. For example, the experiment could start with a block of 5 sets of the 100% condition, then a block with 5 sets of 50%, then a block with 10 sets of 75%SS and finally a block with 10 sets of 75%PT. Each experimental session was comprised of a total of 40 sets × 15 to 25 items ≈ 800 trials, and took about an hour to complete.

There were three groups (A, B, and C) of participants. The experimental session of group A (participants 1 to 5) was organized as follows: 6 sets with $\%\text{correct}_{\text{target}} = 100\%$, 6 sets with $\%\text{correct}_{\text{target}} = 50\%$, 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (PT), 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (SS). The session of group B (participants 6 to 13) was organized as: 6 sets with $\%\text{correct}_{\text{target}} = 100\%$, 6 sets with $\%\text{correct}_{\text{target}} = 50\%$, 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (SS), 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (PT). Two weeks later 6 participants were retested. This group C (participants 1, 7, 8, 9, 11, 13) received 8 sets with $\%\text{correct}_{\text{target}} = 50\%$, 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (SS), 7 sets with $\%\text{correct}_{\text{target}} = 75\%$ (PT). So group A received the phase transition model instructions first, and groups B and C received the sequential sampling model instructions first.
Data analysis

Data were inspected visually for bimodality, analyzed using the distributional RT analysis program of [Dolan, Van der Maas, and Molenaar (2002)], and the mode testing program of [Hartelman, Van der Maas, and Molenaar (1998)]. Using the distributional RT analysis program, mixtures of 1, 2, and 3 Ex–Gaussian components were fitted to RT distributions obtained in each condition. The choice for the Ex–Gaussian distribution is pragmatic since, as explained earlier, the present phase transition model does not make any critical distributional predictions. The BIC was used to determine the number of components. The number of components is however not always equal to the number of modes, since components can have equal means, but differ in variance. Therefore, when the BIC favored the two or three component mixture, and the data and fitted distribution is clearly bimodal (and not trimodal), we concluded that the reaction times in that condition were bimodally distributed. Finally, we confirmed the results from the mixture analysis using a kernel density mode–testing program (e.g., [Hartelman et al., 1998; Silverman, 1981, 1986]).

Results Experiment 2

Penalty scores

In the 75% task, the phase transition model instructions (75%PT) yielded lower penalty scores PS than the sequential sampling model instruction (75%SS).8 Participants in group A (75%PT first) had a mean PS of .86 (SD = .33) in the 75%PT condition and a mean PS of .98 (SD = .39) in the 75%SS condition. For participants in group B and C (75%SS first), these means were .78 (SD = .28) and .96 (SD = .36), respectively. An ANOVA with PS as dependent variable and instruction (phase transition model vs. sequential sampling model) and group (A versus B and C) as independent variables showed that the phase transition model instructions yielded lower penalty scores PS than sequential sampling instructions (F(1) = 16.5, p < .001).

Distributional analyses

Figure 7.17 shows histograms of the RT data of Experiment 2 per experimental manipulation (50%, 100%, 75%PT, and 75%SS) and per participant group (A, B, C). In each group, the data were aggregated over participants. The rightmost column of histograms displays the data when aggregated over all participants in all groups using the ‘V incentizing’ technique described by [Ratcliff, 1979]. Individual participants’ RT distributions can be found on the first author’s website.

As expected, the distributions of RT in the 50% and in 100% condition were clearly unimodal for almost all participants, as is suggested by the group–averages shown in the upper two rows of panels in Figure 7.17. In contrast, visual inspection of the RT distributions of individual participants in the 75%PT and 75%SS condition, suggested bimodality in the data of the majority of participants. This bimodality is also suggested by the group distributions in the lower two rows of Figure 7.17.

These results were checked with the mode testing method. The distributions of about half of the participants in conditions 75%PT and 75%SS were identified as bimodal.

8Although the majority of participants understood the task well, a small minority found it difficult to engage in guessing. These participants found it hard to ignore the primary task of discriminating words from nonwords. After some training, however, all participants were able to perform reasonably well (i.e., attained a low penalty).
For the majority of participants, the mixture analyses also supported a 2-component solution. However, individual-subject analyses were plagued by computational problems due to outliers, sensitivity to starting values, and occasional failures to convergence.

More robust results were obtained by aggregating the data of participants across groups A, B, and C. As can be seen in the rightmost histograms of Figure 7.17, the aggregated RT distributions were unimodal in the 50% and 100% conditions and bimodal in both 75% conditions. These conclusions were confirmed by mode testing using kernel density estimates, using an alpha level of .05. (For details about this analysis, see Hartelman et al., 1998)

We also inspected the results of the 10% lowest penalty scores in conditions 75%PT and 75%SS separately, because sequential sampling models predict that (close to) optimal behavior would involve intermediate, unimodal behavior. Visual inspection showed that the bimodality of the distribution in the condition was preserved when only data were included of sets with the 10% of lowest penalty scores. This was checked with the mode testing method. We found that in both the 75%PT and the 75%SS condition the hypothesis of only one mode was rejected (h-crit=86.5, \( p < .001 \), and h-crit=67.5, \( p < .05 \), respectively).

Finally, we checked whether fast responses were inaccurate and slow responses were accurate. As expected, in the 75% conditions, the probabilities correct of responses of RTs below and above 450 ms were respectively .53 and .82.

Discussion Experiment 2

The data of Experiment 2 generally support the hypothesis that the instruction to respond at intermediate levels of accuracy (75%) lead to bimodality of behavior, regardless of the instructions given. With both instructions, participants managed to respond at 75% correct by alternating between two modes, which are presumably the guess mode and the stimulus controlled mode. Instructing participants to try to reach 75% by adjusting their response criteria did not result in data that are consistent with the continuous SAT predicted by sequential sampling models. Furthermore, the instruction based on the phase transition model lead to lower penalty scores than the instruction based on sequential sampling models. As noted before, this is important, since, when the sequential sampling model account were correct, instructions according to the sequential sampling model would yield the lowest penalty scores. These findings strongly suggest that no intermediate mode of processing is available. Furthermore, both instructions resulted in bimodal RT distributions.

One could argue that grouping data of participants could be the source of spurious bimodality. However, this would only be the case, when some of the participants always guessed and others always responded accurately (which would still evidence the absence of an intermediate mode of processing). Both the Vincentized distributions and the distribution of RT for the 10% best sets suggest that this is not the case. If anything, averaging might have partly masked bimodality, since the locations of the two modes of processing vary over participants.

For some participants we were unable to convincingly demonstrate bimodality. This could be due to a lack of power (the mode testing method is known to be very conservative, e.g., Fisher & Marron, 2001), or to the fact that some participants find it difficult to engage in guessing behavior. We suspect that more training is required for these participants.
7.9 Concluding Remarks

In this paper, we presented a model that departs radically from much current theorizing about response times (e.g., Ratcliff & McKoon, 2008). The phase transition model predicts a sudden collapse in the accuracy of responding when the participant is instructed to speed up, whereas most models predict a continuous trade-off between speed and accuracy.

It should be noted, however, that in sequential sampling models, it is not precisely specified how manipulations of response strategy, such as payoffs and deadlines, relate mathematically to the effective boundary values. This issue is addressed by several models, most of which describe how performance is monitored and optimized (e.g., Bogacz, 2007). One model that describes an autonomous mechanism to adapt response thresholds is Vicker’s self-regulating PAGAN model. This model provides an algorithm that adjusts threshold settings on the basis of discrepancies between experienced and desired response confidence (Vickers & Lee, 1998, Vickers & Lee, 2000, Vickers, 1979). Another model that describes how performance is optimized is the neural network model proposed by Simen, Cohen, and Holmes (2006) in which response thresholds are adjusted to maximize reward rate. One of these or related models could describe the relation between manipulations and threshold setting as a cusp function. One could argue that this implies that our current results are perfectly in line with sequential sampling models.

Our objection to this line of reasoning is twofold. First, although the precise relation between boundary manipulations and the effective boundary setting is rarely specified, it is safe to say that a continuous function is assumed implicitly. A cusp function for boundaries does not really explain the inaccessible region at around 75% and we do not see a conceptual justification for this extension. To the contrary, in the standard explanation of the SAT in sequential sampling models, it is essential that subjects can select any boundary setting.

Second, sequential sampling models, such as the diffusion model, have not been designed to account for guessing. In these models, guessing occurs when boundary separation approaches zero. In that case, no evidence accumulation occurs and the response process reduces to residual processes combined in $T_{er}$. One could argue that $T_{er}$ incorporates the guessing process, so that guessing is just normal decision making without one (important) component. It is unclear, however, how such an account could explain inaccessibility and hysteresis, phenomena that require competition or conflict between guessing and stimulus controlled processing. In two state models, guessing and stimulus controlled processing compete for the same cognitive resources such as attention, motor preparation, and stimulus encoding. This competition results in intermediate behavior that is inherently unstable: one either attempts to answer correctly or one guesses.

Therefore, we believe that the current findings are best explained in terms of a phase transition between two states or processes. Such a two-state explanation should be connected to models that describe how participants select response strategies. Relevant models would be, for example, those of Rieskamp and Otto (2000) and B. R. Newell and Lee (2010) that describe how response strategies may be selected and adjusted according to environmental variables such as pay-off and difficulty.

Yet, the phase transition model is clearly too simple to explain all empirical facts. Figure shows the response time distributions of the 50%, 75% and 100% conditions. The 100% distribution’s peak is shifted to the right compared to the accurate and slow

---

9Other sequential sampling models, of the accumulator type, are more promising in this respect. Guessing could, for instance, be modeled by excitation instead of inhibition of accumulators.
component of the 75% distributions. This is not predicted by the current formulation of the phase transition model. In a sequential sampling framework however, this can be explained naturally by assuming two different criterion settings; one that results in very slow but almost 100% correct responses and one that results in, e.g., 90% correct and somewhat quicker responses. Therefore, it would be of great value to formulate our model by integrating nonlinear sequential sampling models (Heath, 2000; Roe, Busemeyer, & Townsend, 2001; P. L. Smith, 1995; Usher & McClelland, 2001) in the phase transition model. It could be feasible to do so because mathematically, stochastic catastrophe models are very similar to continuous time diffusion models (Cobb & Watson, 1981).

In the light of this integration of both frameworks, it is interesting to observe that the qualitative form of the continuous trade-off (e.g., A2 in Figure 7.1) may well be described as a fold catastrophe (see Figure 7.18).

Considering that a cusp catastrophe consists of two coupled fold catastrophes, as illustrated in Figure 7.18, it seems plausible to specify the upper fold as representing a sequential sampling process and the lower fold as representing a model of simple response time. Since simple response times can also be modeled as a sequential sampling process (Luce, 1986), a complete phase transition model, that includes a sequential sampling explanation for both the stimulus controlled mode and the guess mode, seems within reach.
Figure 7.17: Experiment 2: For each group and each condition, the data were aggregated over participants. The condition’s accuracy targets (% correct target) are shown in the left margin of the figure. The affixes PT and SS stand for sequential sampling and phase transition model instructions, respectively. The histograms of these aggregated sets are displayed in the left three columns. Strong evidence for bimodality is found in both versions of the 75% task. The data of all participants (from all groups) were then aggregated using the ‘Vincentizing’ method described by Ratcliff (1979). The resulting histograms are plotted in the rightmost column.
Figure 7.18: The cusp catastrophe consists of two coupled fold catastrophes. This figure shows that the upper fold 1 could represent the SAT according to sequential sampling models and that the lower fold 2 could represent a model for simple RT. In this figure, the y-axis represents accuracy, but the same pattern holds for RT.
Summary and Discussion

This thesis documents several rather diverse explorations in the small world of speeded two-choice decision making. Despite their diversity, all explorations share two common elements: first, that quantitative modeling is an essential tool to understand behavioral data, and, second, that behavior is adaptive and flexible, even in elementary cognitive–perceptual tasks. Before taking some distance to place the results in a broader perspective, I first summarize and discuss the main conclusions of this thesis.

8.1 Summary of Results

Practice Effects

Extended practice on the same task almost invariably results in a speed–up of performance. These practice effects have been studied for many years (Woodworth & Schlosberg, 1954; Newell & Rosenbloom, 1981; Logan, 1992). Long before experimental psychologists became interested in how information processing of information becomes more efficient with practice, factory owners must have been interested in how their workers improved on the tasks they were hired to do. These factory owners did probably not only care about the average speed of their workers, but were also concerned with their accuracy and steadiness of performance. From this perspective it appears surprising that the field of research on practice effects in psychology has mostly focused on mean response time and has largely ignored accuracy and the spread of response time (from now on: RT). Chapters 2 and 4 of this thesis represent efforts to paint the full picture of practice effects and to account not just for the speed–up in mean RT, but also for changes in accuracy and the effects on the entire distribution of RT.

In the first of the two chapters, each of four participants practiced on the same 400–trial lexical decision task for 25 times. The data of this experiment show a complicated pattern of results: Participants instructed to respond accurately showed stable accuracy throughout the 25 practice blocks. However, the effects on RT were dramatic: not just mean RT decreased strongly, but the entire distribution became more and more peaked. For participants instructed to respond fast, accuracy increased strongly over practice, while the distribution of RT was relatively stable. A diffusion model analysis allowed us to disentangle the effects on both dependent variables. We found that participants' ability to discriminate the stimuli improved both under speed stress and accuracy stress. In addition, we found that participants instructed to respond accurately became less
cautious with practice. In contrast, participants instructed to respond fast did not become less cautious, probably because they were forced to take risks right from the start of practice. Furthermore, we found that non–decision time, a measure for the duration of peripheral and motor processes, strongly decreased for participants who were instructed to respond accurately. These findings strongly suggest that practice is a multifaceted phenomenon that can not be usefully abstracted in terms of mean RT alone.

The most unexpected finding of this large practice study was the strong practice effect on non–decision time. In particular, we were left with the question whether this effect was due to learning on the general task of lexical decision making, or that the effect was caused by an increase in familiarity with the stimuli. To discriminate between these two explanations, we designed the second practice experiment reported in chapter 3. In this experiment, participants practiced on stimulus lists that were unique and with a stimulus list that was presented repeatedly. In this design, stimulus–specific effects should only be found on the repeatedly presented stimulus lists, whereas task–related effects should be evenly expressed in uniquely presented and repeatedly presented stimulus lists. The results of this study show that the practice effects on both non–decision time and rate of information processing are partly task–related and partly stimulus–specific. The practice effects on response bias and caution—the components of processing that are under control of the participant—appeared to be task–related.

The studies in chapters 2 and 3 show that practice effects are clearly more than just a speed–up of mean RT. Several processes underly the practice effect, some of which are task–related and some of which are stimulus–specific. One of the most notable findings is that gradual changes in participants’ ability were accompanied by gradual changes in response caution. In this way, participants were able to maintain a relatively stable level of accuracy while their ability on the task increased.

**Effect of Errors**

When participants err in an RT task, they tend to slow down on the subsequent trial. The most common explanation of this post–error slowing (PES) phenomenon is that participants continually monitor their performance and interpret errors as a sign that the chosen response threshold was too liberal (Rabbitt & Rodgers, 1977; Cohen et al., 2000). Consequently, participants heighten their threshold following an error in order to increase the probability of a correct response on the next trial. The heightened threshold leads to fewer errors but also causes slower responding (i.e., the PES phenomenon). This explanation of PES is so appealing that it is often just assumed to be correct. Consequently, the magnitude of PES is interpreted as a direct measure of cognitive control. Conclusions about cognitive control are then based on correlations between the amount of PES and physiological measures (Li et al., 2006; Danielmeier et al., 2011) or differences in PES between clinical groups (Shiels & Hawk, 2010). Although the explanation of PES in terms of cognitive control is very appealing, there exist several different explanations of PES. In chapter 4 we show that each of the existing explanations of PES maps onto one of the diffusion model parameters. Thus, the application of the diffusion model allows us to decide between existing theories of PES.

In chapter 4, we set out to test competing theories of PES in a very large lexical decision study. This data set, that comprised of over 28,000 data points for each of 39 participants, contained enough error trials (and consequently enough post–error trials) to fit the full diffusion model to each participant’s data set individually. The results showed that the PES effect in this data set was caused by an increase in response caution after committing an error, as predicted by the cognitive control explanation.
In chapter 5, we focused on the difference in PES between young and elderly participants. Elderly participants, who are known to respond more cautiously than young participants in general (Salthouse, 1978; Strayer et al., 1987; Ratcliff et al., 2010, 2006b), are also known to have coarser control of their speed–accuracy trade–off, presumably leading to stronger PES (Rabbitt, 1979; G. A. Smith & Brewer, 1995; Band & Kok, 2000). To obtain a better understanding of this age–related difference in PES, we analyzed the data of an experiment in which young and elderly participants responded to random dot motion stimuli. The results show that both young and elderly participants wasted time on irrelevant processes after committing an error. Apparently, participants took time to think over the error they committed or wasted some time in frustration. Furthermore, elderly participants processed information more slowly and responded more cautiously after committing an error.

The results of chapter 5 do not completely agree with those of chapter 4. Part of this discrepancy may lie in the different tasks that were used in both experiments. In a lexical decision task, as was used in chapter 4, participants often realize that they have committed an error after pressing the response button. In this situation, it makes sense to increase response thresholds, since waiting longer to respond would have led to the correct response. In contrast, when performing a random dot motion task, the only cue that tells the participants that they committed an error is the feedback message. Participants might therefore reason that they would have committed an error, regardless of their threshold setting. Similar reasoning might explain the additional effect on non–decision time that we found in chapter 5: When participants get an error message when they have no clue that their response was wrong, they might get much more frustrated than when they know what they did wrong.

The results from both chapter 4 and chapter 5 underscore that participants’ behavior is far from static. Participants’ behavior changes constantly. Behavior is not just changing due to strategic adjustments, performance is changing as well due to processes that are not under control of the participant, such as changes in the rate of information processing.

Chapter 6 describes a methodological obstacle that we found when we analyzed the PES effect. The obstacle is inherent to the traditional method to quantify PES. This traditional method subtracts the mean RT following correct responses from the mean RT following error responses. We show that this measure of PES can be confounded by global fluctuations in performance over the course of a task. Fluctuations in participants’ motivation may lead to spurious PES, whereas fluctuations in response caution may mask PES or even lead to spurious post–error speeding. The simple solution is to compare post–error trials to pre–error trials, ensuring that both parts of the comparison originate from the same locations in the data. Simulations show that our method is indeed robust to global fluctuations. An empirical data set shows that the confound can be found in real data.

From Accurate Responding to Guessing: A Phase Transition

In most sequential sampling models of RT, participants are assumed to continuously trade off speed for accuracy: Continuously increasing the pressure on accuracy leads to continuously increasing response thresholds; continuously increasing pressure on speed leads to continuously decreasing response thresholds. In chapter 6, we set out to test whether the speed–accuracy trade–off is indeed continuous and we contrast this assumption to predictions that are derived from a new phase transition model for RT. This phase transition model assumes that behavior is governed by two modes of processing: one accurate, stimulus controlled mode, and a fast guessing mode. Two–state systems,
like the one postulated by our phase transition model, can be modeled using catastrophe theory. Catastrophe theory is a mathematical theory that applies to dynamic systems in which continuous changes of environmental variables lead to sudden changes in observed behavior (e.g., Zeeman, 1976). From catastrophe theory, we derived two signature predictions of the phase transition model: bimodality and hysteresis. We tested these two predictions in two experiments.

The first hypothesis, bimodality, states that two irreconcilable modes of processing exist. In our phase transition model, this means that participants can either process the stimuli and respond relatively accurately (the stimulus controlled mode) or guess quickly, without processing the stimulus (the guess mode). Intermediate behavior is unstable. In the experiment that was set out to test this prediction of bimodality, participants were rewarded for targeting 75% correct (intermediate accuracy) while maintaining speed. The results showed that participants were not able to show intermediate behavior, but mixed fast guesses with accurate responses, yielding a bimodal RT distribution.

The second hypothesis, hysteresis, states that both modes of processing are inherently stable. As a result, when participants are performing in the stimulus controlled mode, and rewards pressure them to continuously speed up at the cost of accuracy, they will be able to maintain relatively high accuracy until, at a certain point, performance collapses to guessing behavior. On the other hand, when participants are in the guessing mode and rewards pressure them to respond more and more accurately at the cost of speed, they will resist to shift to the accurate stimulus controlled mode until a relatively high reward for accuracy is offered. Thus, the relative reward for accuracy at which this shift to the stimulus controlled mode takes place is higher than the reward for accuracy at which the shift from the stimulus controlled mode takes place. This hypothesis of hysteresis was tested in an experiment in which we gradually changed the relative payoff for speed and accuracy, from 100% payoff for speed (when participants should guess) to 100% payoff for accuracy (when participants should respond correctly). Hysteresis was found for all but two of our 11 participants.

These experiments show that the speed–accuracy trade–off is not continuous, as predicted by most sequential sampling models. In contrast, performance over the entire range of the speed–accuracy trade–off appears to be governed by two separate modes of processing. Our model accounts for the phase transitions between those two modes of behavior. Yet the phase transition model is clearly too simple to explain all empirical facts. Many studies have showed that people are able to trade off speed for accuracy continuously within a certain range. Thus, the stimulus controlled state should comprise a continuum of behavior in the higher regions of accuracy. In a sequential sampling framework, this can be implemented naturally by assuming a continuously changing response criterion. Therefore, it would be of great value to formulate our model by integrating (nonlinear) sequential sampling models (Heath, 2001; Roe et al., 2001; P. L. Smith, 1995; Usher & McClelland, 2001) in the phase transition model.

8.2 Discussion

Critical Tests of the Diffusion Model

In four chapters of this dissertation, the diffusion model is applied to decompose effects on RT and accuracy into psychological processes that are assumed to underly behavior. The value of such a decomposition of behavior relies on (1) the validity of the interpretation of model parameters (2) the reliability of the parameter estimates and (3) the robustness
to misspecification of the model. The model’s performance in the light of each of these requirements is discussed below.

Validity

The validity of the interpretation of the diffusion model parameters critically depends on the specificity of the mapping between the model parameters and the assumed underlying psychological processes. Voss et al. (2004) show that this parameter–to–process mapping is indeed very specific: speed–accuracy manipulations affected boundary separation $a$, stimulus difficulty affected drift rate $v$, the ease of executing the motor response affected non–decision time $T_{cr}$ and the relative reward for the response options affected starting point $z$. Ratcliff (2002) performed a similar successful test on the specificity of influence on boundary separation and drift rate. In the same study, Ratcliff (2002) showed that, although the diffusion model was able to fit many patterns in real data, it did not fit plausible but fake data. The latter result shows that the model is not over–parameterized and cannot just fit any data set.

Reliability

In order to fit the data to an experimental condition and to reliably estimate the model parameters, one needs many data points. More specifically, one needs enough error responses to fit the error RT distribution. When there are not enough data, the parameters of the diffusion model are estimated poorly, and this is particularly true for the variability parameters $\eta$, $s_z$, and $s_t$. These variability parameters are of least relevance in the interpretation of the model fits. However, in order to improve the reliability of parameter estimates with smaller sample sizes, the psychological interpretation of the diffusion model allows one to combine experimental conditions and constrain parameters over conditions. For example, when two experimental conditions only differ in the difficulty of the stimuli, one can constrain all parameters but drift rate to be equal across those two conditions. Alternatively, one can think of situations in which all parameters but response bias or response caution are restricted across experimental conditions.

More information about the reliability of the diffusion model parameter estimates comes from studies showing that parameters are recovered properly when the diffusion model is fitted to a data set that is generated from the model itself (Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007; Ravenzwaaij & Oberauer, 2009). The study by Ratcliff and Tuerlinckx (2002) lists some situations in which parameters cannot be estimated properly. In particular the influence of observations from contaminant processes, such as fast outliers, negatively affect parameter estimation. However, the study also shows that with a correction for contaminants and a cutoff for very fast RTs these problems are attenuated.

Robustness

Donkin, Brown, Heathcote, and Wagenmakers (2011) studied whether the conclusions based on the diffusion model are robust against eventual misspecification of the model. To do so, they studied whether conclusions based on the diffusion model were the same as conclusions based on the linear ballistic accumulator model (LBA, Brown & Heathcote, 2008). The comparison of conclusions is interesting, since the LBA model assumes a related but different choice process than the one assumed by the diffusion model (see Figure 8.7). The LBA model describes a two–choice decision process as a race between two accumulators, one for each response option. When the first accumulator reaches the
decision threshold $b$, the associated response is initiated. As in the diffusion model, the rate of information accumulation (drift rate) in each accumulator varies between trials. However, an important difference with the diffusion model is that the LBA model has no within–trial variability in drift rate. Another important difference with the diffusion model is that the LBA model posits an absolute response threshold instead of a relative response threshold.

Despite these differences, the interpretation of key parameters is the same for both the LBA model and the diffusion model: In the LBA model, drift rate $d$ quantifies stimulus difficulty or participant ability, threshold $b$ quantifies response conservativeness and thus regulates the speed–accuracy trade–off, and starting point $k$ quantifies response bias. As in the diffusion model, the distribution of RT is shifted by non–decision time $T_{er}$.

![Linear Ballistic Accumulator Model](image)

Figure 8.1: The linear ballistic accumulator model describes a two–choice decision process as a race between two accumulators. The rate of information accumulation (drift rate) in each accumulator varies between trials. However, in contrast to the diffusion model, there is no within trial variability in drift rate. Figure reprinted from Forstmann et al. (2008), Figure 2(B).

To study the agreement between LBA and the diffusion model, both models were fitted to the same data sets. Results showed that the conclusions based on the diffusion model were similar to the conclusions based on LBA: When one parameter was affected in the diffusion model, it was also affected in the LBA model and the relative magnitudes of effects were generally the same in both models. These results suggest that conclusions based on the diffusion model do not depend on the correctness of the model itself. In other words, these conclusions appear to be robust to misspecification.

**Challenges in RT Modeling**

The diffusion model’s ability to stand the critical tests described in the section above have allowed it to become one of the most successful models in experimental psychology. The model has been applied very successfully to many experimental paradigms. However, there are situations that the model cannot currently handle. This section discusses two situations that pose a challenge to the diffusion model: situations in which behavior is based on multiple processes and situations in which multiple responses should be accounted for.
8.2. Discussion

Multiple Processes

In chapter 7, we studied the transitions from accurate to guessing behavior. The diffusion model would describe this transition by bringing the response boundaries closer and closer, until even the smallest unit of information can trigger a response. There are several reasons why we believe this is not a proper account of the transitions found in our data. First, our study shows that the transition from accurate behavior to guessing is abrupt and not continuous as the diffusion model predicts. Also, the guessing behavior seems to be qualitatively different from accurate behavior. When in the fast guessing mode, people often press the same button many times in a row, behavior that is inconsistent with the prediction of a diffusion model in which boundary separation is zero. Furthermore, with zero boundary separation, the diffusion model simplifies to the uniform distribution that is assumed for non-decision time $T_{\text{er}}$, whereas the guessing RT distribution is certainly not uniform. To solve this problem, in a complete phase transition model of RT, pressure on speed should transform the stimulus–response process from a diffusion–like process to a plausible model of simple RT (e.g., P. L. Smith, 1995).

There are certainly more situations in which the diffusion model in its basic form fails to provide a natural account of behavioral phenomena. Often, the problem lies in the fact that the diffusion model describes a decision process as a time–homogenous process: the drift rate in the model is assumed to be constant over the course of a trial. However, many popular experimental tasks in psychology are designed to study conflicting processes. One such task is the Simon task (Craft & Simon, 1970). In this task, the position of the stimulus on the screen, which the participant is instructed to ignore, is found to influence behavior. Pratte et al. (2010) show that the diffusion model is not able to account naturally for these data. However, it should be noted that the diffusion model provides a very general description of a simple decision process. White and Ratcliff (2011) have shown that the very basic assumptions of the diffusion model can be extended naturally to a more complex model that accounts for data of the Eriksen flanker task (Eriksen & Eriksen, 1974).

Multiple Responses

In the diffusion model, the relative evidence for each of two response options is compared to two response boundaries. In other words, evidence in favor of one response means evidence against the other response. One drawback of this assumption is that tasks with more than two response options cannot be easily modeled with the diffusion model. This drawback does not apply to models that assume that each response has a counter that races against the counters of other responses, such the LBA model and the leaky competing accumulator model (Usher & McClelland, 2001).

In addition to this practical and theoretical advantage, one could favor race models over the diffusion model because race models are neurologically more plausible (Usher & McClelland, 2001; Bogacz, 2007; Usher, Olami, & McClelland, 2002). The leaky competing accumulator model assumes a competition between separate accumulators that each represent a response. The evidence in one accumulator can inhibit the accumulation of evidence in the other accumulator and the accumulated evidence leaks over the course of a trial. The latter assumption mimics the decay of activation of neurons in the absence of input.

A phenomenon that can potentially be modeled using the leaky competing accumulator model is double responding. Double responding refers to the observation that participants now and then press two response buttons almost at the same time. An in-
tuitive explanation of this phenomenon is that the trailing accumulator is too close to initiation of the motor response to still be inhibited. This explanation can probably be modeled naturally in the leaky competing accumulator model by assuming a relatively low reciprocal inhibition on a certain trial. However, fitting the leaky competing accumulator model is problematic. Future advances in numerical techniques might possibly solve this practical problem.

8.3 Concluding Remarks

This dissertation sheds light on the temporal dynamics of behavior in speeded decision making. Participants on RT tasks learn, get distracted, speed up, slow down, get confused, get bored, and eventually may start guessing. One can safely say that participants’ behavior is dynamic. It seems therefore obvious that mean RT or accuracy are at best limited summaries of performance. This thesis shows that a study of the dynamics of speeded decision making requires an assessment of all the data, most notably the changes in entire RT distributions. More generally, this thesis shows that the analysis of psychological data should start from well–formulated ideas about the processes that might have generated the data. When the model that formalizes these ideas provides a reasonable account of the observed data, this constitutes a first step towards a deeper and more comprehensive understanding of behavior than is possible from the assessment of summary measures alone.


142


Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to


account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.


Chapter 9

Nederlandstalige Samenvatting

Dit proefschrift is een verslag van mijn vier jaar durende ontdekkingstocht door reactietijdenland. Ik heb in die tijd wat uithoeken aangedaan, maar toch zijn er twee belangrijke punten die in dit proefschrift steeds terugkomen: Ten eerste blijkt telkens weer hoe nuttig of zelfs essentieel het is om data formeel te modelleren in plaats van gebruik te maken standaard statistieken. Ten tweede blijkt dat het gedrag van proefpersonen niet stabiel is, zoals veel analysetechnieken veronderstellen, maar eerder zeer flexibel en adaptief.

In dit Nederlandstalige hoofdstuk geef ik een korte samenvatting van mijn belangrijkste bevindingen. Voordat ik tot deze bevindingen kom, leg ik eerst uit waarom een formeel model nodig is en beschrijf ik het diffusiemodel, dat ik in dit proefschrift meerdere keren heb toegepast.

9.1 Reactietijden en het Belang van een Formeel Model

Hoe meer hersenen moeten denken over een bepaalde beslissing, hoe langer ze hiervoor nodig hebben. Dit is de eenvoudige logica die aan de basis ligt van de grote populariteit van reactietijd als maat van cognitief gedrag. Enerzijds worden reactietijden gebruikt om te bepalen hoe lang een bepaald cognitief proces duurt, door op een slimme manier de reactietijd op verschillende taken te vergelijken. Anderzijds worden reactietijden gebruikt om af te leiden of bepaalde factoren een positief effect hebben op de cognitie, door de reactietijd van verschillende condities of groepen mensen te vergelijken. Er schuilt echter een lastig probleem in de vergelijking van reactietijden: de samenhang van reactietijd met accuratesse.

Om dit verband te illustreren gebruik ik een veelgebruikte taak, de lexicale-decisietaak. In deze taak ziet een proefpersoon telkens een lettercombinatie op een computerscherm. Wanneer deze lettercombinatie een bestaand woord is, moet de proefpersoon zo snel mogelijk op (bijvoorbeeld) de linker van twee knoppen drukken. Wanneer het een niet bestaand “pseudowoord” is moet de proefpersoon zo snel mogelijk op de rechter knop drukken. Deze taak zou bijvoorbeeld gebruikt kunnen worden om te onderzoeken of ouderen even snel woorden herkennen als jongeren. Stel nu dat je dit onderzoek uitvoert, en je vindt dat de ouderen weliswaar langzamer zijn dan de jongeren, maar dat ze ook minder fouten hebben gemaakt. Wat moet je nu concluderen uit deze data?

Het probleem dat hierboven is beschreven is een gevolg van de speed–accuracy trade–off: het fenomeen dat reactietijd en accuratesse tegen elkaar kunnen worden verhandeld. Hoe sneller je wilt zijn, hoe meer fouten je op de koop toe moet nemen. Hoe accurater je
wilt zijn, hoe meer tijd je zult moeten nemen. In het bovenstaande voorbeeld zou je dus kunnen concluderen dat ouderen slechter zijn in het onderscheiden van woorden, maar je zou ook kunnen concluderen dat ze voorzichtiger reageren dan jongeren.

Er zijn verschillende manieren om met dit probleem om te gaan. Regelmatig wordt de keuze gemaakt om een van de twee variabelen, reactietijd of accuratesse, te negeren. Deze vereenvoudiging kan echter een vertekend beeld opleveren, zoals bovenstaand voorbeeld laat zien. Gelukkig zijn er verschillende modellen die een betere oplossingen bieden door gebruik te maken van beide variabelen. Het bekendste model, dat ik ook in dit proefschrift meerdere keren heb toegepast, is het diffusiemodel. Dit model is in meerdere hoofdstukken van dit proefschrift uitgebreid uitgelegd (zie bijvoorbeeld de introductie). Daarom beschrijf ik hier alleen de essentie.

Het diffusiemodel veronderstelt een min of meer plausibel proces volgens welke beslissingen in een tweekeuzetaak genomen worden. Dit proces wordt gevoed door ruizige informatie over de stimulus (bijvoorbeeld een woord in de lexicale-decisietak). De informatie die wijst op het ene antwoord (“het is een woord”) wordt vergeleken met de informatie die wijst op het andere antwoord (“het is een psuedowoord”). Wanneer het verschil in evidentie voor beide antwoordopties groot genoeg is wordt een beslissing genomen. Dit veronderstelde beslissingsproces is in de basis heel eenvoudig. Toch genereert het proces (als je het bijvoorbeeld met een computer simuleert) een reactietijdverdeling voor zowel goede als foute antwoorden die heel precies lijkt op menselijke reactietijdverdelingen. Belangrijk is dat het exact vastligt hoe de vorm van deze reactietijdverdelingen en de accuratesse afhangen van de instelvariabelen van het proces (de parameters). Deze laatste eigenschap maakt het diffusiemodel een formele model. Dit maakt het mogelijk om de boel om te keren: uit een empirische dataset kun je de parameters bepalen die deze reactietijdverdeling en bijbehorende accuratesse beschrijven. Nu is de belangrijkste kracht van het diffusiemodel dat verschillende parameters zeer specifiek samenhangen met een van de psychologische componenten van een beslissingsproces. Een van de parameters, de drift rate, hangt bijvoorbeeld samen met de moeilijkheid van de stimuli of het vermogen van de proefpersoon om de betreffende beslissing te nemen (onderscheidingsvermogen). Een andere parameter, boundary separation hangt sterk samen met hoe voorzichtig een proefpersoon is met het geven van de respons. Deze precieze reflectie van gedrag in de verschillende parameters van het model maakt het mogelijk om op basis van reactietijd–én accuratessedata te ontleden in welke componenten van het beslissingsproces verschillende condities van elkaar verschillen.

In het bovengenoemde voorbeeld met ouderen en jongeren zou een analyse met het diffusiemodel bijvoorbeeld kunnen uitwijzen dat jongeren een iets beter vermogen hebben om woorden van pseudowoorden te onderscheiden dan ouderen. Daarnaast kan blijken dat het verschil in reactietijd en de hogere accuratesse van ouderen voortkomt uit het feit dat ze voorzichtiger zijn dan jongeren.1

Naast deze verschillen in onderscheidingsvermogen en voorzichtigheid kan een analyse met het diffusiemodel ook verschillen in bijvoorbeeld sensomotorische componenten van het beslissingsproces en a priori voorkeur voor een van de antwoordopties aan het licht brengen. Zoals je hieronder zult lezen, heb ik in een aantal hoofdstukken gebruik gemaakt van van het diffusiemodel om reactietijddata te ontleden tot de verschillende onderliggende beslissingscomponenten.

---

1Onderzoek van Roger Ratcliff, die het diffusiemodel voor reactietijddata heeft ontwikkeld, laat inger- daad een dergelijke interpretatie van leeftijdsverschillen in reactietijd en accuratesse zien. Zie hoofdstuk
9.2 Oefening Baart (niet alleen) Kunst

Door de jaren heen is er een flinke hoeveelheid onderzoek gedaan naar de effecten van oefening op allerhande cognitieve en perceptuele taken. Een van de meest algemene bevindingen is dat oefening mensen sneller maakt. Deze versnelling is lange tijd de belangrijkste focus geweest van onderzoek naar oefeneffecten. Effecten op accuratesse en de verdeling van reactietijden zijn daarmee grotendeels genegeerd. Hoofdstukken 4 en 5 beschrijven twee studies waarin we hebben geprobeerd zowel reactietijd als accuratesse te gebruiken om een beter beeld te schetsen van oefeneffecten.

In het eerste van deze twee hoofdstukken lieten we proefpersonen oefenen op een lexicale-decisietoak. De taak die we voor dit experiment gebruikten bevatte 200 woord- en 200 pseudowoordstimuli. De gehele taak werd 25 keer geofend (proefpersonen namen dus alles bij elkaar 10.000 beslissingen). De diffusiemodelanalyse op de data van deze studie laat zien dat het oefeneffect niet simpelweg een afname van reactietijd is, maar een samenspel van verschillende factoren. Niet alleen werden proefpersonen beter in het onderscheiden van woorden en pseudowoorden, ze werden ook steeds onvoorzichtiger. Deze combinatie van factoren leidde ertoe dat proefpersonen steeds sneller werden, terwijl accuratesse gelijk bleef. Een onverwacht effect was dat proefpersonen ook sneller werden in de sensomotorische component van reactietijd.

De resultaten van deze studie riepen de vraag op of de verschillende facetten van het oefeneffect stimulusspecifiek zijn (en dus te maken hebben met het herhaald aanbieden van dezelfde woorden en pseudowoorden) of eerder taakgerelateerd. Om deze vraag te beantwoorden voerden we de studie in hoofdstuk 4 uit. Hierin lieten we proefpersonen herhaaldelijk oefenen op één stimuluslijst die steeds werd afgewisseld met nieuwe stimuluslijsten. Stimulusspecifieke oefeneffecten zouden alleen op moeten treden op herhaald aangeboden stimuluslijsten, terwijl taakgerelateerde oefeneffecten zowel op herhaald aangeboden lijsten als op nieuwe lijsten zouden moeten optreden. Deze studie liet zien dat de effecten op onderscheidingsvermogen en de niet-beslissingscomponent gedeeltelijk taakgerelateerd en gedeeltelijk stimulusspecifiek zijn. Het oefeneffect op voorzichtigheid bleek echter even sterk op beide stimuluslijsten en dus taakgerelateerd te zijn.

Deze twee studies laten zien dat de focus op veranderingen in gemiddelde reactietijd een beperkt beeld geeft van oefeneffecten. Door gebruik te maken van het diffusiemodel konden we laten zien dat het effect van oefening verschillende facetten heeft.

9.3 Leren van Fouten?

Wanneer proefpersonen in een reactietijdtaak een fout maken, reageren ze vaak langzamer op de volgende stimulus, een fenomeen dat bekend staat onder de naam post-error slowing. De meest gangbare uitleg van post-error slowing is dat proefpersonen constant hun prestatie in de gaten houden en dat ze een fout interpreteren als een teken dat ze de volgende keer wat voorzichtiger moeten reageren. Deze interpretatie van post-error slowing is erg aansprekend en wordt regelmatig voor waar aangenomen. Andere verklaringen van het fenomeen zijn echter wel degelijk denkbaar. Het zou bijvoorbeeld zo kunnen zijn dat proefpersonen afgeleid worden doordat ze een fout hebben gemaakt en dat ze daardoor wat trager reageren op de volgende stimulus. In hoofdstukken 4 en 5 passen we het diffusiemodel toe om te onderzoeken welke processen ten grondslag liggen aan de vertraging na fouten.

In de eerste studie gebruikten we een zeer grote bestaande lexicale-decisiedataset.

De resultaten van deze twee studies laten zien dat gedrag van proefpersonen zeer dynamisch is. Niet alleen strategische aanpassingen veranderen het gedrag, maar ook de snelheid van sensomotorische processen en informatieverwerking blijkt te fluctueren tijdens een taak.

Tijdens onze analyse van post–error slowing stuitten we op een probleem met betrekking tot de manier waarop traditioneel de mate van vertraging na een fout wordt uitgerekend. In deze traditionele berekening wordt de gemiddelde reactietijd op trials na een correcte respons afgetrokken van de gemiddelde reactietijd na een foute respons. Dit verschil wordt dikwijls als maat van cognitieve controle gebruikt. In hoofdstuk 6 laten we zien dat deze methode gevoelig is voor een artefact van globale fluctuaties in gedrag. Zo kunnen fluctuaties in motivatie leiden tot een schijnbare vertraging na een fout, terwijl fluctuaties in voorzichtigheid kunnen leiden tot een schijnbare versnelling na een fout. Voor dit probleem beschrijven we een eenvoudige oplossing, waarbij reactietijden na de fout worden vergeleken met reactietijden vóór de fout. Met simulaties laten we zien dat deze methode inderdaad robuust is tegen het beschreven artefact.

9.4 Van Accuraat Reageren tot Gokken

De meeste reactietijdmodellen, zoals ook het diffusiemodel, veronderstellen dat de verhandeling van reactietijd en accuratesse continu is. Dit wil zeggen dat, wanneer de tijdsdruk geleidelijk wordt opgevoerd, een proefpersoon langzaamaan steeds sneller gaat reageren en steeds meer fouten gaat maken, net zo lang totdat accuratesse is gedaald tot kansniveau. In hoofdstuk 4 onderzoeken we deze aanname van continuïteit in de speed–accuracy trade–off. Tegenover deze continuïteitsaanname stellen we de voorspelling van een facetransitienmodel voor reactietijden. Dit model veronderstelt dat het gedrag van een proefpersoon in een reactietijdtaak wordt bepaald door twee afzonderlijke responsmodi: een accurate, relatief langzame modus, en een gokmodus, waar responsen snel zijn, maar accuratesse op kansniveau. Dit model is gebaseerd op catastrofetheorie. Catastrofetheorie biedt een wiskundige beschrijving van systemen waarin kleine veranderingen in omgevingsvariabelen leiden tot abrupte veranderingen in het geobserveerde gedrag. Twee belangrijke voorspellingen die volgen uit catastrofetheorie zijn bimodaliteit en hysteres. Deze twee voorspellingen hebben we in twee experimenten onderzocht.

De eerste voorspelling, bimodaliteit, houdt in er twee elkaar uitsluitende responsmodi zijn. In het facetransitienmodel betekent dit dat proefpersonen ofwel de stimulus verwerken
en relatief accuraat reageren, ofwel snel gokken zonder de informatie in de stimulus te verwerken. Gedrag tussen deze twee modi is instabiel. In het experiment dat we opzetten om deze voorspelling te testen werden proefpersonen aangemoedigd om zo dicht mogelijk bij de 75% correct en zo snel mogelijk te reageren. Deze opdracht vereist gedrag dat tussen de twee staten in licht. Proefpersonen bleken echter niet in staat deze opdracht uit te voeren door consequent relatief onvoorzichtig en snel te reageren, maar ze wisselden relatief langzame, accurate responsen af met snelle gokken. Deze bevindingen bevestigden de voorspelling van bimodaliteit.

De tweede voorspelling, hysterese, houdt in dat beide responsmodi van zichzelf stabiel zijn. Dit zou ertoe moeten leiden dat een proefpersoon die accuraat reageert en wordt aangespoord steeds sneller te reageren in staat zal zijn om accuraat te blijven reageren tot op een bepaald moment zijn gedrag naar gokken overspringt. Andersom luidt de voorspelling dat wanneer een een proefpersoon aan het gokken is, en de druk op accuratesse wordt opgevoerd, deze proefpersoon pas bij een relatief hoge druk op accuratesse weer terugkeert naar accuraat gedrag. Deze voorspelling testten we in een experiment waarin we mensen door middel van een uitbetalingsregel steeds van heel accuraat reageren tot gokken brachten en van gokken weer tot accuraat reageren. De sprongen van de accurate modus naar de gokmodus bleken inderdaad bij een andere instelling van de uitbetalingsregel plaats te vinden dan de sprongen van de gokmodus naar de accurate modus. Deze bevindingen bevestigen de voorspelling van hysterese.

9.5 Conclusie

Dit proefschrift schetst een uitgebreid beeld van de dynamiek van het menselijk gedrag in eenvoudige reactietijdtaken. Proefpersonen die zulke taken uitvoeren kunnen leren, versnellen, vertragen, afgeleid worden, zich vervelen en zelfs overgaan tot gokken. Kortom, hun gedrag is zeer dynamisch. Daarom kunnen gemiddelde reactietijd en accuratesse slechts een oppervlakkige indruk geven van het gedrag. Om de dynamiek van het gedrag in al zijn facetten bloot te leggen, moeten we gebruik maken van alle informatie die in de data aanwezig is, inclusief accuratesse en de vorm van de verdeling van reactietijden. In ruimere zin laat dit proefschrift zien dat de analyse van psychologische data altijd moet steunen op een nauwkeurig geformuleerde theorie over het proces dat aan die data ten grondslag ligt. Wanneer zo’n nauwkeurige theorie een passende beschrijving van de gevonden data geeft, is dit een eerste stap richting een meer alomvattend begrip van gedrag.
Dankwoord

Het moeilijkste aan het schrijven een proefschrift zijn de laatste loodjes. Dit is niet alleen omdat laatste loodjes altijd het zwaarst zijn. Veel erger is dat ik de enige ben die het kan afmaken. Hoe later het loodje, hoe duidelijker het wordt dat ik vier jaar lang heb mogen baden in een weelde van intellect, behulpzaamheid, humor en gezelligheid. Nu schrijf ik m’n ultieme laatste loodje: het dankwoord.

De belangrijkste persoon aan wie ik deze fantastische tijd als AiO te danken heb is Han. Toen ik in vrij matig tempo mijn doctoraal Methodenleer bij Han had afgerond, gaf hij mij de kans om te laten zien dat ik ook in minder dan een jaar een studie kon uitvoeren. Al tijdens mijn doctoraal, maar meer nog tijdens mijn AiO-project, ben ik fan geworden van Han zijn originaliteit. Geen onderzoeker ben ik nog tegengekomen die zo weinig dogmatisch is. Ieder bestaand vraagstuk probeert hij van een nieuwe kant te bekijken. Geen wetenschappelijke gewoonte is hem heilig, behalve de wetenschappelijkheid zelf. Klopt het, kan het en hebben we er wat aan? Ik voel me bevoorrecht bij de geweldige vakgroep Methodenleer onder Hans leiding gewerkt te mogen hebben.


Ik zou nu nog een hele lijst namen kunnen geven van goeie vrienden, fijne co–auteurs, bevriende collega’s en kamergenoten die ik wil bedanken voor hun steun, kennis, vriendschap en kritiek. Ik wil daaraan ontsnappen door twee archetypen van steun en toeverlaat te beschrijven. Ten eerste de Ruud. De Ruud helpt je realiseren dat alles wat je doet onzin is. Hij is een belangrijk toevluchtsoord als je aan je dagelijkse bezigheden wilt ontsnappen en een vleugje van de buitenwereld wilt proeven. De aanstichter van menig AiO–dip. Gelukkig ook degene die je helpt om die dip snel te verdringen. Ten tweede is er de Raoul. De Raoul doet je realiseren dat wat je doet belangrijk is. Hij is je toevluchs-
oord op de momenten dat je net iets te veel de betrekkelijkheid inziet van wat je doet. Een therapeut voor de AiO-dip. Tot in de kroeg principieel en altijd behulpzaam. De labels zijn uiteraard zo gekozen omdat de betreffende personen zeer goede voorbeelden zijn, maar als je je in één of beide van deze archetypen herkent, kan ik je verzekeren dat ook jij een heel belangrijke bijdrage hebt geleverd aan mijn promotie.