The Importance of Choice and Response Times

Chris Donkin - BPsych(Hons)/BMath

Thesis submitted for Doctorate of Philosophy, January 2010

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

I hereby certify that this thesis is in the form of a series of published papers of which I am a joint author. I have included as part of the thesis a written statement from each co-author, endorsed by the Faculty Assistant Dean (Research Training), attesting to my contribution to the joint publications.

Signed:

Chris Donkin

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

12 October 2009

To Whom It May Concern:

This letter outlines the contributions made by Chris Donkin to the papers which will make up his PhD Thesis by publication. The letter is co-authored by Chris' two supervisors Dr. Scott Brown and Professor Andrew Heathcote as both of us have been involved in all six of the publications which make up Chris' thesis.

Donkin, C., Brown, S.D., Heathcote, A. (2009). ChoiceKey: A real-time speech recognition program for psychology experiments with a small response set. *Behavior Research Methods*, *41*, 154-162.

This project was led by Chris, 75% contribution. Chris wrote, developed and tested the software as well as taking the lead in the write-up.

Donkin, C., Heathcote, A., Brown, S. & Andrews, S. (2009). Non-Decision Time Effects in the Lexical Decision Task. In N. A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. ISBN 978-0-9768318-5-3

This project was led by Chris, 70% contribution. Chris did the majority of mathematical modeling and led the write-up of the project.

Donkin, C., Averell, L., Brown, S.D. & Heathcote, A. (accepted 25/5/2009). Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator, *Behavior Research Methods*

This project was jointly developed by Chris Donkin and Lee Averell. Chris and Lee each contributed 40% to the project. Chris developed the software for the paper. Chris and Lee divided the write-up of the project evenly.

Donkin, C., Brown, S.D. & Heathcote, A. (accepted 25/6/2009). The over-constraint of response time models: Rethinking the scaling problem, *Psychonomic Bulletin & Review*

This was a jointly developed project. Chris took the lead in write-up and did much of the mathematical modelling work, 50% contribution.

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

Donkin, C., Brown, S.D., Heathcote, A. & Marley, A.A.J. (2009). Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research, 73*, 308-316.

This project was led by Chris, 70% contribution. Chris did the model fitting and took the lead in the write-up of the project.

Brown, S.D., Marley, A.A.J., Donkin, C. & Heathcote, A. (2008). An integrated, principled account of absolute identification. *Psychological Review, 115*, 396-425.

This project was led by Dr. Brown. Chris particularly contributed to write-up and background experimental work. 20% contribution.

Regards,

.

Dr. Scott Brown

Professor Andrew Heathcote

School of Psychology, Psychology Building, University Avenue

Endorsed by AS(RT)

University of Victoria

PO Box PO Box 3050 STN CSC
Victoria British Columbia V8W 3P5 Canada
Tel (250) 250-472-2067 Fax 250-721-8829
E-mail ajmarley@uvic.ca

Psychology

October 13, 2009

To Whom It May Concern:

I understand that Chris Donkin is proceeding with a PhD Thesis by publication. This letter outlines the contributions that he made to the two papers with which I have been involved.

Donkin, C., Brown, S.D., Heathcote, A. & Marley, A.A.J. (2009). Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research*, *73*, 308-316.

This project was led by Donkin, 70% contribution. Donkin did the model fitting and took the lead in the write-up of the project.

Brown, S.D., Marley, A.A.J., Donkin, C. & Heathcote, A. (2008). An integrated, principled account of absolute identification. *Psychological Review*, *115*, 396-425.

This project was led by Dr. Brown. Donkin particularly contributed to write-up and background experimental work. 20% contribution.

A A. J. Marley
Adjunct Professor
http://www.uvic.ca/psyc/marley/

Research Professor (part-time)
Centre for the Study of Choice
University of Technology Sydney
http://www.censoc.uts.edu.au/

Endorsed by AJRT

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

17 November 2009

To Whom It May Concern:

This letter outlines the contributions made by Chris Donkin to the papers which will make up his PhD Thesis by publication. The letter is co-authored by Lee Averell a college of Chris' involved in one of the publications which make up Chris' thesis.

Donkin, C., Averell, L., Brown, S.D. & Heathcote, A. (accepted 25/5/2009). Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator, *Behavior Research Methods*

This project was jointly developed by myself and Chris Donkin with equal contributions of 40% each. Chris developed the software for the paper. Myself and Chris divided the write-up of the project evenly.

Regards,

Mr Lee Averell

School of Psychology, Psychology Building, University Avenue

Endorsed by ADRT

The University of Sydney

Faculty of Science
School of Psychology

NSW 2006 AUSTRALIA

Sally Andrews
Professor of Cognitive Psychology

Griffith Taylor Building(A19)
Telephone +61 2 9351 8297
Facsimile +61 2 9351 2603
sallya@psych.usyd.edu.au

2nd November, 2009

To whom it may concern

I am a co-author on one of the papers contributing to Chris Donkin's PhD Thesis by Publication:

Donkin, C., Brown, S.D., Heathcote, A. & Andrews, S. (2009). Non-Decision Time Effects in the Lexical Decision Task. In N. A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. ISBN 978-0-9768318-5-3.

My co-authorship on the paper reflects my contributions to discussion of the issues that motivated the mathematical modeling reported in the paper, and the implications of the outcomes. I made no direct contribution to the modeling work or to the write-up of the paper, which was led by Chris Donkin supported by his supervisors.

Yours sincerely

Sally Andrews

Endorsed by ADRT

Acknowledgements

Thank you to anyone who has contributed at all to this thesis. From the many unknown soldiers who battled through our mind-numbing tasks, to the other members of the Newcastle Cognition Lab, who put up with my inane ranting and raving and provided support and advice when needed. To my two supervisors, Scott Brown and Andrew Heathcote, I can not express enough my gratitude for all of the help and guidance you have provided. That you both work so hard and lead such rich lives, yet find so much time for your students is something we can all only hope to one day accomplish.

To friends, thank you all. To family, thank you for your patience and the head start you blessed me with. And finally, to my wife Jette, the last three years would not have been possible without your constant love and support. From the bottom of my heart, thank you.

Table of Contents

Abstract

The six published papers making up the main body of this thesis aim to communicate the many benefits provided by a combined, model-based, analysis of choice and response time (RT). Consideration of both choice and RT is important largely because the two naturally trade with each other – fast responses are error-prone while slow responses are more often correct. Quantitative models based on evidence accumulation allow for a combined analysis of choice and RT. These so-named choice RT models take into account both the speed and accuracy of responses to produce quantities associated with performance, response caution, bias, and other elements of decision making. The first section contains tools sufficient to carry out a model-based choice RT analysis. The first chapter of the thesis provides software for the collection of both choice and RT data from vocal responses. The second chapter contains software for applying a particular choice RT model to data. Choice RT models can be used to better understand differences in the way that decisions are made, say between different groups or across different experimental conditions. This powerful ability, however, requires many important assumptions be made. The second section of the thesis deals with issues surrounding the assumptions made about the effect of experimental manipulations on the parameters of choice RT models. The third chapter demonstrates the ramifications of such assumptions, while the fourth chapter shows how careful choice RT users must be when making these assumptions. The third and final section details a process model of both choices and RT in absolute identification, in which the use of a choice RT model as a description of the decision process is integral. The fifth chapter outlines the process model and the sixth chapter tests a prediction of the model which highlights the importance of a quantitative account of both choice and RT in absolute identification.

Introduction

The Speed-Accuracy Trade-Off

Much of experimental psychology uses accuracy and response time (RT) data to make inferences about the processes underlying performance. The overarching theme of this thesis is that both accuracy and RT are important and should be considered together when making these types of inferences. One of the main reasons we must consider both variables is the potential trade-off between how long a response takes to make and the likelihood that the response will be correct. The well-established *speed-accuracy trade-off* states that responses made quickly are more likely to be incorrect than responses which are slower (e.g. Johnson, 1939; Luce, 1986; Pachella, 1974; Schouten & Bekker, 1967; Wickelgren, 1977). This phenomenon can make it very risky to interpret just one of the two variables without considering the other. For example, imagine our aim was to compare two different populations (group A and B, say) on their performance in a particular task. Imagine that group A was able to respond, on average, in 500ms, faster than group B, who had a mean RT of 1000ms. It is tempting to infer that group A were better at the task than group B. What if, however, we subsequently found out that group A made more errors (15% incorrect responses) than group B (5% incorrect responses). Because group A were faster but made more errors than group B it is possible that both groups performed the task equivalently, but that group B was more cautious. In other words, it is possible that if group A were more cautious, such that they too made errors only 5% of the time, that their mean RT would also be 1000ms.

The previous example is a simple demonstration of why experimental psychologists should collect and consider both accuracy and RT data. Unfortunately, even when both accuracy and RT data are collected, it is difficult to quantify the

difference between group A and group B in terms of both RT and accuracy. For example, imagine that group B remained 500ms slower than group A but now made errors 10% of the time, only 5% better than group A. Would we now think of group A as better than group B? What if group B made incorrect responses on 14% of trials – only 1% better, but still 500ms slower than group A? This question is further complicated when one considers that the trade-off between speed and accuracy looks exponential (McElree & Dosher, 1989), suggesting that the size of the trade-off between the speed and accuracy of responses is a function of how accurate the responses are, or how long they take to make.

The standard approach of submitting accuracy and mean RT to Analysis of Variance (ANOVA) hypothesis tests does not help solve this problem of quantification of differences. At best, an ANOVA might reveal that group A and group B do not differ significantly in one of the response variables. However, because of the shape of the speed-accuracy trade-off function, when accuracy is close to ceiling then even small, non-significant, changes in accuracy values can cause large changes in RT. In any case, our problem of quantifying the difference between groups for a given set of accuracy and RT data remains unsolved. For example, the question remains as to how much better group A is compared to group B when they are equally accurate, but group A performs 500ms faster on average.

Looking just at mean RT and accuracy rates also fails to identify why differences between groups occur. For example, simply observing a 500ms difference in mean RT between two groups gives us little information about why the difference has occurred. In particular, we can not distinguish between an explanation which says the difference in mean RT is due to participants' performance, or simply because one group's members

took longer to press the response buttons. The situation is further complicated when there is a potential trade-off between the speed and accuracy of responses. In such cases there is a question of whether response caution is the sole cause of differences in RT between groups, or if there are also underlying differences in performance between the groups.

The key to understanding the underlying causes of differences in accuracy and RT comes from analysing not just mean RT, but entire RT distributions for correct and incorrect responses (as well as how often errors are made). The shape, scale and location of RT distributions can be used to infer about how decisions are made. For example, if we observed that all responses for one group were 200ms slower than for another group (i.e. a 200ms shift in RT distributions), then we might conclude that the differences were due to some constant factor not related to the decision process itself, but to external processes such as the time taken to encode the stimulus, or the difference in time taken to press the response buttons.

A Model-Based Solution

There are many quantitative, cognitive models which distil accuracy and RT distributions into latent variables representing the processes underlying relatively simple decisions. The most successful models of choice and RT (choice RT models) are the evidence accumulation (or sequential sampling) models, and though many variants of this type of model exist (e.g. the diffusion model, Ratcliff, 1978; the EZ diffusion model, Wagenmakers, van der Maas, & Grasman, 2007; the Poisson counter model, Pike, 1966; Van Zandt, Colonius, & Proctor, 2000; the accumulator model, Smith & Vickers, 1988; the leaky competing accumulator model, Usher & McClelland, 2001; the ballistic accumulator model, Brown & Heathcote, 2005; and the linear ballistic

accumulator model, Brown & Heathcote, 2008), they all share a basic framework for describing decisions – evidence accumulation models assume that participants sample information as they attend to a particular stimulus. This information is then taken as evidence for one of the competing responses. Evidence is accumulated until it reaches some threshold level for one of the potential responses. That response is then chosen, with the time taken for evidence to reach the threshold being the decision time component of the RT (Stone, 1960). Since errors occur even in the most simple of decisions, most evidence accumulator models include variability, or noise, at some level of the decision process, because this variability means that on some trials evidence for incorrect responses will reach threshold before evidence for the correct response.

Choice RT models summarise RT distributions for correct and incorrect responses using a number of latent variables which represent processes underlying how decisions are made. Of these variables, three are common across all variants of evidence accumulation models and are usually of interest when describing how individuals are responding (Wagenmakers et al., 2007). The three variables are *rate of processing*, *response caution* and *non-decision time*.

Rate of processing, often simply called drift rate, refers to the rate at which evidence for a response is accumulated, and is a measure of how well the task is being performed. To see this, consider a small and a large drift rate; a small drift rate means there is little evidence that a particular response should be made. If this response is the *true* correct response on a particular trial, then evidence accumulation happens slowly for the correct response, and therefore increases the probability that any noise in the process will lead to an error. Small drift rates occur when the task is difficult or when the participant is performing poorly. On the other hand, a large drift rate implies that

there is a large amount of evidence indicating that the response should be made. Large drift rates, therefore, result in fast and accurate responses, while small drift rates result in slow and error-prone responses.

Response caution refers to how much evidence is required before a response is made, and is largely responsible for producing a trade-off between the speed and accuracy of responses. In general, by setting a large threshold for how much evidence is required before making a response, a participant will wait longer to make a decision. Waiting this extra time means that the response is more likely to be correct, as noise in the evidence accumulation process will be integrated out with time. When the threshold is set low, however, responses will be faster but more influenced by noise in the system, and hence more likely to be incorrect.

Non-decision time refers to the time taken for all aspects of RT which are not strictly part of the evidence accumulation process. This parameter exists because it is impossible to measure exactly how long it takes for a person to accumulate evidence. Instead, the standard RT recorded by experimenters is typically composed of decision, or accumulation, time as well as other non-decision time components such as the time taken to perceive and encode the stimulus, and the time taken to execute a motor response once a response is selected to be made. The non-decision time is added to the decision time produced by the evidence accumulation process to give a predicted RT.

Though all evidence accumulation models have some instantiation of these three latent variables, their exact form within any particular model varies substantially. The different choice RT models also make considerably different assumptions about what noise is necessary to account for accuracy and RT data. What follows is an overview of some of the more popular choice RT models, with particular focus on two things: how

the three aforementioned latent variables are implemented, and which sources of noise are assumed to be important enough to model.

Review of Choice Response Time Models

Discrete Beginnings

One of the first attempts to model RT distributions was the random walk model (Laming, 1968; Link & Heath, 1975; Stone, 1960). In a random walk process, time passes in discrete time steps of length Δt. During each time step some evidence is extracted from the environment suggesting which of the two possible responses (A or B) is correct. This evidence then increments some evidence accumulation counter, say x, such that if the evidence supports response A the value of x increases, and if the evidence supports response B then x decreases. When x equals some threshold value, say a for response A and 0 for response B, then that particular response is made, and the number of time intervals of size Δt determines the time taken for the decision to be made.

Evidence accumulation begins at some intermediate value, $0 < z < a$. If there is no bias towards either responding A or B then $z = a/2$, the midpoint between the two response threshold values. If there is bias towards one particular response then evidence will start closer to that response threshold value; if there is bias for responding A then $z > a/2$ and if there is bias for responding B then $z < a/2$. During each time step the amount of evidence added to or subtracted from x is sampled from a normal distribution with mean δ and standard deviation s[1]. This δ value is the drift rate parameter in a random walk model because it indicates the rate at which evidence accumulates towards

1 s in the random walk, and diffusion models, is generally set fixed to an arbitrary value. This parameter is fixed because all choice RT models have a scaling property which means that a subset of their parameters can be multiplied by a constant to give the same predictions. All choice RT models, therefore, have to have one parameter be fixed as constant so that parameters can be estimated from data. Chapter 4 of this thesis is concerned mainly with this scaling property and how the scaling parameter can be used.

boundary a or 0. A positive drift rate indicates more evidence for response A, while a negative drift rate suggests more evidence for response B. Drift rates closer to zero lead to slower and more error-prone responses because the accumulation process is influenced predominantly by the variability in drift rate between time steps. In other words, when drift rate is small then x tends to reach thresholds a or 0 due to random fluctuations in the accumulation of evidence rather than due to incoming information from the environment. In contrast, when drift rate is large then variability in drift, s, will have smaller influence and responses will tend to be correct.

The size of a reflects response caution in the random walk model, such that if a is small then little evidence is required to trigger a response, and errors due to drift variability will occur often. On the other hand, if a is large then the effects of variability in drift rate will integrate out and responses will be more often correct. Non-decision time in the random walk model, T_{er}, is added to the decision time to give the standard RT.

The Diffusion Model

Ratcliff (1978) used a continuous time version of the random walk model, such that $\Delta t \rightarrow 0$, to account for performance in recognition memory tasks. The accumulation of evidence in the continuous version of a random walk model mimics a Wiener process, or Brownian motion, and is usually referred to as a diffusion model. To accommodate the fact that error responses tended to be slower than correct responses in recognition memory experiments (where accurate responses were preferred over fast ones), Ratcliff (1978) added the additional assumption that drift rate, δ, varied from trial-to-trial according to a normal distribution with mean v and standard deviation η. The addition of trial-to-trial variability in drift rate allowed Ratcliff's version of a

diffusion model to account for slow error responses (an explanation of why this works is beyond the scope of this review, but see Ratcliff, 1978 for a detailed explanation).

Ratcliff and Rouder (1998; see also Ratcliff, Van Zandt & McKoon, 1999; Smith & Vickers, 1988) showed that error responses from the one experiment could be both faster and slower than correct responses when the decisions were high and low in accuracy, respectively. To accommodate this pattern of error RT using a diffusion model, Ratcliff and Rouder (1998) added an additional source of variability to the diffusion model of choice RT: trial-to-trial variability in the starting point of evidence accumulation. Prior to Ratcliff and Rouder (1998) it was assumed that on each trial that the evidence accumulator began at a fixed point z. Ratcliff and Rouder (1998) showed that a diffusion model could predict fast errors if start-point, z, was allowed to vary according to a uniform distribution with centre z and range s_z. Having both trial-to-trial variability in start point and drift rate allows a diffusion process to produce both faster and slower error RTs for easy and hard conditions, even within a single block of experimental trials. Note that Laming (1968) incorporated start-point variability into a random walk model to account for fast errors.

A third source of trial-to-trial variability was added to the diffusion model of choice RT by Ratcliff and Tuerlinckx (2002) to explain changes across experimental conditions in the very fastest responses made by participants. The authors showed that a diffusion model predicts that regardless of drift rate, the fastest responses made by participants all take a similar amount of time (sometimes called a “flat leading edge” of the RT distribution). They demonstrated that the diffusion model gave much better account of empirical data when non-decision time was allowed to vary according to a uniform distribution with centre T_{er} and range s_t. Allowing non-decision time to vary

across trials also helped the diffusion model account for performance in the lexical decision task, where relatively large changes in the leading edge were observed across stimulus-based conditions (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

A diffusion model with these three sources of trial-to-trial variability is now the most successful and widely used model of choice RT, and is due largely to the work of Ratcliff and colleagues (in recognition, this particular implementation of the diffusion model is often called "the Ratcliff diffusion model", Wagenmakers, 2009). There are alternative diffusion models, such as the Ornstein-Uhlenbeck (OU) model (Busemeyer & Townsend, 1992, 1993; Roe, Busemeyer, & Townsend, 2001). This differs from the already described (Wiener) diffusion model such that the evidence total, x, decays away from response thresholds as decision time increases. Ratcliff and Smith (2004) showed that the OU model did not provide as good an account of data as the Wiener diffusion model, and that estimates of the decay parameter often approached zero, making the OU model equivalent to the Wiener model.

Recently, Wagenmakers et al. (2007) provided simple methods for estimating rate of processing, response caution and non-decision time parameters for a Wiener diffusion model. This method, called the EZ-diffusion model, involves the estimation of the a, δ and T_{er} parameters via method of moments using the mean and variance of RT and the percentage of correct responses. The EZ-diffusion model, unlike the Ratcliff diffusion model, does not assume response bias nor between-trial variability. Further discussion of the EZ-diffusion model is contained in a later section.

Both random walk and diffusion models are examples of single accumulator models, as evidence is tracked in a single accumulator. There also exist a set of multiple

accumulator models which use an accumulator for each possible response. Before moving on to a discussion of multiple accumulator models of choice RT, I wish to note that this distinction between models in terms of the number of accumulators, is largely for explanatory purposes. Recall that in a random walk, or diffusion model, that evidence for response A is perfectly negatively correlated with evidence for response B. Thus, the diffusion model can be re-interpreted as a multiple accumulator model simply by having two accumulators wherein an increase of evidence in accumulator A causes an equivalent decrease of evidence in accumulator B. Note, however, that the diffusion model can not be extended to model choices between more than two alternatives without adjustments to the model's framework and core assumptions.

Multiple Accumulator Models.

The recruitment model (LaBerge, 1962) was one of the first choice RT models to use a separate accumulator for each possible response. In the recruitment model time passes in discrete time periods and on each passage of time a unit of evidence is placed in just one of the available accumulators. Thus, in LaBerge's recruitment model both time steps and the increment in evidence are discrete. The recruitment model fails to account for the shapes of empirical RT distributions for correct and error responses, particularly for conditions in which responses are slow. I will cover two alternatives related to the recruitment model. First, the accumulator model (Smith & Vickers, 1988; Vickers, 1970), which also assumes discrete, equally-spaced time periods, but that the amount of evidence incremented between these time periods is sampled from a continuous distribution. Second, the Poisson counter model (LaBerge, 1994; Pike, 1966, 1973; Smith & Van Zandt, 2002; Townsend & Ashby, 1983; Van Zandt et al., 2000), assumes the opposite, that the amount of evidence accumulated on each trial is fixed but

that the time intervals in which evidence arrives varies from step-to-step.

In the accumulator model (Smith & Vickers, 1988, 1989; Vickers, 1970, 1979) evidence is accumulated at equally-spaced time steps. At each time step how much evidence is drawn from the environment is sampled from a normal distribution. This evidence value is then compared to a criterion value, if the evidence is larger than the criterion then the difference between the criterion and the evidence value is added to counter B, and if the evidence is smaller than the criterion then counter A is increased by the same difference. When the evidence in either counter reaches a response threshold then that response is made, and the time taken to make the response is the number of time steps multiplied by a constant which converts time steps to seconds.

The distance of the mean of the normal distribution of evidence values from the criterion value is equivalent to the drift rate in the diffusion model, in that it reflects the rate of processing – when the mean is very different from the criterion value then accumulation of evidence will be fast. The response threshold parameter in the accumulator model reflects response caution, and non-decision time parameter in the accumulator model is equivalent to that in the diffusion model. Indeed, in all of the choice RT models to be discussed, non-decision time is implemented in a similar way.

Smith and Vickers (1989) showed that an accumulator model with three sources of between-trial variability provided a good account of empirical data. Firstly, the mean of the evidence accrual distribution was assumed to vary from trial-to-trial according to a normal distribution. Secondly, non-decision time was assumed to vary across trials. Thirdly, the response threshold was allowed to vary from trial-to-trial according to an exponential distribution. Other distributions have been used to describe variability in threshold, though the exponential distribution was found to provide the best fit to

empirical data (Ratcliff & Smith, 2004; Smith & Vickers, 1988,1989).

In the Poisson counter model (LaBerge, 1994; Merkle & Van Zandt, 2006; Otter, Allenby, & Van Zandt, 2008; Pike, 1973; Smith & Van Zandt, 2002; Townsend & Ashby, 1983; Van Zandt et al., 2000) it is assumed that equal amounts of evidence arrive on each time step, but that the time steps vary in size. The time between when evidence arrives in each accumulator is assumed to be exponentially distributed with separate rate parameters for each possible response. Because the time between evidence arrival is exponential, the rate at which evidence increases in each accumulator is distributed according to a Poisson process. The evidence accumulation process continues until evidence in one of the accumulators reaches a response threshold.

The rate at which evidence arrives for each accumulator is the rate of processing parameter in the Poisson counter model. Response caution, like in the accumulator model, is summarised using the response threshold parameter. As an aside, in all multiple accumulator models, bias is implemented using different response threshold values for different accumulators. For example, if there was bias for response A then the threshold for accumulator/counter A could be made smaller than the threshold for accumulator B. This would mean that less evidence is required to make response A than response B.

Three sources of between-trial variability have been added to the Poisson counter model (e.g. Ratcliff & Smith, 2004). Firstly, like the diffusion and accumulator model, non-decision time was assumed to vary. Secondly, the rate of arrival of information for each counter was assumed to vary across trials, though the form of this variability differs somewhat from that in previously discussed models (for details see Ratcliff & Smith, 2004). Thirdly, response thresholds were assumed to vary according

to a geometric distribution. It has been shown that despite the addition of these sources of variability the Poisson counter model is unable to produce both fast and slow errors within experimental blocks (Ratcliff & Smith, 2004; Van Zandt et al., 2000). Ratcliff and Smith (2004) did show, however, that the accumulator model and the Wiener and OU diffusion models were capable of predicting this pattern.

In both the Poisson counter model and accumulator model, evidence in any one accumulator accrues independently of evidence in other accumulators. In the single accumulator models (such as Ratcliff's diffusion) the evidence for one response is assumed to be perfectly negatively correlated with evidence for the alternative response. Usher and McClelland (2001) proposed the leaky competing accumulator model, which parameterises the amount of inhibition between accumulators, and hence can mimic both independent and inversely related evidence accumulation. The leaky competing accumulator model also differs from the accumulator and Poisson counter model in that it assumes that the amount of evidence accrued in an accumulator decays over time (cf. McClelland, 1991; Diederich, 1997).

The leaky competing accumulator model (Usher & McClelland, 2001) assumes that evidence from the environment drives an accumulator for each possible response. This input from the environment is assumed to follow a Wiener process, as in the diffusion model. This makes the rate of accumulation of evidence due to input from the environment equivalent to drift rate in the diffusion model, and is, therefore, considered the rate of processing parameter within the model. Like the OU diffusion model, the leaky competing accumulator model assumes that the amount of evidence in any accumulator decays at a rate proportional to its current evidence level. This leakage was justified by Usher and McClelland using both empirical (e.g. Pietsch & Vickers, 1997)

and neural (e.g. Amit, 1989) evidence. To counteract the decay process, the leaky competing accumulator model also includes a self-excitatory process, which means that evidence in any accumulator also tends to increase at a rate proportional to its current amount of evidence. The model also includes competition between accumulators, such that evidence in one accumulator inhibits the rate of evidence accrual in the other accumulator(s), again at a rate proportional to the current amount of evidence. Like the other choice RT models discussed, the leaky competing accumulator model makes a response once evidence in one accumulator reaches a set response threshold, which acts as the response caution parameter in the model.

Usher and McClelland (2001) showed that the leaky competing accumulator model does not require trial-to-trial variability in drift rate to produce slow error RTs. Ratcliff and Smith (2004) showed that this was largely due to inhibition between accumulators; such that removing competition between accumulators means that the model can no longer predict slow error RTs. Usher and McClelland also showed that the model is able to predict fast error RTs by assuming that the start point of evidence in any accumulator varies from trial-to-trial (cf. Ratcliff & Rouder, 1998).

The leaky competing accumulator model can mimic, as a result of its architecture, the Ratcliff (i.e. Wiener) diffusion model and the OU diffusion model. In particular, the leaky competing accumulator looks like a Wiener diffusion model when inhibition is high and leakage and self-excitation are close to zero. The model mimics an OU diffusion model when both inhibition and leakage are high. This is largely because the rate of accumulation of evidence due to input from the environment follows a Wiener process. However, because the leaky competing accumulator model consists of a race between multiple accumulators, the model never truly becomes either form of the

single accumulator diffusion model. We now consider two recent models of choice RT which assume no variability in the accumulation of evidence (i.e. that accumulation of evidence is ballistic).

Ballistic Models.

Brown and Heathcote (2005) showed that a simplified version of the leaky competing accumulator model, the ballistic accumulator (BA) model, was able to account for all benchmark choice RT phenomena – the shape of RT distributions, the speed-accuracy trade-off, as well as both fast and slow errors. The only difference between the BA and Usher and McClelland's (2001) leaky competing accumulator model is that there is no moment-to-moment variability in the evidence accumulation process. In other words, evidence from the environment was not assumed to follow a Wiener process, but was assumed to be noiseless (hence, "ballistic"). Brown and Heathcote (2005) showed that with between-trial variability in the rate at which evidence accumulates (drift rate) and the start point of evidence accumulation, passive decay and self-excitation of accumulated evidence, and lateral inhibition between accumulators that the BA model was able to accommodate empirical data from a simple discrimination task.

Further to this, Brown and Heathcote (2008) showed that an even simpler version of the BA model, the linear ballistic accumulator (LBA) model, in which accumulation was assumed to be free of leakage, excitation and competition, was equally capable of capturing the important empirical choice RT data. The LBA assumes that evidence accumulates for each response at a fixed, linear rate (the drift rate) until evidence in one accumulator reaches a response threshold. The model assumes two sources of between-trial variability – in the start point of evidence and in drift rate. The

mean of the drift rate distribution is the rate of processing parameter in the LBA. Response caution in the LBA is a function of the difference between the response threshold and the maximum position from which evidence accumulation in any accumulator could possibly begin. Brown and Heathcote (2008) demonstrated, quite surprisingly, that despite not including leaky or self-exciting accumulation, competition between accumulators, between-trial variability in non-decision time, and moment-to-moment variability in evidence accumulation, the LBA is capable of accounting for the shape of RT distributions, the speed-accuracy trade-off, as well as the relative speed of errors.

Despite their differences in architectures, most of the choice RT models just described have parameters which share an interpretation about the processes underlying simple decisions. These models are often applied to data in order to better understand the effect of an experimental manipulation. Which parameters must vary across the different conditions of the manipulation in order to explain the observed data gives insight into which aspects of the decision process are thus affected. An example may make this clearer, imagine we are interested in whether how easily a word can be identified is influenced by its frequency of use, and so we do a lexical decision task using low and high frequency stimuli. This results in accuracy and RT data which differ for low and high frequency words. Which parameters of a choice RT model need to differ between low and high frequency conditions to explain the observed differences in RT distributions reveal which latent variables (e.g. response caution) are affected by the frequency of a word.

The general approach of using the parameters of quantitative models to describe differences that underlie empirical data has been dubbed “cognitive psychometrics”

(Batchelder, 1998; Batchelder & Riefer, 1999; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). Choice RT models have been used extensively for this purpose (see below for some examples), and the popularity of this particular use of choice RT models seems to be increasing. Indeed, one of the fundamental aims of this thesis is to help provide the tools needed to carry out cognitive psychometrics with choice RT models. For example, Chapter One contains software for collecting data suitable for choice RT-based cognitive psychometrics (Donkin, Brown & Heathcote, 2009a).

Increasing Availability

In recent years, with the benefits of cognitive psychometrics becoming more apparent to those outside the field of quantitative psychology, there have been several valiant efforts to make the model fitting process more accessible. Some early attempts included written guides and tutorials on fitting RT distributions (Ratcliff & Tuerlinckx, 2002; Smith, 2000; Van Zandt, 2000). Taking a slightly different approach, Wagenmakers et al. (2007) offered the EZ-diffusion model as a simple way to estimate parameters for a choice RT model. Wagenmakers et al. provided relatively simple formulae that transform mean RT, variance of RT and the proportion of correct responses into estimates of the drift rate, response threshold and non-decision time of a Wiener diffusion process. To calculate these parameters Wagenmakers et al. assume that the EZ-diffusion model must be unbiased and have no between-trial variability (i.e. in drift rate, start point or non-decision time). Such a simplification means that the model no longer gives a full account of benchmark choice RT data. In practice, however, this cost is offset by the fact that researchers in applied areas outside of quantitative psychology benefit greatly from being able to model their data using relatively simple calculations which require no iterated fitting.

As Wagenmakers et al. (2007), Ratcliff (2008) and Wagenmakers, van der Maas, Dolan, and Grasman (2008) discuss, there are some unwanted downsides to having the EZ-diffusion model be so much simpler than the full Ratcliff diffusion model. Wagenmakers et al. (2008) and Grasman, Wagenmakers, and van der Maas (2009) addressed some of these issues in their respective versions of the model, Robust-EZ and EZ2. In particular, the Robust-EZ version of the model answered the complaint made by Ratcliff that the parameters returned by the EZ-diffusion model were sensitive to the existence of outliers. Wagenmakers et al. (2008) made the EZ-diffusion model into a mixture model which modelled contaminant responses, which allowed for clean parameter estimation. Grasman et al. (2009) provided the EZ2 model which models response bias. The EZ2 fitting software also allows for parameters to be fixed as constant across experimental conditions – something missing from the original EZ-diffusion model which forced the experimenter to estimate all parameters across all experimental conditions. Importantly, all of these changes came without the cost of a large increase in the complexity of the fitting software. Moreover, the EZ2 software remains very simple to use and fits are generally quick enough to be done using an interactive web page. Indeed, as Wagenmakers et al. (2008) point out, the EZ-diffusion and its subsequent improvements exist simply to make choice RT models more accessible to a wider audience.

Around the same time as the EZ-diffusion model became available, software which made it easier to use the full Ratcliff diffusion model also began to appear: DMAT (Vandekerckhove & Tuerlinckx, 2007, 2008) and fast-DM (Voss & Voss, 2007, 2008). Vandekerckhove, Tuerlinckx, and Lee (2008) also offered a hierarchical diffusion model (HDM), and included code which used Markov Chain Monte Carlo Bayesian

sampling methods for estimation of diffusion model parameters. DMAT is a toolbox within Matlab, fast-DM is a set of platform-independent C code, and HDM uses WinBUGS. The three methods for parameter estimation differ in their implementation, but their aim is shared – to help make the fitting of choice RT models more accessible to those without firm grounding in choice RT modelling, a group which have historically avoided such techniques because of their prohibitive difficulty.

Chapter Two of this thesis contains software similar to the aforementioned packages for the diffusion model, but using Brown and Heathcote's (2008) LBA model (Donkin, Averell, Brown & Heathcote, 2009). The LBA is similar to the EZ-diffusion model in that makes relatively few assumptions regarding how evidence accumulates – that it is linear and independent between accumulators with only two sources of between-trial variability. Unlike the EZ-diffusion model, however, the LBA accounts for the full range of benchmark choice RT data. The simplification of the model also leads to analytic solutions for the likelihood of the time taken for the first accumulator to reach threshold. Having a full model of choice RT whose likelihood can be expressed in closed form - including all forms of assumed between-trial variability - is essentially unique to the LBA model, and makes the model relatively simple to apply to data. For example, in Chapter Two there are methods for applying the LBA to data using a Microsoft Excel spreadsheet.

Application of Choice Response Time Models

Cognitive Psychometrics

The increasing availability of methods for fitting choice RT models means that cognitive psychometric applications have also become increasingly popular. This is particularly evident when we look at the range of experimental paradigms to which

cognitive psychometrics with choice RT models have been applied. For example, conclusions using choice RT models have been drawn in simple discrimination tasks (Usher & McClelland, 2001; Ratcliff, 2002), lexical decision tasks (Grasman, Wagenmakers, & van der Maas, 2009; Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, et al., 2008), recognition memory (Ratcliff, 1978), the implicit association test (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007), the accessory stimulus effect (Jepma, Wagenmakers, Band, & Nieuwenhuis, 2009), and the effects of practice (Dutilh, Wagenmakers, Vandekerckhove, & Tuerlinckx, 2009).

The approach of using the parameters of choice RT models to describe differences between experimental conditions has also been extended to make comparisons between different groups performing the same task. For example, to investigate the effects of aging (Ratcliff, Thapar, & McKoon, 2003, 2004, 2007; Ratcliff, Thapar, Gomez, & McKoon, 2004), the effects of depression and anxiety (White, Ratcliff, Vasey, & McKoon, 2009, in press), the differences between normal and impaired readers (Ratcliff, Perea, Colangelo, & Buchanan, 2004), and the effects of IQ on performance (Ratcliff, Schmiedek, & McKoon, 2008).

Ratcliff et al.'s (2004) application of the diffusion model to the lexical decision task is a particularly interesting example of how the application of a choice RT model can reveal information about processes underlying a particular paradigm. It is also relevant to this thesis because Chapter 3 contains a re-analysis of some of the data from Ratcliff et al.. In the lexical decision task a participant is presented with a string of letters and asked to respond with whether the string forms a "word" or a "non-word". A diffusion model explanation of how the lexical decision task is performed is as follows: Suppose a participant is presented with CHAIR as a stimulus. The participant perceives

and encodes the stimulus, then searches their lexicon for the stimulus. These processes are assumed to make up the pre-decision component of non-decision time (with button press making up the majority of post-decision T_{er}). Evidence coming from the lexicon as to whether or not the stimulus is a word is then accumulated until there is enough evidence for either a "word" or "non-word" response. In the current example, since CHAIR is a frequently used word, we expect that evidence will race quickly towards the "word" response threshold. If the stimulus presented were CHEIR, however, then we would expect that evidence will on average head towards the "non-word" response threshold.

Ratcliff, Gomez, and McKoon (2004), across nine experiments, manipulated the frequency of use of the word (across high, low and very low frequencies) as well as whether the stimulus presented was a word or a non-word. Consistent with the usual findings, high frequency words were identified faster and more accurately than low and very low frequency words. This difference is generally taken to indicate that lower frequency words are more difficult, and hence take longer, to find in a person's lexicon. Based on Ratcliff et al.'s use of the diffusion model, however, it was concluded that changes in word frequency could be entirely attributed to changes in drift rate, and not other parameters such as T_{er}. A larger drift rate for words of higher frequency suggests that when the word was of higher frequency, the lexicon produced more evidence to indicate that the stimulus was a word than when a lower frequency stimulus was presented. This finding led Ratcliff, Gomez, and McKoon (2004) to conclude that the word frequency effect is not due to differences in time taken to retrieve words from the lexicon, as this would have meant that T_{er} would be affected by frequency, but instead on the basis of how like a word the stimulus is. In particular, they concluded that lower

frequency words provide less evidence that they are words and hence have lower "wordness" values.

Fitting a choice RT model requires that several assumptions be made. Section Two of this thesis deals with the effects these assumptions can have on the conclusions drawn from choice RT models. In Chapter Three – Donkin, Heathcote, Brown, and Andrews (2009) question the assumption that drift rate is the only parameter of the diffusion model which changes across word frequency conditions. We conclude that a different assumption is more appropriate – that non-decision time also varies across word frequency conditions. This alternative assumption leads to a very different interpretation of the effects of a word frequency manipulation in the lexical decision task. In Chapter Four – Donkin, Brown and Heathcote (2009) show that an assumption which tends to be given little consideration when fitting choice RT models can have relatively large implications for both the predictions and the conclusions that the models provide.

Other Applications

Choice RT models have a number of applications other than just in cognitive psychometrics. For example, in the last couple of years choice RT models have been used to help identify signals in imaging studies (e.g. Forstmann et al., 2008; Ho, Brown, & Serences, 2009). Choice RT models have also been implemented in larger theories as a model of the decision process. For example, Smith and Ratcliff (2009) proposed the visual short-term memory (VSTM) model of cued signal detection. The VSTM theory accounts for the effects of cueing, masking and uncertainty on accuracy and RT distributions using a set of non-linear inputs fed dynamically into a diffusion model (Wiener and OU) decision model. In other words, the choice RT model - in this case, the

diffusion model - acts as a model of how decisions are made based on a stream of evidence which comes directly from the VSTM system.

In Chapter Five of this thesis Brown, Marley, Donkin, and Heathcote (2008) use a choice RT model as the decision stage in a complete theory of absolute identification. In absolute identification participants are tasked with deciding which particular stimulus, from a set of *N* stimuli that vary on only one dimension, has been presented on any one trial. Brown et al (2008) provide the Selective Attention Mapping Ballistic Accumulator (SAMBA) model as a complete account of both choices and RTs in absolute identification. Like the VSTM, SAMBA contains a detailed theory of how the evidence for all *N* responses are produced. These evidence values are fed into a choice RT model which makes a decision regarding which stimulus has been presented. In Chapter Six we test a prediction which arises as a result of the architecture of SAMBA. In particular, the spacing of stimuli in SAMBA has a large effect on RT but very little effect on accuracy. Empirical results from Lacouture (1997) support this prediction, and the model accounts for data across several experiments not with free parameters, but as a natural consequence of its framework.

Overview and Discussion of Main Body

The main body of the thesis is a collection of six papers which together seek to impress upon the reader the importance, availability and potential benefits gained from a quantitative, model-based approach to the combined analysis of accuracy and RT data (see Table 1 for a full list of the papers). There are three sections: the first (Chapters 1 and 2) offers techniques for the collection and quantitative modelling of choice RT data. The second section (Chapters 3 and 4) concerns the assumptions made when using models of choice RT. Finally, the third section (Chapters 5 and 6) demonstrates the use

of a choice RT model in a complete theory of absolute identification, and solidifies the theme of the thesis with a demonstration of the importance of a combined quantitative approach to accuracy and RT data within the absolute identification paradigm. Note that because the thesis is a collection of published papers, there is a reference section after each chapter (including this introduction).

Section One: Methods and Quantitative Techniques

The aim of the first section of the thesis is to offer computer-based methods for collecting and analysing choice RT data. The section is broken into two papers, Chapter One – Donkin et al. (2009a) "ChoiceKey: A real-time speech recognition program for psychology experiments with a small response set", provides a method for collecting choice and RT data for psychology experiments involving rapid choices. Voice key software has traditionally been used in experiments where the population of participants are unable to use a vocal mode of response over the more standard button press, or in cases where vocal responses are preferred. The use of a voice key, however, requires that the choice made is either ignored, or that responses are manually coded by an experimenter. ChoiceKey removes this extra labour by recording the RT while simultaneously recognising the response made from a pre-defined set of possible responses. Simply put, ChoiceKey collects accuracy and RT data from verbal responses.

Chapter Two – Donkin, Averell et al. (2009) "Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator model", offers methods for analysing choice and RT data which go beyond standard approaches such as an Analysis of Variance on mean RT or the proportion of correct responses. Choice RT models such as the LBA (Brown & Heathcote, 2008), when applied to accuracy and RT data, give a numerical summary of the latent variables underlying simple decisions.

For example, the application of a choice RT model can give information about how much information a participant is extracting from the environment, how cautious a participant is in their responding, as well as the time taken for the non-decision aspects of RT (e.g how long it takes to execute the motor-response). Additionally, since the estimation of these latent variables is based on both the choice made as well as the time taken to make that choice, it is not affected by the potential trade-off between the speed and accuracy of a response. This second chapter offers methods for applying the LBA to choice and RT data using a range of techniques (including Microsoft Excel, the statistical language R, and the Bayesian sampling software WinBUGS).

Table 1 Publications making up my thesis, including corresponding chapter and section references.

Section	Chapter	Reference
1	1	Donkin, C., Brown, S., & Heathcote, A. (2009). ChoiceKey: A real time speech recognition program for psychology experiments with a small response set. *Behavior Research Methods, 41*, 154-162.
	2	Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator. *Behavior Research Methods, 41,* 1095- 1110.
2	3	Donkin, C., Heathcote, A., Brown, S., & Andrews, S. (2009). Nondecision time effects in the lexical decision task. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.
	4	Donkin, C., Brown, S., & Heathcote, A. (2009b). The over-constraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16*, 11291135.
3	5	Brown, S., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115*, 396425.
	6	Donkin, C., Brown, S., Heathcote, A., & Marley, A. A. J. (2009). Dissociating speed and accuracy in absolute identification: The effect of

Section Two: Applications and Assumptions

Many assumptions are made in any application of a choice RT model to accuracy and RT data. Each of these assumptions requires a considered decision on the part of the researcher. This is especially important as the use of choice RT models to analyse data continues to expand. Section Two contains two papers which deal with assumptions about the way that experimental manipulations can influence the parameters of choice RT models.

In particular, Chapter Three – Donkin, Heathcote, et al. (2009) "Non-decision time effects in the lexical decision task" is an example of how changes in these assumptions can have large ramifications for the paradigm under investigation. The paper involves a re-analysis of lexical decision task data from Ratcliff, Gomez, and McKoon (2004), highlighting a systematic problem with the diffusion model's account of the word frequency effect. This misfit is corrected by adding the additional assumption that the time taken to retrieve a word from the lexicon is affected by that word's frequency. Such inference about the processes underlying the lexical decision task would not have been possible without a combined, quantitative analysis of accuracy and RT data.

In Chapter Four – Donkin et al. (2009b) "The over-constraint of response time models: Rethinking the scaling problem" my co-authors and I suggest that one particular assumption about how manipulations can influence parameters has been largely overlooked. The chapter shows that this assumption is not benign and like other assumptions it causes changes in the predictions and psychological interpretations of the

model. The paper encourages the need for the appropriate and considered use of choice RT models as a tool for extracting information from RT and accuracy data.

Section Three: A Psychological Process Model – Absolute Identification

The third section of the thesis is concerned with the importance of both RT and accuracy in fully understanding a particular paradigm called absolute identification. Absolute identification is particularly relevant to the overarching aim of the thesis since historically, very little attention has been paid to RT phenomena: few RT data exist, and prior to Brown, Marley, et al. (2008) a complete theoretical account of both accuracy and RT phenomena was missing. It was common opinion amongst absolute identification researchers that RT phenomena were relatively uninteresting as they simply mirrored accuracy phenomena – accurate responses were fast and inaccurate responses were slow, so it was sufficient to consider just response accuracy, trusting that RT would take care of itself. The two papers making up this final section of the thesis are extensions of the more general quantitative approach to accuracy and RT contained in the previous two sections.

In Chapter Five – Brown, Marley, et al. (2008) "An Integrated Model of Choices and Response Times in Absolute Identification" a choice RT model is used as the decision stage in a process model of the absolute identification task. Much of the SAMBA model of absolute identification presented in Brown et al. (2008) is concerned with the development of the inputs to this decision process. The use of the Ballistic Accumulator (Brown & Heathcote, 2005) as a model of how decisions are made in absolute identification tasks allows SAMBA to account for the wide range of benchmark phenomena observed in both accuracy and RT. The paper presented in Chapter Five gives a detailed overview of the SAMBA model and demonstrates its

account of benchmark accuracy and RT data.

In Chapter Six – Donkin, Brown, et al. (2009) "Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing", my co-authors and I test a prediction from SAMBA – that there is a dissociation between accuracy and RT when the spacing between stimuli is varied. Donkin, Brown, et al. (2009) present data from an experiment in Lacouture (1997) in which increasing the spacing between stimuli caused an increase in accuracy without a corresponding decrease in RT, confirming SAMBA's prediction. This paper highlights the importance of consideration of both RT and accuracy in absolute identification, since they are not a simple transformation of each other. It also presents a strong challenge to any alternate theoretical account of the task, especially since SAMBA accounts for the dissociation as a natural result of its architecture and not via free parameters. The prediction is yet another example, and hopefully cements in the reader's mind, the importance of first considering, and then quantitatively modelling, both accuracy and choice RT data.

References

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment* , *10* , 331–344.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6* , 57–86.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112* , 117-128.

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57* , 153-178.

Brown, S., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115* , 396–425.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23* , 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making. *Psychological Review, 100* , 432–459.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41* , 260–274.

Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods*, *41* , 1095–1110.

Donkin, C., Brown, S., & Heathcote, A. (2009a). Choicekey: A real-time speech

recognition program for psychology experiments with a small response set. *Behavior Research Methods*, *41* , 154–162.

Donkin, C., Brown, S., & Heathcote, A. (2009b). The over-constraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16* , 1129–1135.

Donkin, C., Brown, S., Heathcote, A., & Marley, A. A. J. (2009). Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research*, *73* , 308–316.

Donkin, C., Heathcote, A., Brown, S., & Andrews, S. (2009). Non-decision time effects in the lexical decision task. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Dutilh, G., Wagenmakers, E.-J., Vandekerckhove, J., & Tuerlinckx, F. (2009). A diffusion model account of practice. *Psychonomic Bulletin & Review*, *16* , 1026–1036.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). The striatum facilitates decision-making under time pressure. *Proceedings of the National Academy of Science*, *105* , 17538–17542.

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology.*

Ho, T., Brown, S., & Serences, J. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29* , 8675–8687.

Jepma, M., Wagenmakers, E.-J., Band, G. P. H., & Nieuwenhuis, S. (2009). The effects of accessory stimuli on information processing: Evidence from electrophysiology and a diffusion model analysis. *Journal of Cognitive Neuroscience*, *21* , 847–864.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology*, *34* , 1–53.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion–model analysis. *Journal of Personality and Social Psychology*, *93* , 353–368.

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27* , 375–396.

LaBerge, D. A. (1994). Quantitative models of attention and response processes in shape identification tasks. *Journal of Mathematical Psychology*, *38* , 198–243.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: a response-time analysis. *Psychological Research*, *60* , 121–133.

Laming, D. R. J. (1968). *Information theory of choice–reaction times*. London: Academic Press.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40* , 77–105.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

McClelland, J. L. (1991). Stochastic interactive activation and the effect of context on perception. *Cognitive Psychology*, *23* , 1–44.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General,*

118, 346-373.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology General* , *135* , 391–408.

Otter, T., Allenby, G. M., & Van Zandt, T. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research*, *45* , 593–607.

Pachella, R. G. (1974). The interpretation of reaction time in information–processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale (NJ): Lawrence Erlbaum Associates.

Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information-processing mechanism. *The Journal of General Psychology*, *124* , 231–339.

Pike, A. R. (1966). Stochastic models of choice behaviour: Response probabilities and latencies of finite Markov chain systems. *British Journal of Mathematical and Statistical Psychology*, *21* , 161–182.

Pike, A. R. (1973). Response latency models for signal detection. *Psychological Review*, *80* , 53–68.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85* , 59–108.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9* , 278–291.

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin & Review*, *15* , 1218–1228.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review*, *111* , 159–182.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition*, *55* , 374-382.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science*, *9* , 347–356.

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, *36* , 10-17.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two–choice reaction time. *Psychological Review*, *111* , 333–367.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical–decision task. *Psychology and Aging*, *19* , 278–289.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65* , 523–535.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50* , 408–424.

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75-90 years old. *Psychology and Aging*, *22* , 56–66.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9* , 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261-300.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment* , *14* , 184–201.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multi–alternative decision field theory: A dynamic artificial neural network model of decision–making. *Psychological Review*, *108* , 370–392.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica*, *27* , 143–153.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44* , 408–463.

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116* , 283–317.

Smith, P. L., & Van Zandt, T. (2002). Time–dependent poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, *53* , 293–315.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two–choice discrimination. *Journal of Mathematical Psychology*, *32* , 135–168.

Smith, P. L., & Vickers, D. (1989). Modeling evidence accumlation with partial loss in expanded judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *15* , 797–815.

Stone, M. (1960). Models for choice–reaction time. *Psychometrika*, *25* , 251–260.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. London: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108* , 550–592.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14* , 1011-1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40* , 61-72.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). Hierarchical Bayesian diffusion models for two–choice response times. *Manuscript submitted for publication*.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7* , 424-465.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7* , 208-256.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13* , 37–58.

Vickers, D. (1979). *Decision processes in visual perception*. London: Academic Press.

Voss, A., & Voss, J. (2007). Fast–dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39* , 767–775.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52* , 1–9.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff

diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21* , 641–671.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58* , 140–159.

Wagenmakers, E.-J., van der Maas, H. J. L., Dolan, C., & Grasman, R. P. P. P. (2008). Ez does it! extensions of the ez-diffusion model. *Psychonomic Bulletin & Review*, *15,* 1229–1235.

Wagenmakers, E.-J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ–diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14* , 3–22.

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion*, *23* , 181-205.

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (in press). Sequential sampling models and psychopathology: Anxiety and reaction to errors. *Journal of Mathematical Psychology.*

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41* , 67–85.

ChoiceKey: A real-time speech recognition program for psychology experiments with a small response set.

Christopher Donkin, Scott D. Brown, & Andrew Heathcote

The University of Newcastle
Newcastle, Australia

Abstract

Psychological experiments often collect choice responses using button presses. However, spoken responses are useful in many cases, for example: when working with special clinical populations, or when a paradigm demands vocalization, or when accurate response time measurements are desired. In these cases, spoken responses are typically collected using a voice key, which usually involves manual coding by experimenters in a tedious and error-prone manner. We describe an open source speech recognition package for MATLAB, ChoiceKey, which can be optimized by training for small response sets and different speakers. We show ChoiceKey to be reliable with minimal training for most participants in experiments with two different responses. Problems presented by individual differences, and occasional atypical responses, are examined, and extensions to larger response sets are explored.

Many psychology experiments require participants to complete hundreds of trials using a small response set. For example, memory experiments often require participants to respond with only 'old' or 'new' (e.g., Rubin, Hinton & Wenzel, 1999) and choice response time (RT) tasks often require participant to respond make one of two responses, such as 'one' and 'two' or 'high' and 'low' (e.g., Ratcliff & Rouder, 1998). The typical method for collecting responses for this type of experiment is a button press, usually via a keyboard, mouse, specially developed button-box, or a touch screen. However, there are a number of reasons that an experimenter might instead prefer to collect spoken responses. For these cases we offer an open source speech recognition package, ChoiceKey. In the following we show that ChoiceKey reliability identifies a small number of response alternatives, and that it gives precise estimates of vocal RT, but first we discuss a few of the circumstances in which one might prefer vocal responses over alternative data collection methods.

Response time measurement is an important aspect of many psychology experiments, and the precision and accuracy of RT estimates from different response tools has been documented extensively in this journal. Keyboard responses are often imprecise due to buffering issues (Plant, Hammond & Turner, 2004; Shimizu, 2002; Voss, Leonhart & Stahl, 2007), as are mouse-button clicks (Beringer, 1992; Crosbie, 1990; Plant, Hammond & Whitehouse, 2003). Precise RT measurements can be obtained using button-boxes connected via the PC parallel port (e.g., Stewart, 2006; Voss et al.), but these solutions require specialised hardware, which can be expensive and is not always well supported. We show that ChoiceKey also yields precise measurements of RT, with the advantages that it is simple to set up and inexpensive, as it requires only a microphone and a sound card, which are now standard equipment for

most PCs.

Aside from RT measurements, making responses via buttons can be problematic because it requires the participant to learn a response-to-button mapping. Although some of these mappings are relatively natural, such as ‘left’ and ‘right’ using the left and right arrows of the keyboard, other response sets have no intuitive button mapping. For example, Rubin et al. (2004) mapped the responses ‘old’ and ‘new’ to keys chosen by the experimenters, and participants had to learn this mapping and maintain it throughout the experiment. If participants were able to speak aloud the responses ‘old’ or ‘new’, the learning of this mapping could be avoided.

Experimental research with clinical populations who are unable to give manual responses via a button press might also benefit from the ability to easily collect spoken responses and the associated RTs. In particular, an automatic speech recognition program might benefit experimenters working with people with schizophrenia, people with intellectual disabilities, or people with psychomotor disabilities. Trewin and Pain (1999) have shown that people with these types of psychological and/or motor disabilities display a large number and wide range of errors when using a mouse or keyboard. The use of spoken responses may help avoid some of these measurement errors.

Even if participants are able to give manual responses, Vidulich and Wickens (1985) show that when central processing is required for a task that is verbally oriented spoken responses are most appropriate. It is also possible that the need for spoken responses is implicit given the experimental procedure or paradigm being used, such as in Stroop-like tasks that investigate the cause of interference due to response modality (Simon & Sudalaimuthu, 1979; Wang & Proctor, 1996). In these situations ChoiceKey

offers a way of both identifying response and recording RT.

Amongst others, Lacouture and Marley (2004) allowed participants to respond vocally in an absolute identification experiment. Participants gave spoken responses via microphone and reaction times were obtained using a "voice key", which is a device that measures RT by measuring the time sound energy crosses a threshold. However, a voice key requires response choices to be manually coded by an experimenter, a task which is inevitably time consuming, tedious, and error-prone. Speech recognition software can help alleviate these problems.

Speech recognition software is in common use; most people have had the experience of ordering a pizza, or giving personal details, by speaking into a phone. However, the accuracy of these systems can be far from perfect, and is unlikely to be acceptable for experimental measurement. Microsoft Windows and Macintosh OS X both come with inbuilt speech recognition functionality that can be adapted by training to individual users' voices. However, we found these inbuilt speech recognition packages to be far too inaccurate for use in experiments, even under ideal conditions with only two different responses and extensive training. This is likely because the programs are intended to recognise a very large number of different responses in environments where the cost of an incorrect recognition event is low.

As an alternative, we offer an exemplar-based speech recognition program that is trained exclusively to recognise only those responses that are to be used in an experiment. The program, ChoiceKey, is an open source library for the software package MATLAB that can be called by a MATLAB script that controls the experiment. Appendix A outlines an example script for a simple experiment where one of two stimuli is presented, and the participant is required to name it. ChoiceKey was

developed using the Data Collection and Voicebox toolboxes under MATLAB v7.5.0 (R2007b) and Reynolds, Quatieri and Dunn's (2000) Gaussian mixture models for speaker identification. Details about the contents of the ChoiceKey library are outlined in Appendix B.

The underlying structure of ChoiceKey is based on leading models of speaker verification. Bimbot, Bonastre, Fredouille, Gravier, Magrin-Chagnolleau, et al. (2004) offer a detailed and complete discussion of the extensive work in this area. A graphical summary of how ChoiceKey works is presented in Figure 1. Sound card outputs are captured using the inbuilt MATLAB Data Collection Toolbox. Audio capture begins when input from a microphone reaches a threshold, and terminates one and a half seconds later. Both the threshold and recording time can be altered by the user. The recorded data are first passed through front-end processing, transforming the time-varying amplitude input from the microphone into a set of features represented as a vector of "cepstral" coefficients. These feature vectors are then modelled statistically to create a set of training exemplar models. Later, during the experiment, participants' responses are turned into feature vectors and the likelihood that these vectors came from each training exemplar model is calculated. The most likely model is the chosen response. We now discuss each of these aspects in more detail.

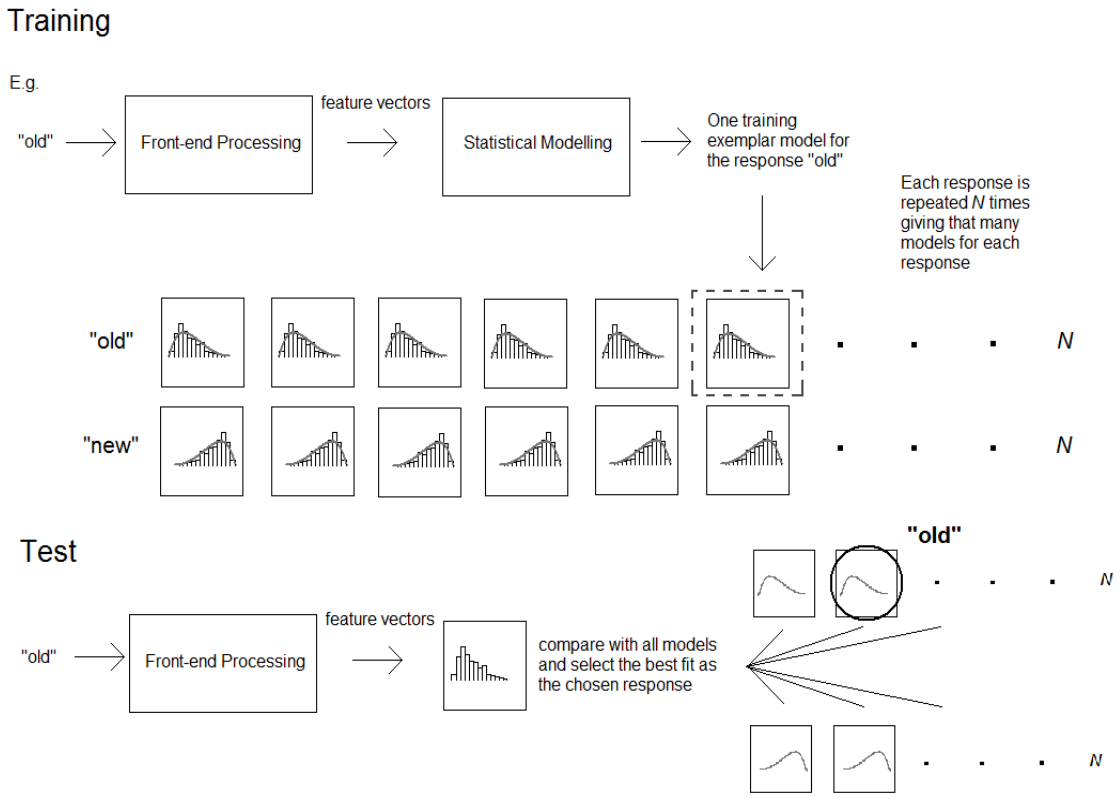

Figure 1 Graphical representation of ChoiceKey's operation. During training, the response "old" is given on a particular trial. Features are extracted and a statistical model is created for that training exemplar. The model is then stored with the rest of the models created in training. During testing, the participant speaks the word "old" during one of the trials of the experiment. The features of the word are extracted and then compared against all the models created at training. The most likely model is chosen, which happens to be one of the trained exemplars for the response "old", so ChoiceKey chooses that response.

Front-end Processing

In the front-end processing stage, the sound input is broken up into 20ms windows using a 10ms frame-rate, ensuring 50% overlap between segments. Only those windows containing enough sound energy to be considered not silent are kept. This is done relative to the noise in the signal so as to lower the probability that speech is discarded. The Mel-scale cepstral feature vectors are then calculated for each of the 20ms windows. This is done by first taking the fast Fourier transform of the speech segment. The resulting spectrum is smoothed using a series of band-pass frequency filters which are convolved with the spectrum to get an average value for each

frequency band. These filters are spaced on the Mel scale, which has the property of being close to the frequency scale of the human ear (Stevens, Volkman & Newman, 1937). A discrete cosine transformation is applied to the log of the values produced by the frequency filters to yield cepstral coefficients.

Reynolds, Quatieri and Dunn (2000) suggest that all but the 0^{th} cepstral coefficient are best used in speaker recognition. However, for ChoiceKey we desire speech, not speaker, recognition, and we have achieved greater accuracy by retaining the 0^{th} coefficient. Reynolds et al. also suggest a number of normalisation transformations be made to compensate for mismatched microphone conditions between training and testing. We did not use these transformations as we assumed that, because training ChoiceKey typically takes less than five minutes, training and testing would be done under identical conditions.

Statistical Modelling

The cepstral coefficients are modelled using Gaussian mixture models (GMM), which have been shown to be successful in the domain of speaker recognition (Reynolds, 1992). GMMs have the desirable properties of being able to capture the behaviour of a distribution without assuming a very specific (e.g., Gaussian) form. They are also computationally simple, facilitating real-time processing. A GMM's density is the weighted linear combination of *M* Gaussian densities, each parameterized by a mean and variance term for each cepstral coefficient vector. The number of Gaussian densities used, *M*, can be altered by the user. During development we found that five Gaussian densities gave the best overall performance. However, individual differences in the optimal value of *M* did exist, so improved individual accuracy may be obtained by setting *M* based on an individual's data.

The Decision

During ChoiceKey training, the participant will speak aloud each of *k* response words *N* times. Each of these training words is modelled using a GMM, giving *N* exemplar models for each of the *k* responses at the end of training. On any particular trial of the experiment proper, the participant will make a new and unknown response. The Mel-scale cepstral feature vectors are calculated for this new response. ChoiceKey then calculates the log-likelihood of observing the cepstral feature vectors given the parameters of the GMM for each of the *N* exemplars and *k* responses. The exemplar with the largest log-likelihood is selected as the given response.

During development we tried a range of alternative response selection rules, such as selecting the response set with the largest summed log-likelihood across all exemplars, and more sophisticated classifiers, such as backpropogation neural networks and support vector machines. For the small response-set sizes in our experiments the more sophisticated selection rules did not provide any benefit, but this may not be the case for larger response-set sizes. The simple selection rule used by ChoiceKey has the advantage of reduced computational cost, particularly during training. In other settings, ChoiceKey can be easily adapted by the user to implement alternative training and decision algorithms.

Using ChoiceKey

A typical experiment using ChoiceKey involves a short training session (less than five minutes), where participants speak aloud the words in the response set (e.g. old/new) a number of times (typically between 10 and 30), training ChoiceKey to identify their voice. The experiment then proceeds as normal, with responses made by voice, using ChoiceKey to return the response that it calculates to be the one most likely

spoken by the participant, and the response time (RT). We now report the results of experiments examining the accuracy of these measurements. The first experiment investigates the accuracy of the RT measured by ChoiceKey, by comparing it with reaction times manually measured from audio waveforms recorded in real time. In the second experiment we investigate how accurately ChoiceKey identifies responses from a variety of response sets.

Experiments

Response Time

An AMD computer with an AthlonXP 64-bit 2.33 Ghz processor and 2Gb of RAM, running Windows XP SP2 with a SoundBlaster Live! v5.10 sound card was set up to play a loud tone through external speakers. After the tone played, data capture within ChoiceKey was initialised. After an interval the word ‘respond’ appeared on screen and the participant spoke aloud the number ‘one’ into a headset-mounted SONY DR-220 microphone. The speaking of the word ‘one’ was intended to trigger recording, and for all 200 trials the triggering worked as required. The intervals between the tone and response were varied from 250 to 2000ms in intervals of 250ms, with each interval occurring 25 times, yielding 200 response times that spanned the range of RTs usually observed in simple psychological tasks.

Throughout the course of the experiment, a second laptop was set up nearby with its external microphone making a real-time recording of the entire proceedings under Adobe CS3 Soundbooth. This recording was later opened in Soundbooth as a waveform and the time between the tone and the “one” response was determined manually. This process gave us an accurate estimate of response time that should

correlate highly with ChoiceKey's RT measurement.

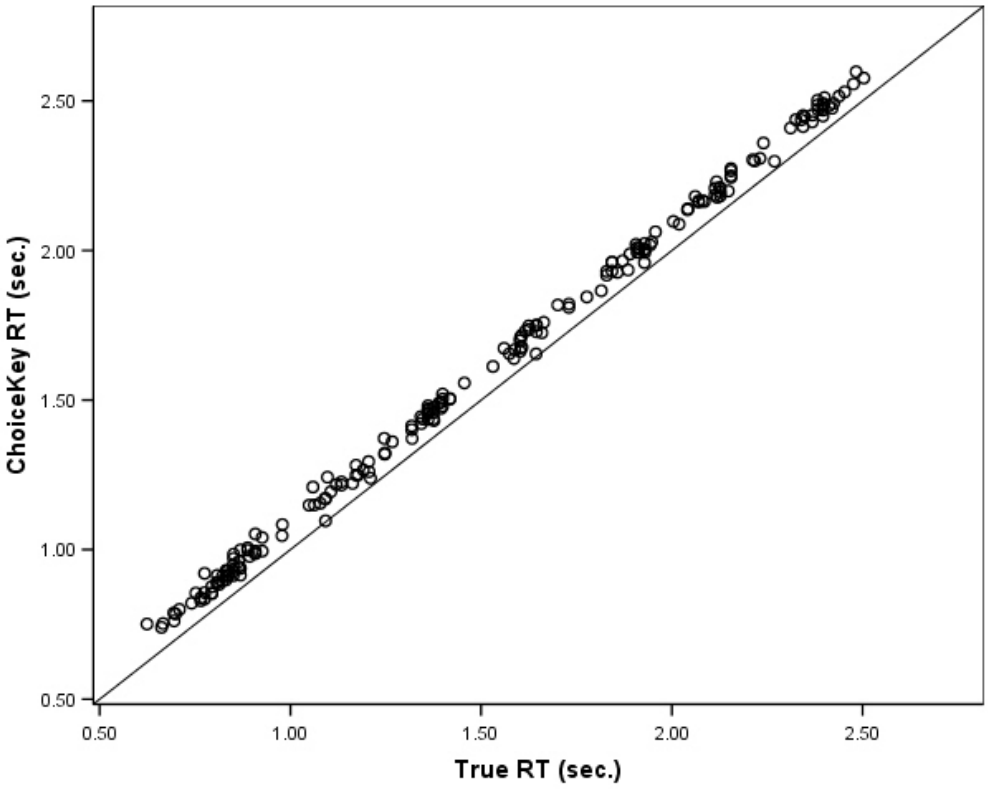

Figure 2 RT as recorded by ChoiceKey (y-axis) as a function of RT calculated manually from sound waveforms (x-axis). The solid diagonal line represents perfect measurement of RT. ChoiceKey gives a precise, but slightly biased estimate of RT.

Figure 2 shows the response times recorded by ChoiceKey plotted against response times derived from the waveform. The two measures of response time were highly correlated (r=.999). Response times from ChoiceKey were used as the response variable in a linear regression, with true response time used as the predictor variable. The slope of the regression line was 0.999, $t(199) = 344.83$, $p < .001$, and the intercept was 90ms $t(199) = 18.875$, $p < .001$, suggesting that ChoiceKey gives a precise, but slightly biased estimate of response time.

Rastle and Davis (2002) discuss biases in voice-key RT measurement as a function of the onset characteristics of different response waveforms. If experimenters

are concerned about obtaining absolutely unbiased measurements of RT, the above procedure can be carried out for all responses separately. More likely to be of concern to users of ChoiceKey, however, are differences between biases in RT measurement for different responses. Unless this issue is addressed, differences in RT may be attributed to differences between stimuli, when the real cause is differences in the time taken for ChoiceKey to trigger the onset of recording.

To address this issue without the time and effort required by manual scoring of waveforms, ChoiceKey includes a function called *callib.m*. On each trial, this function presents participants with a "+" sign and ask them to make one response repeatedly for a block of trials of length determined by the experimenter. The process is then repeated for each response. The function returns the mean RT for each response. Any differences in RT due to onset biases for the different responses can then be identified and corrected.

Identification

Identification accuracy was investigated using data from 24 participants who read aloud the following three sets: {1,2,3,4}, {old,new} and {high,low}. Participants were 12 male and 12 female first year psychology students. Words in each set were spoken 50 times in a random order and the order of each set was counterbalanced across subjects. Responses were collected using the same hardware used in the previous experiment. As envisaged for a standard experiment, the first 10 responses spoken by participants were used to train ChoiceKey. The remaining 40 responses were used to test ChoiceKey's identification accuracy. Results from the response set consisting of numbers one to four were partitioned into six different response sets of size two. These sets, along with {old,new} and {high,low} were used to test ChoiceKey's two-choice

identification accuracy. Table 1 shows the distributions of the number of errors out of the 80 identifications made by ChoiceKey.

Table 1 The number of errors (out of 80 identifications) made by ChoiceKey for individual participants for each response set. For example, the first cell indicates that for 16 out of 24 participants there were zero errors in identification for the word set {old,new}. The mean proportion of correct identifications for each response set is reported in the rightmost column.

		Number of errors (out of 80 identifications)					
		0	*1*	*2*	*3*	*>3*	*Mean Proportion Correct*
Response Set	*old,new*	16	3	3	2	0	.99
	2,4	15	5	1	1	2	.98
	1,2	14	5	0	3	2	.96
	1,3	11	6	1	4	2	.97
	3,4	12	4	4	3	1	.98
	1,4	9	2	2	9	2	.95
	high,low	6	4	5	8	1	.96
	2,3	6	2	4	7	5	.92

Very few identification errors were observed for the majority of participants and responses. For most response sets ChoiceKey made zero or one error out of 80 for almost all participants. For the response set {old,new}, ChoiceKey made no errors in identifying responses for 16 out of 24 participants. The response sets {2,4} and {1,2} were also identified with very few errors, with either zero or one error being made for 20 and 19 participants, respectively. The average accuracy of ChoiceKey's identification was highest for the response sets {old,new}, followed closely by {2,4} and {3,4}. Individual differences in identification far outweighed any differences observed as a function of age or gender.

The results of Table 1 indicate that with only 10 training exemplars ChoiceKey is able to perform well for some response sets, but that others are less discriminable. For example, in the response sets {high,low} and {2,3}, ChoiceKey was able to identify all

responses correctly for only a quarter of the participants. Not only were some response sets less discriminable, but even for the responses sets in which ChoiceKey *was* almost perfectly accurate for the majority of participants, a small proportion of participants remained whose responses were difficult to discriminate. For example, the response set {1,2} leads to either one or no errors for 19 out of 24 participants, suggesting it as a good candidate for use with ChoiceKey. However, for one participant 35 responses out of 80 were identified incorrectly. This suggests that with minimal training ChoiceKey is not viable for some participants.

One approach to solving this problem is to first screen participants based on a pre-experimental test of ChoiceKey's accuracy. The function *traintest* provides an estimate of ChoiceKey's identification accuracy. Interestingly, for the participant with very low accuracy for the response set {1,2} all other combinations of number responses accuracy were also low. However, no errors in identification were observed for the response set {high/low} for this participant, suggesting that speaker identification accuracy varies substantially as a function of response set.

An alternative approach to dealing with low identification accuracy, either for particular participants or particular response sets, is to use more than the ten training exemplars. To evaluate this strategy, 12 of the 24 participants completed an extra 50 responses for the word sets {old,new} and {high,low}. These extra responses were used to test the effect of varying the number of responses per word used in training ChoiceKey. The number of training exemplars used was varied from 1 to 30, always with the last 70 exemplars used to test ChoiceKey's accuracy.

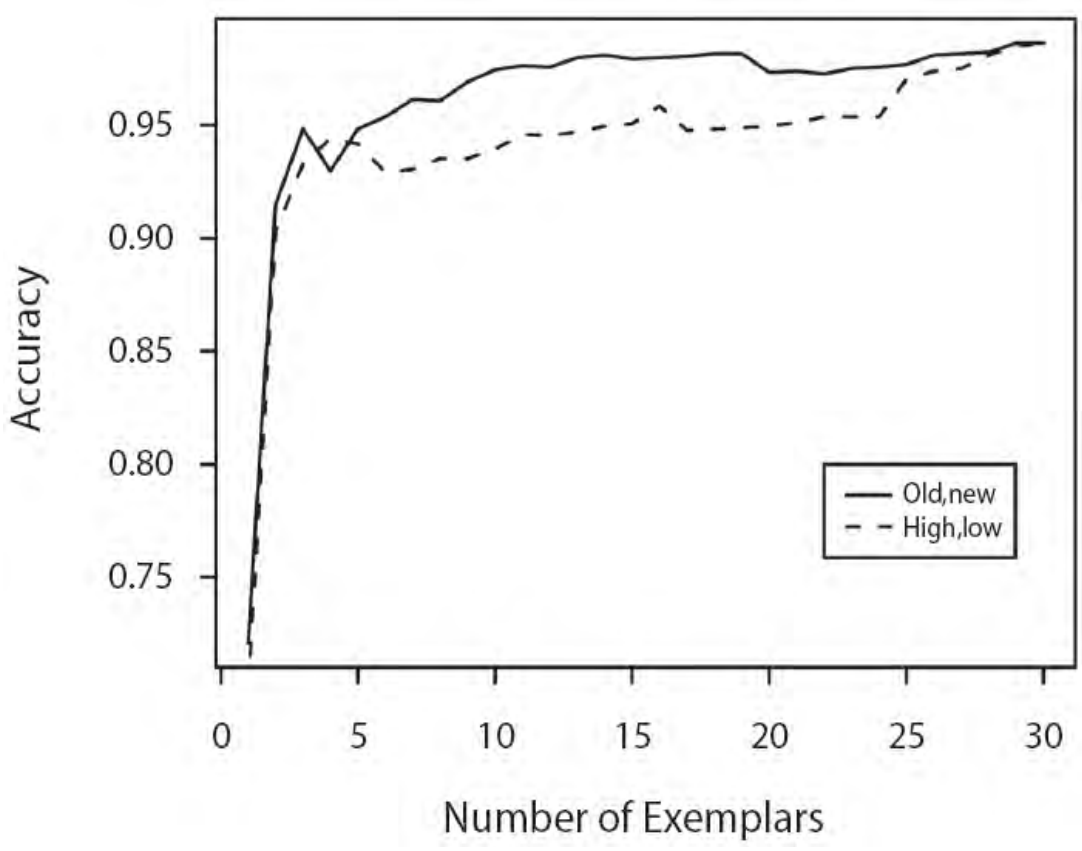

Figure 3. ChoiceKey identification accuracy averaged over participants and plotted as a function of number exemplars used in training. Accuracy was high for the response set {old,new} even when few training exemplars were used, but more exemplars were required for acceptable performance with the {high,low} response set.

Figure 3 shows the average percentage of accurate identifications made by ChoiceKey as a function of the number of training exemplars. For the response set {old,new}, increasing the number of training exemplars beyond 10 did not increase accuracy. This is likely due to a ceiling effect, as discrimination between the words was already close to perfect with 10 training exemplars. For the response set {high,low}, improvement was less rapid, but the same accuracy as for {old,new} was achieved with a set of 30 exemplars, suggesting that even difficult-to-discriminate word sets can be used with ChoiceKey, as long as sufficient training is provided.

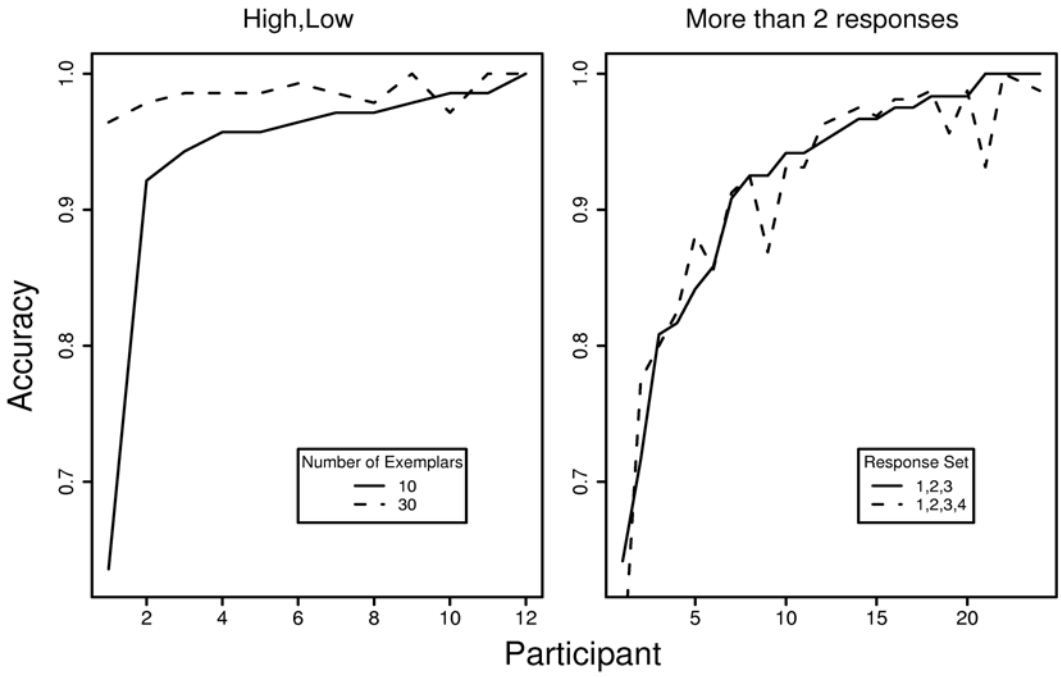

Figure 4. ChoiceKey recognition accuracy for individual participants for the response set {high/low} when 10 and 30 exemplars were used in training (left panel), and when the response sets {1,2,3} or {1,2,3,4} were used with 10 training exemplars. Participants are ordered according to accuracy in the 10 exemplar and {1,2,3} cases for left and right panels, respectively.

Increasing the number of training exemplars improved accuracy for all participants, even those who had very low accuracy with fewer training exemplars. The left panel of Figure 4 compares individual participant accuracy for the response set {high,low} with 10 and 30 training exemplars. Participants are ordered along the x-axis by accuracy for the 10 training exemplar case, and the same order is used for the 30 exemplar case to highlight individual improvement. For 23 out of 24 participants, accuracy either increased or remained perfect with the increase in training set size. Participants whose accuracies were lowest with 10 training exemplars showed the largest increase, bringing performance for almost all participants up to acceptable levels.

With only ten training exemplars, accuracy was much worse for responses sets of more than two words. The right panel of Figure 4 shows accuracy for individual

participants for the response sets {1,2,3} and {1,2,3,4} when ChoiceKey was trained with ten exemplars. Participants are ordered by accuracy for the smaller response set on the x-axis. Average accuracy was roughly equivalent for both response sets, although some participants were noticeably less accurate for the larger response set. Only a quarter of participants were more than 95% accurate, suggesting that a larger training set is required for the remaining participants.

These results suggest that experimenters who wish to use ChoiceKey for response sets larger than two should use the *traintest* function to screen participants and/or calibrate training set size to achieve the desired level of accuracy. The latter strategy should be used with caution, however, as we did not test whether larger set sizes display the same improvement in accuracy with training set size as we found with set size two.

A second limitation related to the use of larger set sizes is the associated computational cost, which increases as a polynomial function of both the response set size and the number of training exemplars. Sufficient time must be available between collecting the training data and commencing the experiment to process the training data. For example, when response set size was increased from two to four, the time taken to identify a single response, given 10 training exemplars, increased from 50 to 100 milliseconds. When 40 training exemplars were used, however, identifying a single response took 100 and 300 milliseconds for two and four responses, respectively.

Discussion

ChoiceKey is a measurement tool for the MATLAB environment that allows the collection of vocal choice responses. It is designed for use in experiments with a small number of possible responses. ChoiceKey provides precise estimates of vocal onset time, and can be easily calibrated to eliminate onset differences between responses

(Rastle & Davis, 2002). For a variety of common response pairs, it can reliably identify most participants' responses with high accuracy after only minimal training, such as 10 training exemplars per response, which takes only around two minutes for a binary choice task.

However, we found that some response pairs are identified with lower accuracy than others (e.g., {high, low}). Although experimenters could simply use response pairs which are reliably identified with high accuracy (e.g., {old,new}), doing so removes the potential benefit of reduced response learning offered by spoken responses. An alternative strategy is to use a larger training set, which improves accuracy. For example, with 30 exemplars accuracy was equally good for {old,new} and {high,low}.

A second issue is that identification accuracy is low for some participants, suggesting that there may be a need for pre-test screening of ChoiceKey's accuracy for each participant. This problem could be addressed through the use of more training exemplars, at least for the participants in our experiment. Increasing the number of responses used to train ChoiceKey from 10 to 30 not only increased identification accuracy for one of the response sets with the poorest performance, {high,low}, it also improved identification accuracy for those participants whose responses were most poorly identified when only 10 training exemplars were used.

Even after extended training, ChoiceKey did not perfectly identify all responses from all participants. These errors appeared to be asymptotic (i.e., they did not disappear with increased training). Such asymptotic errors are likely due to atypical responses, background noise (in the environment or in computer hardware), or both. We minimized the latter source of error by using a quiet testing environment and a high quality sound card and microphone. However, it is likely that even when background

noise and hardware errors are minimized, participants will sometimes say words in a way that was not encountered in training, causing misidentification. Fortunately, the proportion of such asymptotic errors in testing was low (around 1%).

It is arguable that this low error rate may not be too different from the rate of errors due to participants pressing the wrong response button, and it may even be less than the button-press error rate when the response mapping is unfamiliar or insufficiently practiced. Similarly, ChoiceKey's low error rate may be comparable to errors made by the experimenter manually coding responses in real-time. Where perfect vocal choice identification is required we recommend that responses be recorded and scored off-line. Even where an error rate of 1-2% is acceptable it may be prudent to record some responses and perform an off-line check of ChoiceKey's scoring as a quality control measure.

Apart from lowering background noise and using high quality hardware, we were not able to identify any other measures that reliably increased ChoiceKey's accuracy. For example, we were unable to identify an obvious reason why certain word pairs show lower discriminability than others. We therefore advise users that they choose the most natural response set for the task, and if necessary, increase the number of training exemplars until the desired level of identification accuracy is achieved. Similarly, there appears to be no clear pattern to the type of voice that ChoiceKey is able to identify with high accuracy (e.g., male vs. female voices). We suggest the same course of action – pick the most natural methodology and if pre-test screening shows a high error rate with any individual participant either use more training exemplars or exclude that participant's results from analysis.

Vocal responses are particularly advantageous with larger response sets, where

button responding is naturally more error prone due to the greater difficulty of learning a larger response mapping. Large response sets are also more likely to introduce differences in RT due to differences in response production time (e.g., differences between fingers). When more than 10 responses are required a unique finger cannot be assigned to each response, requiring either finger combinations, further increasing learning difficulty, or a movement response (e.g., moving a finger or mouse cursor from a "home" button to a response button). In both cases, RT variability and the potential for differences in production time are increased.

Unfortunately, we found that creating a speech recognition system which is highly accurate for large response sets is very difficult. When the response set was extended beyond two alternatives, we observed a large drop in accuracy, to about 90% on average for three or four different responses. We also found large individual differences, and that a majority of participants had low accuracy when ChoiceKey was trained with only 10 exemplars per response. These results indicate that ChoiceKey should be used cautiously with greater than two response alternatives, and that likely extended training will be required to obtain high accuracy. Fortunately, because ChoiceKey is open source and implemented in the flexible MATLAB language, users may easily explore such extensions.

Some directions for these future extensions have already been discussed. For example, it may be possible to optimise certain parameters affecting identification accuracy individually for each participant (e.g., the number of Gaussian density mixtures to use in the GMM). Future improvement may also lie in an alternative form of statistical modelling of the features of the recorded speech segment. Our use of GMM as a model of these features was based on their success in the field of text-

independent speaker recognition. Text-independent speaker identification involves the recognition of a voice irrespective of the spoken utterance. Text-dependent speaker identification involves recognising voices based on a particular set of spoken words. Hidden Markov models (HMMs) are often used to model the features of the spoken response in text-dependent speaker identification, as they incorporate temporal information from the sound segment, whereas GMMs do not (Bimbot et al., 2004). It is possible that the lower accuracy observed for certain word pairs might be due to our use of time-independent modelling of spoken features. For example, the words “high” and “low” certainly sound dissimilar in real time, but collapsing their features to a single point in time, as in a GMM, may increase their similarity. Using HMMs instead of GMMs in ChoiceKey might lead to higher identification accuracy for word pairs whose features overlap significantly on a time-independent scale, or for larger sets of words.

To summarise, ChoiceKey can easily be used to collect spoken responses and precise response times without the need for manual coding of responses associated with voice keys. We have shown that with 30 training exemplars, ChoiceKey is inaccurate on only around 1% of trials for only around a quarter of participants. We doubt whether this error rate is much different from errors made using other forms of response collection (i.e., pressing the wrong button when using a keyboard or mouse). Some may worry that the time taken to train ChoiceKey using 30 exemplars might be restrictive or offer no benefit over the time taken for participants to learn response-button mappings; however, with only two responses this training would take only two minutes assuming two seconds per response. We also note that once trained, unlike a participant, ChoiceKey will not forget its training. At present, the only major downfall of ChoiceKey is that it is limited to highly accurate measurement when using 2 responses.

References

Beringer, J. (1992). Timing accuracy of mouse response registration on the IBM microcomputer family. Behavior Research, Methods, Instruments & Computers, 24, 486 – 490.

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal of Applied Signal Processing, 4,* 430-451.

Crosbie, J. (1990). The Microsoft mouse as a multipurpose response device for the IBM PC/XT/AT. *Behaviour Research Methods, Instruments, & Computers, 11,* 305-316.

Lacouture, Y. & Marley, A. A. J. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception & Psychophysics, 66,* 1206-1226.

Plant, R. R., Hammond, M., & Turner, G. (2004). Self-validating presentation and response timing in cognitive paradigms: How and why? *Behaviour Research Methods, Instruments, & Computers, 36,* 291-303.

Plant, R. R., Hammond, M., & Whitehouse, T. (2003) How choice of mouse may affect response timing in psychological studies. *Behaviour Research Methods, Instruments, & Computers, 35,* 276-284.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 307-314.

Ratcliff, R., & Rouder, J. N. (1998). Modelling response times for two-choice decisions.

Psychological Science, 9, 347-356.

Reynolds, D.A. (1992). A Gaussian mixture modeling approach to text-independent speaker identification. In: *PhD thesis*, Georgia Institute of Technology (1992) September .

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10,* 19-41.

Rubin, D., Hinton, S. & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25,* 1161-1176.

Shimizu, H. (2002). Measuring keyboard response delays by comparing keyboard and joystick inputs. *Behaviour Research Methods, Instruments, & Computers, 34*, 250-256.

Simon, J.R., & Sudalaimuthu, P. (1979). Effects of S-R mapping and response modality on performance in a Stroop Task. *Journal of Experimental Psychology: Human Perception & Performance, 5*, 176-187.

Stewart, N. A PC parallel port button box provides millisecond response time accuracy under Linux. *Behavior Research Methods, 38,* 170-173.

Stevens, S. S., Volkman, J. & Newman, E. (1937). A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, *8*(3), 185-190.

Trewin, S., & Pain, H. (1999). Keyboard and mouse errors due to motor disabilities. *International Journal of Human Computer Studies, 50*, 109-144.

Wang, H., & Proctor, R. W. (1996). Stimulus-Response compatibility as a function of stimulus code and response modality. *Journal of Experimental Psychology:*

Human Perception & Performance, 22, 1201-1217.

Vidulich, M. A., & Wickens, C. D. (1985). Stimulus-central processing-response compatibility: Guidelines for the optimal use of speech technology. *Behavior Research Methods, Instruments, & Computers*, *17*, 243-249.

Voss, A., Leonhart, R., Stahl, C. How to make your own response boxes: A step-by-step guide for the construction of reliable and inexpensive parallel-port response pads from computer mice. *Behaviour Research Methods. 39,* 797-781.

Appendix A

Example Experiment

We provide MATLAB code for a mock experiment (*example.m*), where participants are asked to determine on each trial which of two tones differing in loudness is presented. The purpose of this code is not only to help the user collect responses using ChoiceKey, but also to show how MATLAB can be used to control a simple experiment. To begin the experiment, "*example"* (the name of the .m file) should be entered into the MATLAB Command Window. The experiment starts by calling of the *train* function, which allows for the recording of responses and the subsequent training of ChoiceKey. After the training, a test of ChoiceKey's identification accuracy is performed using the *traintest* function. After the function reports accuracy and number of errors the enter key must be pressed to continue to the experiment.

When the experiment begins, participants are prompted to press any key to continue. The function *getkeywait* has been included as it is a handy way to get MATLAB to wait to accept and then return keyboard responses. A fixation cross is then presented for 300 ms, followed by the presentation of the stimulus. In this experiment a tone is played; however, this can be easily adapted to any other stimuli, such as strings of characters, using code similar to that used to display the fixation cross. Similarly, images can be displayed using the *imread* and *image* functions in MATLAB.

The function *test* then allows the participant to speak their response, and will return the response which ChoiceKey calculates to be most probable given its training, as well as the response time. Feedback is displayed for one second, as either the word "Correct" or the correct response, depending on whether the participant was correct or

incorrect, respectively. At the end of each block the block number, trial number, stimulus presented, response time and given response are all recorded to a text file. Participants are either given a break of fixed duration, or thanked for their participation if they have completed all blocks. Following is pseudocode for the example experiment:

```
#Train ChoiceKey using the train function
train()
#Test the trained version of ChoiceKey
traintest()
#The experiment
For (k in 1:number of trials)
        #Show the stimuli
        showstimuli()
        #Collect the response and use ChoiceKey to get RT and response
        test()
end
```

Appendix B

The file *sr.zip* is available from the Psychonomic Society's electronic archive www.psychonomic.org/archive. The zip file contains the functions, and their respective .m files, required for ChoiceKey to run. The end user will normally only be concerned with the following four functions:

The *train* function

Typically the first function employed is *train*. Participants are first given the complete response set, followed by a series of presentations of each word individually. During each presentation participants are asked to read aloud the presented word. These initial utterances form an exemplar set which ChoiceKey uses to make all future identifications. Only one parameter must be set for the *train* function, the response set: "words". Optional parameters are the number of responses per word to use for training, "ex", the duration of recording, "duration", the frequency of recording, "Fs", and "trigger", the input energy required before audio capture begins. The default number of exemplars that ChoiceKey uses is 10. After the initial exemplars are recorded there will be a short pause of around 30 seconds to a minute, depending on the size of the response set, while ChoiceKey extracts the features and builds a Gaussian mixture model for each exemplar (Reynolds et al., 2000). We experimented with different numbers of features (Gaussians) and found 5 (the default value) to be best with our response sets. Both a smaller and larger number of Gaussians decreased accuracy, and larger values increased computational time.

The *traintest* function

The *testtrain* function provides a test of ChoiceKey's identification accuracy for each participant. An extra set of exemplars for each word in the response set are recorded and then used to provide an estimate of expected identification accuracy for the participant. The *testtrain* function requires the response set, "words", the number of responses per word to use in testing, "ntest", and the results of the training, "mu", "sigma", and "c", to be given. The optional parameters are the same as for the *train* function as well as an additional parameter "silent", which defaults to "F" (false). The *traintest* function returns the proportion of correctly identified responses and displays it, and the number of errors in identification, in the MATLAB Command Window if "silent" is not set to "T" (true).

The experimenter may choose to use this feedback to decide if the expected identification accuracy is too low, and whether or not to use ChoiceKey with this participant, or possibly to use the extra responses just recorded as additional training exemplars. After the accuracy is displayed onscreen (if "silent" is set to "F") the experimenter is asked to decide whether or not to use the additionally recorded responses as the exemplars in training ChoiceKey. If silent is set to "T", the participant's accuracy is written to a text document called acc.txt. If "yes" is chosen after the prompt then a pause will occur while ChoiceKey is trained on the new responses.

The *callib* function

The *callib* function offers the experimenter a method for estimating the differences in identifying onset time for different responses. Participants are instructed to respond with one of the words from the response set for a block of *n* trials each time they are presented with a neutral stimulus (a "+" sign). The process is repeated for each

word in the response set. The function writes to a text file, callib.txt, the average RT for each response. The *callib* function requires as input the response set, "words" and the number of recordings per response, "nrec". The optional parameters "duration", "Fs" and "trigger", default to 1.5 seconds, 44100Hz and 0.05, respectively.

The *test* function

After ChoiceKey has been trained for a participant's voice, the *test* function can be called whenever response collection is required. This function will record the participant's response and return the most likely spoken response, given the set of exemplars recorded in the training stage. The time taken to make the response is also returned. As input, the *test* function needs the three parameters "mu", "sigma", and "c" returned by the *train* function. Optional parameters also include "duration", if different from 1.5 seconds, "Fs", if different from 44100Hz, and "trigger", if different from 0.05.

Once the function is called, the audio capture device is activated and waits until audio input reaches the "trigger" threshold, after which it records audio signal for a set duration. The time taken from stimulus onset to the audio signal reaching threshold is recorded as the response time, and returned to the user. The trigger threshold value default of 0.05 worked well in our testing; however, this value can be changed by the user. For example, if background noise is present (e.g., from a computer fan), the threshold may be increased to reduce the false alarm rate due to triggering by background noise. However, setting this value too high may result in responses being missed by ChoiceKey.

Getting more from accuracy and response time data:

Methods for fitting the Linear Ballistic Accumulator

Chris Donkin, Lee Averell, Scott Brown & Andrew Heathcote

The University of Newcastle,
Newcastle, Australia

Abstract

Cognitive models of the decision process provide greater insight into response time and accuracy than standard Analysis of Variance techniques. However, they can be mathematically and computationally difficult to apply. We provide instructions and computer code for three methods of estimating the parameters of the Linear Ballistic Accumulator (LBA), a new and computationally tractable model of decisions between two or more choices. These methods vary in their flexibility and user accessibility, and include a Microsoft Excel worksheet, scripts for the statistical program R, and code for implementation of the LBA into the Bayesian sampling software WinBUGS. We also provide some scripts in R which produce a graphical summary of the data and model predictions. Finally, in a simulation study we explore the effect of sample size on parameter recovery for each of our different methods.

Many tasks used in experimental psychology involve participants making relatively simple decisions, for which the experimenter measures the time taken and the accuracy of their responses. In many cases the difficulty of the task is also manipulated within subjects. The resultant interaction between speed, accuracy and difficulty is complicated and presents significant challenges for standard analysis techniques even in the simplest case of two response alternatives. Results from an experiment conducted by Ratcliff and Rouder (1998) are a good demonstration of the range of effects that can occur, even within data from a single participant. They also demonstrate the well established tradeoff between decision speed and accuracy, whereby a participant can improve their accuracy by increasing the time taken to make a decision. The complex interdependence of accuracy and response time (RT) draws into question the common practice of analysing accuracy and RT separately – for example, using two separate ANOVAs.

For more than 30 years, mathematical psychologists have been developing cognitive models to account for the wide range of choice RT phenomena. These models use a small number of decision process variables to account for both the accuracy of responses and the complete distribution of associated RTs. Many models exist, (e.g. Brown & Heathcote, 2005; 2008; Busemeyer & Townsend, 1992; Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlincx, 2002; Smith & Ratcliff, 2004; Van Zandt, Colonius & Proctor; 2000; Vickers, 1970) each differing in their assumptions about the exact nature of the underlying processes. However, most share the same basic framework. They assume that when making a decision the participant repeatedly samples information from the environment and this information is used as evidence for

one of the potential responses. Once the evidence in favour of one response reaches a threshold the decision process is terminated and that response is made. The time taken to make the response equals the time to accumulate the required amount of evidence plus some time taken for non-decision processes, such as perception, and the execution of a motor response. These cognitive models all provide estimates of three key parameters: the rate at which evidence for a particular response is accumulating (*drift rate*), how much evidence is required before making a response (*response threshold*), and how much time is taken for non-decision aspects of the task (*non-decision time*). These quantities are estimated taking into account the interaction between speed and accuracy in the decision being made. This unified account can be much more informative about the decision process than independent analyses of accuracy and RT.

The community using cognitive process models of choice RT has been steadily growing. These models have been used to describe the underlying neurology of simple decisions (e.g.: Carpenter, 2004; Forstmann, Dutilh, Brown, Neumann, von Cramon, Ridderinkhof, & Wagenmakers, 2008; Gold & Shadlen, 2001; Hanes & Carpenter, 1999; Mazurek, Roitman, Ditterich, & Shadlen, 2003; Ratcliff, Cherian, & Segraves, 2003; Reddi, 2001; Roitman & Shadlen, 2002; Schall, 2001; Smith & Ratcliff, 2004). They have also been used to provide insight into the cognitive processes underlying a wide range of simple choice tasks, including elements of reading (Ratcliff, Gomez & McKoon, 2004), recognition memory (Ratcliff, 1978), and visual discrimination (Ratcliff, 2002; Smith & Ratcliff, in press), as well as more complex decisions, such as purchasing a car (Busemeyer & Townsend, 1992). Ratcliff, Thapar and McKoon (2001; 2003) used a decision model to identify which factors associated with aging were responsible for the observed slowing of older participants in simple discrimination

tasks. In this application, independent, separate analysis of accuracy and RT would not have identified these factors, as they caused a tradeoff between speed and accuracy. Wagenmakers, van der Maas, and Grasman (2007) also suggested that variables estimated by these models (such as the rate of accumulation of information, drift rate) be used to describe data in preference to accuracy and RT. Their approach is akin to considering 'intelligence' in terms of an intelligence quotient rather than performance in individual tasks. Wagenmakers et al. give a good explanation of how the analysis of data using decision models can reveal patterns in data that would have otherwise escaped the notice of the experimental psychologist, in a manner similar to psychometrics.

Despite their ability to provide insight into the processes underlying decisions, the application of decision models has been mostly limited to those already within the field of cognitive modelling. This is because the models have been notoriously difficult to apply to data, requiring complicated computer programming and mathematics to implement (Smith, 2000; Van Zandt, 2000). The models have grown more complex as the range of phenomena they can account for has grown (e.g., compare the original Ratcliff diffusion model, Ratcliff, 1978, to more recent versions in Ratcliff & Rouder, 1998 and Ratcliff & Tuerlickx, 2002). Fortunately, there have also been attempts to reduce the complexity of choice RT models. Wagenmakers et al.'s (2007) *EZ* diffusion provides simple formula for directly estimating the three key decision parameters based on three easily estimated statistics. However, the model underlying the *EZ* diffusion approach is not comprehensive. For example, it fails to account for the latency of error responses. Such criticisms do not apply to the "complete" decision making models, such as Ratcliff's. However, the price of this explanatory power is greatly increased

mathematical and computational complexity.

Brown and Heathcote (2008) proposed the linear ballistic accumulator (LBA) model as the simplest complete model of choice RT. The LBA is simple in the sense that Brown and Heathcote were able to analytically derive the model's probability density function (*pdf*), which makes efficient estimation tractable using a range of techniques. Despite its relative simplicity, the LBA can account for the same breadth of empirical two-choice RT phenomena as the Ratcliff diffusion model. In contrast to the diffusion model, the LBA can also be applied to choices amongst more than two alternatives. Even though the model is in its infancy, it has already begun to be applied to experimental data sets (see, e.g., Donkin, Brown & Heathcote, 2009b; Forstmann et al., 2008; Ho, Brown, Serences, 2009). Forstmann et al. showed that an experimental manipulation of speed and accuracy emphasis produced changes in behaviour and brain activity which closely agreed with appropriate parameters from the LBA.

In recent years, the options available for estimating parameters for the Ratcliff diffusion model have been increasing, and have become more user-friendly. Vandekerckhove and Tuerlinckx (2007; 2008) developed "DMAT", a MATLAB program, which uses methods developed by Ratcliff and Tuerlinckx (2002) to apply the Ratcliff diffusion model. Vandekerckhove, Tuerlinckx, and Lee (submitted) have implemented the diffusion model into the sampling program for Bayesian inference, WinBUGS. Voss and Voss (2007; 2008) offered "FastDM", standalone C code that also implements the Ratcliff diffusion model. Wagenmakers et al. (2007) offered a spreadsheet in Excel, a web applet, and some scripts in the statistical language R that could all be used to obtain *EZ* diffusion estimates.

The current paper is motivated by the observation that most previous software

offerings for applying choice RT models to data have focused on Ratcliff's diffusion model. Here we provide a similar range of options for estimating the parameters of an alternative choice RT model, the LBA. We first review the details of the LBA, and then describe estimation software for it developed in Microsoft Excel, R and WinBUGS. We then describe additional software developed in R to produce a visual summary of data and model predictions. Rather than providing a comprehensive parameter estimation environment for a particular paradigm, our aim is to illustrate the different approaches in a way that allows users to flexibly extend the analysis to a range of paradigms.

Overview of the LBA model

Consider a participant who has been presented with a string of letters and asked to decide whether the stimulus is a word or a non-word – this decision is represented in the LBA as shown in Figure 1. Each possible response ("word" and "non-word") is assigned to an independent evidence accumulator. Evidence accumulation starts at a value randomly sampled (separately for each accumulator) from the interval $[0, A]$ at the beginning of each trial. The participant gathers information from the stimulus which is then used to increment the evidence in either accumulator. Brown and Heathcote (2008) made the simplifying assumption that evidence accumulation occurs linearly, at a rate termed the *drift rate*. Drift rate is an indication of the quality of the stimulus: the larger the drift rate, the faster the accumulation of evidence occurs. For example, because higher natural language frequency words are easier to classify as a word, a string of letters that forms a frequently used word, such as "house", would likely have a higher drift rate than the word "siege", since it is used less often. The drift rates for each accumulator vary from trial to trial according to a normal distribution with mean drift rates v_W (for words) and v_{NW} (for non-words). For simplicity, we assume a common

standard deviation, *s*, for these distributions. Once evidence in one accumulator reaches a threshold, *b*, the response associated with that accumulator is made. Changing the threshold parameter changes the amount of evidence required to make a response. For example, a lower *b* produces less cautious responses and an increased *b* more cautious responses. The relative values of *b* for different accumulators can model response bias – an a priori preference for one response over the other. Response latency is given by the time taken for the first accumulator to reach threshold plus the time taken for non-decision aspects of the task such as the motor response and stimulus encoding. Non-decision time is assumed to have negligible variability, so is estimated by a single parameter, T_{er}.

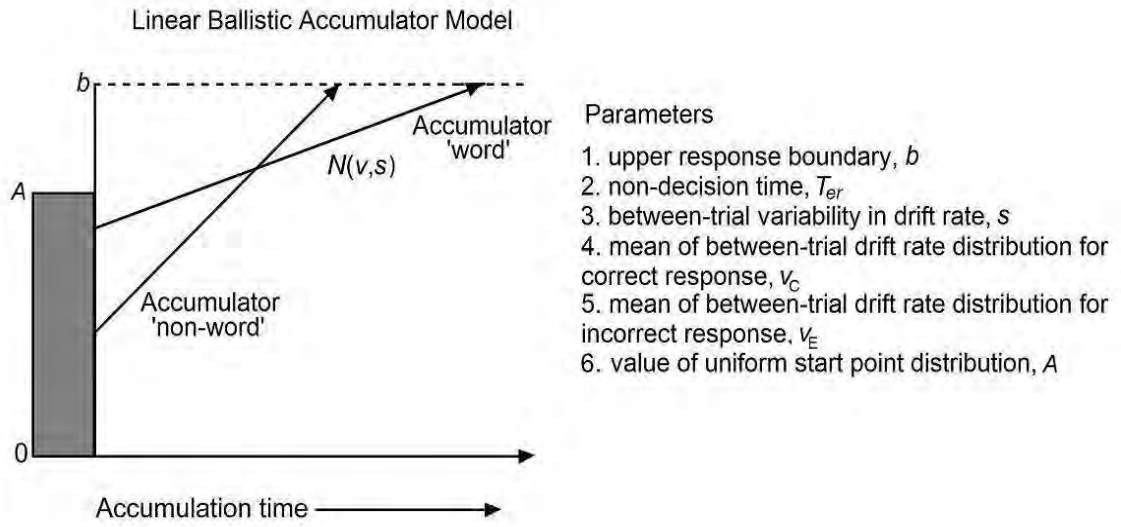

Figure 1 Graphical representation of a single decision made by the LBA model.

One way of estimating LBA parameters from data involves the search for a set of parameters (*b, A,* v_W, v_{NW}, *s,* T_{er}) that produce predictions for accuracy and RT which closely resemble the data. The resemblance between data and model is quantified by an *objective function*. A variety of different objective functions are commonly used with RT data, including maximum likelihood (Ratcliff & Tuerlinckx, 2002), chi-squared

(Ratcliff & Smith, 2004) and quantile maximum products estimation (QMPE; Heathcote & Brown, 2004; Heathcote, Brown & Mewhort, 2002). The search for a set of parameters which optimise the objective function begins with the choice of parameters at some initial values, called a *start point*. This is followed by a computer-driven search that changes parameters until a set is identified which provides a better value for the objective function than other nearby sets. Bayesian estimation takes an alternative approach that provides a distribution of estimated parameter sets rather than a single set. Variability in the distribution quantifies uncertainty about estimation and a measure of the distribution's central tendency, such as the mean, provides a point estimate. Our aim here is not to detail or compare these different methods. Instead, we take advantage of the ready availability of efficient and general purpose search algorithms (*solver* in Excel and *optim* in R) and Markov Chain Monte Carlo methods for generating Bayesian estimates (WinBUGS) to provide accessible options for estimating LBA parameters given a set of data.

Methods for estimating LBA parameters from data

We use a single set of simulated accuracy and RT data to illustrate the different methods of applying the LBA model to a two choice task. The design from which the simulated data are derived is typical of designs that the LBA has been applied to: three different conditions that vary in decision difficulty – easy, medium and hard. This set can be thought of as data from a single participant in an experiment with three within-subjects conditions. The first few lines of the simulated data are shown in Figure 2, with each line representing one choice trial. The first column codes the easy, medium and difficult decision conditions, labelled 1 to 3. The second column codes the accuracy of the response made, using 0 for incorrect and 1 for correct. The third column contains the

response latency of the decision in milliseconds (msecs). We provide an R script that simulates data in the same format as our example dataset ('makedata.r'). For this example, we sampled data from an LBA model with the following parameters: $s = 0.25$, $A = 300$, $T_{er} = 200$, $b = 400$, $v_E = 0.9$, $v_M = 0.75$, $v_H = 0.6$.

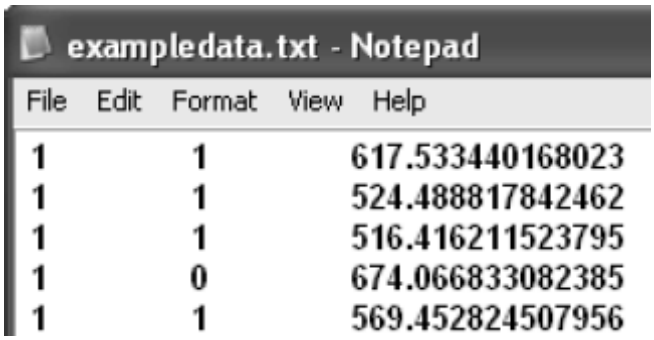

Figure 2 The first 5 lines of data from our simulated data set. The first column contains the experimental condition (1-3), the second the accuracy (0=incorrect, 1=correct), the third RT (in milliseconds)

The parameters v_E, v_M and v_H refer to the drift rates for correct responses. We use the traditional parameterization which fixes drift rates for the error responses to be equal to one minus the drift rate for correct responses (although see Donkin et al., submitted, for a different approach). Hence the drift rates for incorrect responses in our example data set were 0.1, 0.25 and 0.4 for easy, medium and hard conditions, respectively. To keep the example simple, we assumed that drift rate for correct (and error) responses was the same regardless of which stimulus was presented on that particular trial. This embodies an additional assumption that drift rates are the same for both stimuli corresponding to each response (i.e., words and non-words). In more general cases there could be four drift rates – a correct and error drift rate for each of the two responses (e.g., old and new responses in a recognition memory task, Ratcliff, 1978). We also assume that only drift rate changes between the three difficulty conditions, with all other parameters constant. This assumption is standard in paradigms where the stimulus

factors that are used to manipulate difficulty are mixed randomly within blocks of trials. We address the more complicated cases later.

Example 1 – Using Microsoft Excel

We use the file 'lba.xls' to fit the LBA using Microsoft Excel (hereafter *Excel*). The Excel LBA worksheet records the data in Sheet 2 and uses the parameter values in Sheet 1 to calculate the likelihood of the data, given the current set of parameter estimates – this likelihood value is our objective function. Excel's inbuilt "Solver" function is used to find parameters which maximise this likelihood. The quality of the fit to data is shown in the histograms presented in Sheet 1.

The likelihood can be thought of as a measure of the quality of the fit of the model to the data, with larger values indicating better fit. The parameters that provide the best fit to the data are those that maximise the likelihood, or equivalently the log-likelihood. We use log-likelihood as the objective function, rather than raw likelihood, because taking logarithms avoids numerical problems associated with multiplying together many very small numbers.

To analyse our simulated data, the three columns of data were pasted directly from exampledata.txt into Sheet 2 in columns A-C, rows 2 (hereafter cells A2-C2) and onwards. Initial parameter guesses were entered into cells B1-B7 of Sheet 1. The natural logarithm of the likelihood of the current set of parameters given the data is shown in cell B9 of Sheet 1. The Solver function, which can be found in the Tools drop down menu, is then applied[1]. Solver can then be used to maximise the log-likelihood by going to Tools > Solver. A new application box will appear ready for the user to simply click Solve. There are numerous options in the Solver function users should feel free to

1 If the Solver option does not appear in the Tools menu go to Tools > Add-Ins and check the box labelled 'Solver Add-in'.

experiment with. However, no such changes are required to fit the LBA to our example data. Although we do not discuss these options in detail here, we note that by default the 'Subject to the Constraints' section is appropriately set up so that condition #1 is the easiest condition, and hence should have the highest drift rate, that condition #2 is the next hardest condition, and so on. A number of other sensible constraints can also be imposed, such as requiring $T_{er}>0$ and $b>A$.

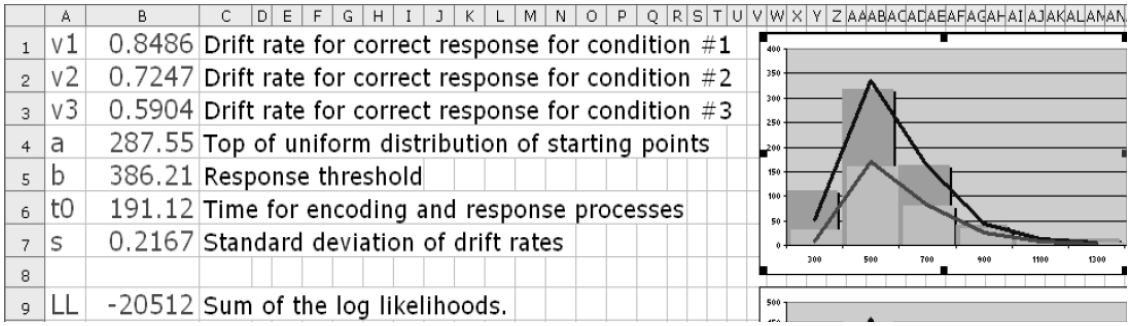

Figure 3 Screenshot of the Excel LBA worksheet. The plot contains the data from one condition as a histogram, with bars showing error and correct responses. Solid lines show predictions of the LBA for error and correct responses, respectively. Predictions are based on parameter values given in row B in the figure.

A visual summary of the quality of the fit is shown by the plots in Sheet 1, an example of which is shown in Figure 3. To create these plots, the user must first place the RTs from their data into Column A of Sheet 3. This can be done by copying and pasting the contents of Column C of Sheet 2. There are three plots shown in Sheet 1, one for each condition. The plots show the correct and error RT histograms for the data and the LBA grouped into bins ranging from 300-1500 ms in increments of 200ms. The three histograms in the Excel sheet show data and predictions from the easy, medium and hard conditions from top to bottom, respectively. The data are shown by the bars of the histogram – the actual spreadsheet uses colour figures, in which red bars represent error responses and blue bars show correct responses. The predictions made by the LBA are shown by solid lines; again, in the actual spreadsheet colours are used to error and

correct responses. The predictions shown in the plots are based on parameter values given in Column B of Sheet 1. Changing the parameter values by hand causes direct changes to the solid lines in the histograms. The LBA provides a good fit to the data whenever the solid lines match closely the bars of the histograms that underlie them, indicating that the RT distributions predicted by the LBA closely resemble those of the data. Once a good fit to the data is achieved, the user can record the parameter values reported in cells B1-B7 of Sheet 1. We do not suggest that these histograms are of publication quality, however, they do provide the user with a method for quickly checking the quality of the fit. We will discuss how to use the parameter estimates to create further graphical summaries in a later section.

Sometimes, Excel's Solver function becomes stuck in a *local maximum*, rather than finding the optimal solution – the *global maximum* – and then the estimated LBA parameters will not provide an accurate account of the data. This occurs when the solver finds a set of parameters, say, solution *A*, that are better than other nearby sets of parameters, but are still worse than the (quite different) parameters which provide the best fit to the data. Although this problem can be difficult to avoid in general, there are some measures that help address it. One method is to start the search again from a different starting point. Solution *A* is likely a global maximum if the Solver function repeatedly finds Solution *A* from a sufficiently wide range of starting points.

The choice of initial estimates for any method of fitting a model to data can be very important –automatic optimisation routines (like Solver) can fail badly if the initial estimates are poor. Choosing good initial estimates is something of an art that often requires experience and experimentation. This process is made relatively easy thanks to the interactive nature of the Excel spreadsheet we have provided. The effects of

changing parameters can be instantly observed by looking at the plots in Sheet 1. The initial estimates need not produce a set of solid lines that closely resemble the data; all that is really necessary is that Solver be given parameters that produce a solid line that has approximately the shape and location of the observed RT distributions. As a rough guide, the parameters used to generate our simulated data would be a good starting point for parameter search for new data sets as they approximate average parameter values from fits to a range of different paradigms.

Given the wide range of possible data sets, it is impossible to create an Excel workbook that will work 'out of the box' for every case. For example, simply changing the number of observations in each condition requires the user to change certain aspects of our worksheets. However, given the flexibility and intuitive nature of data manipulation in Excel, the changes required to adapt the LBA workbook to a new datasets should usually be relatively simple. For example, if the number of observations in each condition does change then the user has to update the entry in C12 of Sheet 1 and make sure columns G-O are the same length as the data which has been entered.

Example 2 - Using R

To estimate LBA parameters from our example data using R, the user begins by extracting the files in 'Rfit.zip' into a directory. The R language is free software available for Windows, Mac OS X and Linux. It can be downloaded from the R homepage http://www.r-project.org/. Once you have downloaded and run the executable install file, follow on-screen instructions to install the R software. Extracting the contents of the 'Rfit.zip' file into a folder (e.g. Desktop\Rfit) provides four files – 'lba-math.r', 'pq-lba.r', 'fit-example.r' and 'exampledata.txt'. The .r files all contain R scripts that are used to estimate parameters for the LBA model from the example data

contained in the .txt file. To begin the estimation process, the user should open the R software and change the working directory to the folder where the zip file was extracted (i.e. Desktop\Rfit), for example by using the File > Change dir option. Typing `source("fit-example.r")` fits the LBA to the data found in the exampledata.txt file. The R prompt will disappear while the software searches for a set of parameters which best fit the data. Once search is finished the estimated parameters are printed on screen.

Using the `source` function in R is equivalent to entering each line contained in the 'fit-example.r' R script directly into the R window. To see what is being run by the source command you can open the 'fit-example.r' file in any text editor. The first command sources the 'pq-lba.r' script, which in turn sources the 'lba-math.r' script. These two scripts set up the functions that encode the basic mathematics of the LBA model. Some variables are then defined - *qps* gives the set of quantiles of the RT distribution to be used by the QMPE method of fitting the LBA and *trim* gives the minimum and maximum values used to censor the RT data. We set *trim* to remove RTs faster than 180 ms and greater than 10 seconds. The R ***read.table*** function reads the contents of 'exampledata.txt' file, maintaining the column structure and giving the variables the names "difficulty", "correct" and "rt". The user can change the name of the first argument of the ***read.table*** call to change what data file being read by R. However, they would need to ensure that the data followed the same structure as that of the exampledata.txt file, or else change other aspects of the R script appropriately.

The next section of script transforms the imported data into the format required for the QMPE fitting technique. Because QMPE is based on quantiles (i.e., order statistics such as the median) it provides more robust estimates than maximum likelihood in small samples (Brown & Heathcote, 2003; Heathcote, Brown & Mewhort,

2002). In the example the estimates are based on five quantiles (.1, .3, .5, .7 and .9) for both correct and error RTs in the different difficulty conditions and storing them in the array, `q`. The number of observations in each quantile bin is also calculated and stored in `pb`. Response accuracy and sample sizes for correct and incorrect responses in each difficulty condition are also calculated and stored in `p` and `n`, respectively. Finally, these variables are finally bundled together in a list and stored in the variable `data`.

Once the data are formatted correctly, the parameter search is called by the ***fitter*** function. There are two arguments required by the fitter function. The first is the `dat` argument, which in our example is the `data` list. The `maxit` argument specifies the maximum number of iterations that should be used when searching for best fitting parameters. Like the Excel workbook described earlier, R finds best fitting parameters by, at each step, systematically changing parameters and keeping changes which provide a better value for the objective function. The `maxit` argument specifies how many steps are taken when trying to find a set of parameters which best fits the data. The parameters which arise out of the ***fitter*** function are placed into the `pars` variable. The last few lines of the script transform the parameter values returned by the ***fitter*** function to those that are familiar to the reader from our explanation of the LBA. Figure 4 shows the R output after fitting the LBA to the example dataset.

```
> source("fit-example.r")
      s       A     ter       b      v1      v2      v3
  0.242 293.715 195.177 392.699   0.875   0.738   0.595
>
```

Figure 4 Screenshot of the use of R to fit the LBA to our example dataset.

Unlike the Excel worksheet, the majority of the code which does the fitting in R is hidden in the 'pq-lba.r' and 'math-lba.r' scripts. The parameter values used to

initialise the search for best fitting parameters are produced automatically as part of the ***fitter*** function defined in the 'pq-lba.r' script. These heuristics can be found, clearly labelled, in the 'pq-lba.r' file. These estimates will work in many situations but in a small number of cases will be inadequate, such that fitting algorithm will be unable to find good parameter estimates. Such cases reinforce the need to use the graphical summary methods given below to check the quality of a fit. To use these scripts with other data sets where only changes in drift rate are extended across conditions, only the file 'fit-example.r' need be edited by the user. In this situation, the user must simply change the `ndrifts` parameter to be equal to the number of conditions in the data.

When other parameters are allowed to vary across conditions then more substantive changes are required. For example, Donkin, Heathcote, Brown and Andrews (submitted) propose that in a lexical decision task both drift rate and non-decision time vary with word frequency. Say we have three frequency conditions and so we wish to estimate three values of drift rate, v, and three values of non-decision time, T_{er}. In such a situation the user would update the ***fitter*** function in 'pq-lba.r' so that starting points are generated for v_1, v_2, v_3 and T_{er1}, T_{er2}, T_{er3}. The ***obj*** function should then also be updated to take into account these changes. Specifically, the `par` vector passed to the ***obj*** function will be two elements longer – it used to contain s, A, T_{er}, b, v_1, v_2, v_3 and now has s, A, T_{er1}, T_{er2}, T_{er3}, b, v_1, v_2, v_3. The ***getpreds*** function expects to receive from ***obj***, for each parameter of the LBA, a vector which is has length equal to `nc`, the number of conditions (3 in this case). This means that where we would have previously have replicated T_{er} nc times (the line: `Ter=rep(par[3],nc)`) we now use three free parameters (`Ter=par[3:5]`), in just the same way as we previously used three separate drift rate estimates (previously `v=par[5:7]`, now `v=par[7:9]`).

To analyse data with more than one factor, further changes to the 'pq-lba.r' file need to be made. For example, we may have a difficulty manipulation which varies between trials and a speed-accuracy emphasis manipulation which varies between blocks of trials. In this case, it's customary to fit an LBA where *v* varies between difficulty conditions and where *b* and *A* vary between emphasis conditions. We begin in the same way by first generating start points for each of the parameters to be estimated within the **`fitter`** function. However, in the **`obj`** function, rather than producing a vector of length `nc` for each parameter, we must now produce a matrix with `nc` rows and 2 columns, one for speed emphasis parameters and accuracy emphasis parameters. This also means that where the **`getpreds`** function used to take one element of the parameter vector (by using a loop over 1:`nc`), it will now have to take one element of the matrix of parameter values (using two loops, one over 1:`nc` and another over 1:2). Obviously, as the design of the data and the LBA model to be fit becomes more complicated, so too will the R code needed. For those readers whose programming skills may be limited we later provide code for WinBUGS which can be more easily adapted to more complicated models.

Multiple Choice Data

One of the advantages of the LBA is its ability to model multiple choice data (see Brown & Heathcote, 2008, for a demonstration). To illustrate we provide code which can be used to simulate a set of data from an LBA with four accumulators, corresponding to a choice between four response alternatives. The code then fits the four-accumulator LBA model to the simulated data to recover the parameters. The data are simulated to mimic an experiment in which a participant is presented with one of four 'random dot kinematograms' (RDKs) – a set of pixels of which a small proportion

move coherently in one direction which must be identified by the participant, while the others move randomly. The difficulty of the task was also manipulated to be easy, medium or hard. Ho, Brown and Serences (submitted) fit the LBA to an experiment using this paradigm. The code necessary for simulating and fitting the multiple choice data is contained in the 'Rmultifit.zip' file. After extracting the files data can be simulated and fit using `source("fit-multi.r")`. The parameters are estimated by maximum likelihood, printed on screen and histograms containing data and model predictions are produced. We used maximum likelihood estimation in this case to illustrate how the R code described in the last section can be adapted for a different objective function.

The 'fit-LBA.r' script is completely self-contained, in that we are not required to source the 'lba-math.r' or 'pq-lba.r' files. We do this because the code required for maximum likelihood estimation is relatively simple compared to that of QMPE. The data are simulated, and starting values are generated for the parameters we wish to estimate using similar heuristics to those used in our other R code. We then define our objective function, ***obj***. Since we are using maximum likelihood our ***obj*** calculates the likelihood of each RT value given a set of parameters *pars*. Finally, we include a set of code for producing histograms of observed and predicted RT distributions for each response in each difficulty condition.

For both simulating and fitting the data we assumed that all parameters were fixed across stimuli, suggesting that participants show no bias for one particular direction of pixel flow. The simulated data used a large drift rate corresponding to the correct response; the size of this drift rate varied for easy, medium and difficult conditions. We reasoned that incorrect responses were more likely in the two directions

orthogonal to the correct response, and less likely in the direction opposite the correct response. To instantiate this, we used only two free parameters for the three error response alternatives: one value to set the fraction of the correct-response drift rate assigned to the orthogonal responses (we used 0.5) and one value to set the fraction assigned to the opposite response (0.25 in our simulated data). We simulated the data, therefore, using 5 parameters associated with drift rate – three drift rates for the correct responses in each condition, and a parameter indicating the proportion of the correct drift rate for perpendicular incorrect responses and a parameter indicating the proportion of the correct drift rate required for opposite direction incorrect responses. When fitting the data, to fully demonstrate the method for estimating parameters from a multiple accumulator LBA model, we made no assumptions about relationships between drift rates. We estimated 12 drift rate parameters – one for each of the four responses in the three difficulty conditions. Figure 5 shows parameter values returned by our maximum likelihood fitter. The parameters reported are close to the parameters used to simulate the data (A=300, b=400, T_{er}=*300,* v_e=.9, v_m=.75, v_h=.6, p_{perp}=.5, p_{opp}=.25). The histograms in Figure 5 also demonstrate that the predictions from the LBA match closely the observed data. The biggest misfit is to RT distributions for error opposite responses. These responses are the most incorrect and are made least often. Therefore, RT distributions for these responses are made up of relatively few observations and hence estimation of parameters for these responses is more erroneous.

```
> source("fit-multi.r")
       A      ter        b     v_e1     v_e2     v_e3     v_e4     v_m1     v_m2     v_m3
310.738 294.111 418.411    0.910    0.516    0.472    0.239    0.759    0.389    0.379
    v_m4     v_h1     v_h2     v_h3     v_h4
   0.245    0.631    0.313    0.330    0.127
```

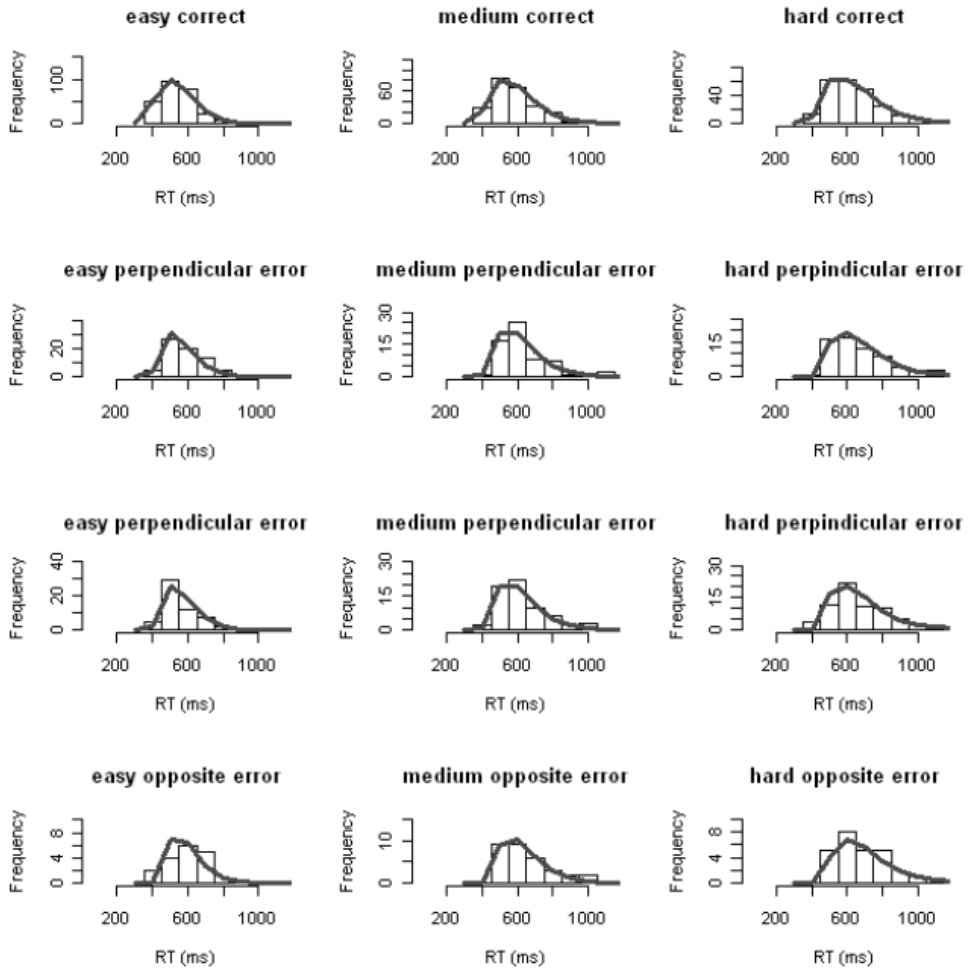

Figure 5 Screenshot of the R code and resultant output used to fit data simulated from a four-accumulator LBA model. Data are represented by the bars of the histogram.

Example 3 – Using WinBUGS

Bayesian analysis in psychological research is rapidly gaining popularity for a range of reasons, such as providing estimates that perform well in predicting new data sets and that take account of model flexibility (Wagenmakers, Lee, Lodewyckx,

Iverson, 2008; see Raftery, 1995, and Wasserman, 2000, for general introductions). Bayesian analysis starts by assuming a "prior" distribution of parameter estimates (i.e, distribution before new data is taken into account). It then combines the prior with the observed data to produce a "posterior" distribution of parameter estimates (i.e., estimates updated by the new data). The process of Bayesian estimation has been made relatively easy by the availability of flexible programs, such as WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000), which use general purpose Markov Chain Monte Carlo (MCMC) methods to obtain samples from the posterior distribution (see Calin & Chib, 1995). We demonstrate how WinBUGS can be used to fit the LBA to data, including instructions for compiling and running WinBUGS, as well as reviewing and saving the results.

WinBUGS makes MCMC methods available to researchers with relatively little programming and mathematical knowledge through a graphical user interface. The Appendix to this paper provides instructions for installing WinBUGS and the WinBUGS Development Interface and BlackBox Component Builder[2]. The latter two programs are used to give WinBUGS access to the LBA probability density function (*pdf*). The zip file 'BugsLBA.zip' contains a compound document (odc file) that defines the LBA pdf ('lba.odc'). As described in the Appendix, a one-time installation procedure is required to enable WinBUGS to sample from the LBA posterior. Once this procedure is complete WinBUGS should always be launched from the BlackBox installation folder.

The zip file also contains a file that defines the WinBUGS model and data

2 WinBUGS requires Microsoft Windows to operate, and although a platform-independent version, OpenBUGS, does exist, the lack of equivalent multi-platform versions of the BlackBox and WBDev software means that our implementation of the LBA into a Bayesian framework is restricted to the Windows operating system.

specific to the current example ('fitlbaexample.odc'). The model and data specifications are contained in separate sections of the fitlbaexample.odc compound document. The model section of 'fitlbaexample.odc' specifies uniform prior distributions (`dunif`) for each of the LBA parameters (*A, b, v, s,* T_{er}). The parameters of the uniform priors were chosen to be relatively uninformative. That is, the range of the uniform priors is chosen to span a broad range of plausible parameter values. When it covers a sufficiently broad range, the prior is not overly influential on the posterior estimates given a reasonable amount of data. For the *A* parameter, for example, the uniform prior distribution ranges from 0.1 to 1[3].

Specification of overly broad priors can cause WinBUGS to fail, so some experimentation can be required to obtain a computationally stable but sufficiently uninformative prior. Relatively uninformative priors produce WinBUGS estimates that do not differ greatly from the estimates obtained from the two methods presented previously. A section containing initialising values (called `inits`) for the MCMC sampling can also be added to fitlbaexample.odc, but this is only necessary when the inits automatically generated by WinBUGS fail. Such failures are most common when priors are broad. Specifying appropriate inits can help to protect against failures of WinBUGS when broad priors are utilized.

As with the previous methods, we estimate different drift rates, *v*, for each of the three conditions. In WinBUGS this is achieved by letting *v* be a vector containing three priors, one for each of the three conditions. In our example code all of the *v* priors are identical and relatively uninformative, however, this need not be the case – different priors for the drift rate for each condition could be imposed if desired. The final line of

3 In our example the priors for the *A, b* and T_{er} parameters are defined in seconds. This means that reaction times given to WinBUGS should also be in seconds. This can be done by simply dividing the reaction times in the exampledata.txt file, which are in milliseconds, by 1000.

the model section connects the RT data, defined as the variable *t* in the next (data) section, with the previously defined parameters (and their priors) via the pdf for the LBA.

The next section of fitlbaexample.odc contains a modified specification of the data contained in exampledata.txt to. In order to allow WinBUGS to handle bivariate data (response time and accuracy) we follow the common practice (Vandekerchove et al., submitted; Voss, Rothermund & Voss, 2004; Voss & Voss, 2007; Voss & Voss, 2008): Let *RT* be the observed response latency for a particular response and let *t* be the data given to WinBUGS. If the response is correct then code *t* = *RT,* and if a response is incorrect then code *t* = -*RT*. This enables both accuracy and RT information to be specified in a single variable. The data section also defines other variables used in the model section. For example, the number of data points, `N`, is defined as 3000. The condition for each response is defined by the entries in the `cond` variable; the first 1000 have '1' indicating that the first 1000 RTs are from condition 1, the next 1000 having '2', and so on.

The following steps compile the model and obtain posterior samples.

1. Open the 'exampledata.odc' file from within WinBUGS (recall that this must be run from the BlackBox directory). Once opened, highlight word "model" at the top of document and select Model> Specification, this will open a dialog box labelled 'Specification Tool'. From inside the Specification Tool select 'check model' and if all parameters are given priors then a message "model is syntactically correct" will appear in the bottom left of the screen.
2. Either a single MCMC chain (default) or multiple chains may be run. In our example we use three chains by typing "3" in the 'num of chains' box. Having

multiple chains helps the user to check whether the MCMC chain converges to the posterior distribution.

3. Highlight the word ‘list’ at the start of the data section and choose ‘load data’ from the specification tool dialog box. A message “data loaded” will appear in the bottom left of the screen. If an error occurs it is most often due to misspecification of variables used in the model section (i.e., `N, nc, cond` variables in our example code).
4. Select ‘compile’ from the specification dialog box; if everything is correct the bottom left of the screen should be displaying the message “model compiled”.
5. Select ‘gen inits’ to have WinBUGS to generate initializing values for each of the three chains. After the initializing values have being generated the bottom left of the screen will have the message “model initialized” indicating WinBUGS is ready to run.

Before beginning MCMC sampling the user must indicate which posterior parameter estimates are to be saved for later analysis. This is done via the Inference> Samples menu, which will bring up the ‘Sample Monitor Tool’ dialog box. The following two steps set up monitoring and run the sampling.

6. Type the variable name into the ‘node’ section. For example, if we wished to monitor the *A* parameter enter ‘a’ into the node section (as this parameter was defined as ‘a’ in the model section). You are then required to choose at what iteration to begin and end the monitoring. The value you put in ‘beg’, which represents the number of iterations of the MCMC chain that are discarded before monitoring, is commonly referred to as the burn-in period. In our examples we used a burn-in period of 10,000 iterations. Since the MCMC chain begins with

inits values which may not represent valid samples from the posterior, a burn-in period is required before the MCMC chain converges to the posterior distribution. The 'end' value represents the length of the MCMC chains; in our example we set this value at 21,000. This will, if the chains converge, result in 11,000 samples from the posterior distribution. Larger values of end cause sampling to take longer to complete, but provide more information about the posterior. The process is then repeated for each parameter the user wishes to monitor. In our example we monitored all seven parameters (*A, b, s,* T_{er}*,* v[1], v[2], v[3]*)*.

7. Select Model> Update and the 'Update Tool' dialog box will appear. Typically you will enter the same number in the 'updates' section as you did in the 'end' section of the Sample Monitor Tool dialog. Here, you have the option of thinning the MCMC chain. Thinning discards iterations in order to reduce autocorrelation amongst the posterior samples[4]. For example, if 'thin' is set to two then every second iteration will be recorded, and so it will take twice as long to obtain the number of iterations specified in the updates section. In our example we set thin at two for every parameter. The refresh section indicates how often WinBUGS updates the screen which indicates to the user how many iterations of the chain have occurred. Setting a large value reduces processing time. Clicking the update button causes sampling to commence.

While WinBUGS is generating samples the words "model updating" will appear at the bottom left of the screen. This process can take a long time and is uninterruptible once

4 MCMC chains typically are strongly autocorrelated. Autocorrelation is not a problem for parameter estimation, except that the information contributed to the estimate by each sample is reduced. However, it can be problematic when the variability of samples is important (e.g., when calculating confidence intervals on estimates).

begun, so it is prudent to double check that all parameters are being monitored and that prior specification is as desired. Once WinBUGS has run through the desired number of iterations a message "Update took *x* secs." will appear in the bottom left hand corner of the screen and results become available for analysis.

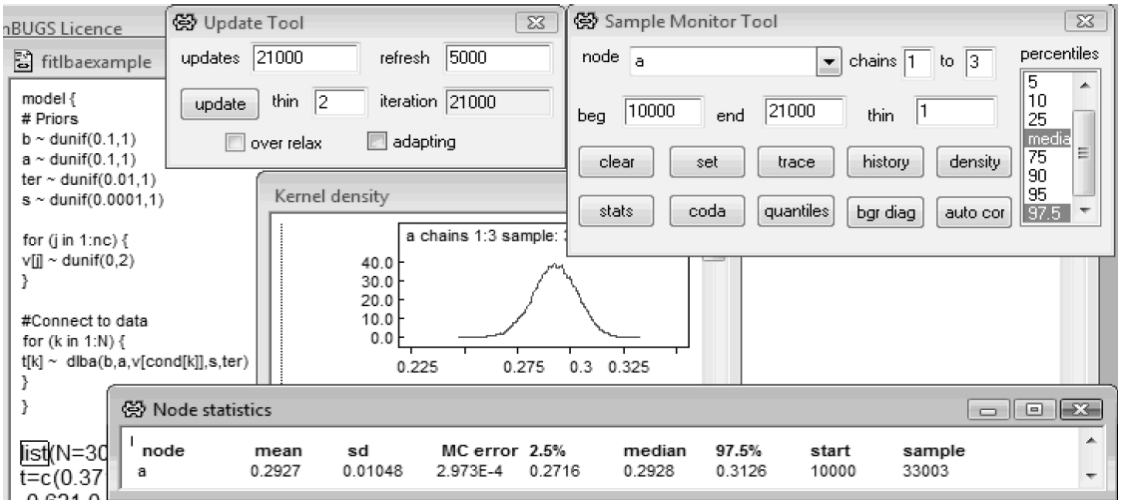

Figure 6 Screen grab from WinBUGS. Shown is the model code, the Update Tool, the Sample Monitor Tool and output from the density and stats options for parameter *A*.

To look at the results for each parameter return to the Sample Monitor Tool. Select the parameter of interest from the node drop down menu. Once a node is selected statistical and diagnostic options are highlighted. Amongst the many available choices we will focus on the 'density', 'stats' and 'compare' options. Figure 6 displays the 'stats' and 'density' outputs – 'Node statistics' and 'Kernel density', respectively – for the A parameter. Clicking on the 'density' option will display a density plot of the parameter of interest. This is a plot of the posterior estimates returned by WinBUGS for each of the iterations that was monitored. Once the MCMC chain has converged, (i.e., when the burn-in period is large enough), this density plot will approximate the marginal posterior distribution of the parameter. The quality of the approximation will increase as the number of iterations, or the length of the MCMC chain, increases. Figure

6 shows that, in our example, where 11,000 iterations were used to generate the posterior distribution, that the majority of the density plot is close to the true value of 0.3.

The ‘stats’ option provides commonly used statistics, such as mean and variance, as well as quantile information, for the chosen parameter. Generally, this summary provides the information used to derive parameter estimates for the LBA. Either the median or mean of the posterior distribution can be used as a parameter estimate. When the statistic is symmetrically distributed, as in Figure 6, there is little difference between these different estimates. The mode of the posterior distribution equals the maximum likelihood estimate of the parameter (e.g., as generated by our Excel worksheet). Although WinBUGS doesn’t directly return the mode of the distribution, the ‘coda’ option can be used to save monitored posterior samples to a text file which can be analysed in another statistical package to obtain the mode.

The WinBUGS ‘compare’ option, found in the inference drop down menu, can be used to obtain graphical representations of credible intervals. A credible interval estimates the range within which, given the data and the prior distribution, the true value of a parameter lies. Selecting the ‘compare’ option casues a dialog box to appear that requires at least one variable name to be entered. Type the variable of interest into the top left dialog box and select ‘box-plot’. This will produce a box-plot where the whiskers represent, by default, the 95% credible interval. The whiskers correspond to the 2.5% (lower whisker) and 97.5% (upper whisker) columns in the node statistics output, as credible intervals are based on the quantiles of the posterior distribution. Figure 7 shows the credible intervals for each of the three drift rates defined in the v parameter, the horizontal line going from one side to the other is the group median. The

plot also shows that the credible regions do not overlap suggesting that the drift rates differ from one and other.

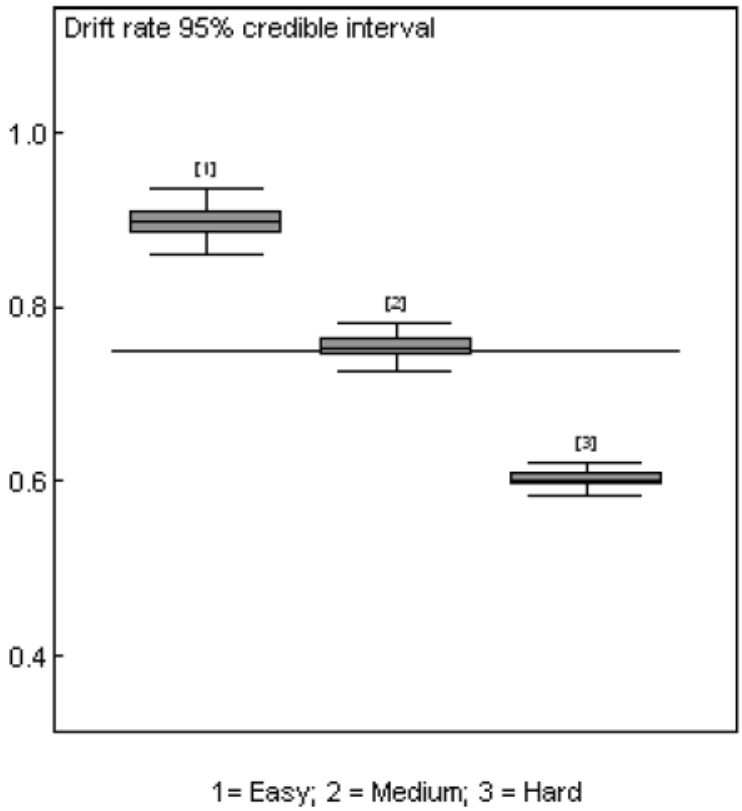

Figure 7 Box-plot representing the 95% credible regions of the drift rate for each of the three conditions, easy (1), medium (2) and hard (3). The line cutting through the centre of the plot represents the median of all three conditions.

The Sample Monitor Tool can also be used to check that the MCMC chain has converged to the posterior distribution using the 'history' option. An example of the output produced by this option is shown in Figure 8. The vertical axis of the plot indicates the parameter estimate for *A* for the iteration of the MCMC given by the horizontal axis – collapsing this plot onto the vertical axis gives the density function shown in Figure 6. Each of the chains is represented by a different grey scale and here we see the three MCMC chains for the *A* parameter in our example overlap greatly. In other words, they all appear to be random samples from the same distribution throughout the entire chain. This suggests that all chains are samples from the posterior

distribution of the parameter *A*. If any of the chains looked systematically different from the others, perhaps showing greater variance, or a different mean, it would suggest a lack of convergence of the MCMC chains to the true posterior distribution. The 'auto cor' option in the Sample Monitor Tool can also be used to check if further thinning is needed. It displays the correlation between parameter estimates for iterations *i* and *i* – *k* for *k* = 1 through 50.

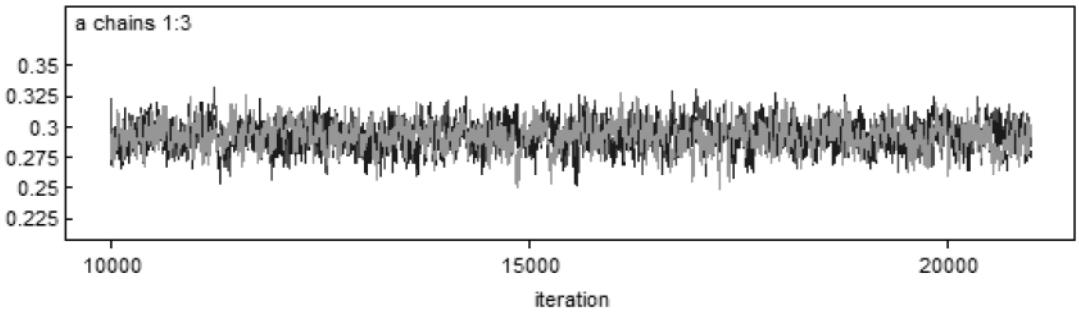

Figure 8 Output produced from the 'history' option for the *a* parameter from our fits of the LBA to example data. Notice that the three chains (indicated by different shades of grey) greatly overlap, indicating that all chains have converged upon the same posterior distribution.

Changing model parameterization is very simple within WinBUGS. The user must define a new prior for each of the parameters they want estimated and makes a small adjustment to the call to the LBA pdf. For example, say we were again fitting lexical decision data and wished to estimate a different non-decision time, T_{er}, for each word frequency condition. Then we need only augment the WinBUGS model specification to have a vector for the parameter T_{er} with a prior distribution for each of the three frequency conditions and make the call to the LBA pdf include this extra information. Specifically, to make our priors, when we would have previously used `Ter ~ dunif(0.1,1)` we instead use a `for` loop to set `Ter[k] ~ dunif(0.1,1)` for k in 1:3, as is done for drift rates. Finally, where we would have previously used `t[i] ~ dlba(b,A,v[cond[i]],s,Ter)` we now would use `Ter[cond[i]]`. To use the

WinBUGS code we provide with multiple choice data would require a substantial change to the code for the LBA pdf, 'lba.odc' along the lines of the R code we provide for fitting multiple choice data, but is beyond our scope here.

Using R to create a graphical summary given parameter values

We also provide R code that can be used to create a graphical summary of the data and model predictions. This process is useful in determining the appropriateness of the parameter estimates returned by our various methods. The code we provide requires that the user first enters the parameters produced by one of the three methods previously described (or indeed, any method). The user must then source the 'makegraphs.r' file within R, which defines two functions for producing two plots – histograms similar to those described in Example 1, and a quantile probability (QP) plot. The `qpplot` and `histplot` functions provide plots that are suitable for checking parameters, and can also be adapted to produce figures suitable for publications (see, for example, Maindonald, 2008, at http://cran.r-project.org/doc/contrib/usingR.pdf).

Ensure, first of all, that the R software is installed (refer to the guide in Example 2 if this is not done). Once installed, extract all of the files in the 'graphs.zip' file into the same folder. This folder should now contain 'pq-lba.r', 'lba-math.r', 'makegraphs.r' and 'exampledata.txt'. Now open R and make sure that the folder that the files were extracted to is set as the working directory in R (again, see Example 2). The user must first enter the parameters into a vector called `pars` in the following order: s, A, T_{er}, b, v_E, v_M, v_H. The units for A, T_{er} and b should be in milliseconds. This means that parameters from the WinBUGS version of the LBA, which are returned in seconds, will have to be multiplied by 1000. Parameter values should be entered as a vector, for example `pars=c(0.25,300,200,400,0.9,0.75,0.6)`. The user should then type

`source("makegraphs.r")`, which does two things: it first reads in the data from the 'exampledata.txt' file, and then defines two functions, ***histplot*** and ***qpplot***.

The ***histplot*** function produces a plot which contains six histograms, one for error responses and one for correct responses for each difficulty level. An example of this plot is shown in Figure 9. The data are represented by the black bars of the histogram, with the predictions of the model shown by the solid line. The top row of the plot shows the correct responses and the bottom row shows histograms for the error responses. The order of difficulty of the conditions is easy to hard from left to right. The ***histplot*** function has five arguments. Two are required: *data*, which must be formatted in the way that is produced by the ***read.table*** function contained within the 'makegraphs.r' script, and *pars*, which must be entered in the exact form given above. There are three optional parameters: *minx* and *maxx* define the smallest and largest RT values shown in the histogram, the third, *bindiff* defines how wide (in msec) the bins of the histogram are. It is essential that *bindiff* divides evenly into the difference between *minx* and *maxx*. To create the plots shown in Figure 9 we used the call: `histplot(data,pars)`.

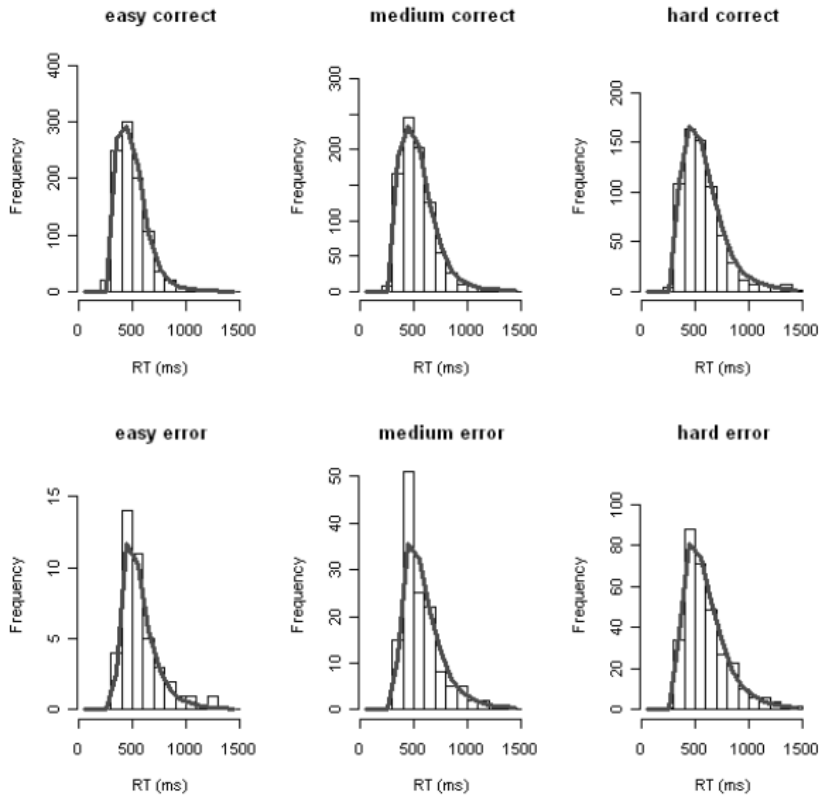

Figure 9 An example of the plot produced by the histplot function. Correct responses are shown in the top row, error responses on the bottom row. Difficulty of the decision goes from easy, medium to hard in order from left to right.

The ***qpplot*** function accepts four arguments. The two that are required – `dat` and `pars` – are of the same form as for the ***histplot*** function. The two optional arguments, `tmin` and `tmax`, define the fastest and slowest RT data points used to obtaining parameter estimates. They are, by default, set at 0 and ∞, respectively, indicating that no data were censored during estimation. Figure 10 shows an example of the QP plot produced by the ***qpplot*** function. The QP plot gives the majority of the important information shown by the histograms, but accomplishes this with one graph. It does this by taking each of the six histograms and summarising them with five quantile values. The quantiles for each of the histograms are placed onto the one plot. This results in the accuracies for the correct and error responses for the three difficulty conditions being indicated by the horizontal position of the six dots across the QP plot.

The right half of the plot (response probability above 0.5) shows the correct responses and the left half (response probability below 0.5) gives information about the error responses. The vertical position of the five points above each of these six accuracies refer to .1, .3, .5, .7, .9 quantiles of the RT distribution for each of the correct and error responses in the three difficulty conditions. The quantile values are the proportion of responses under which a given proportion of RTs in the distribution fall (e.g. the 0.5 quantile is the median). As an example, consider the bottom right point of the plot. The rightmost points of the plot refer to those decisions with the highest accuracy – in other words, the RTs from the correct responses in the easiest condition. Conversely, the left-most points are the error responses in the easiest condition. The bottommost point on the QP plot refers to the .1 quantile of the RT distribution. The .1 quantile of the RT distribution gives the value below which 10% of the RTs fall. Hence, the bottom right point of the QP plot gives the value below which 10% of the RTs for the correct responses in the easiest condition occur. To make the plot shown in Figure 10 we used the call: `qpplot(data,pars)`

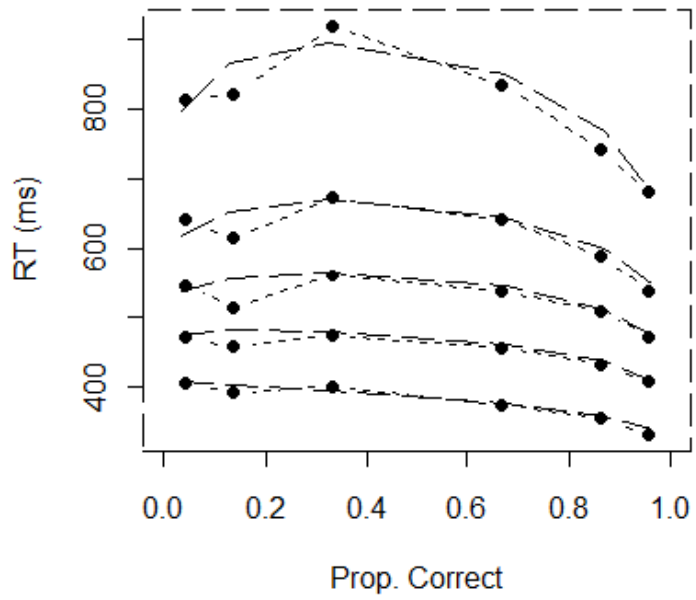

Figure 10 An example of the plot produced by the qpplot function. Proportion correct is shown on the x-axis, reaction time (in ms) shown on the y-axis. Data is shown by the dotted line with filled points, the LBA predictions are shown by the solid line.

The Effect of Sample Size on Parameter Estimates

We have provided four different methods for fitting the LBA to data – one of these specifically for fitting multiple choice data. For the other three cases we have fit the model to a set of simulated data with 1000 observations in each of three conditions. In practice, there are often considerably fewer observations per condition. To investigate how well parameters for a two choice task are recovered by each of our methods for a range of sample sizes, we conducted a simulation study. We simulated ten sets of data for each of four different sample sizes – N = 50, 100, 400 and 1000 observations per condition. The data were simulated using the same parameter values used to generate our example data, and are shown in Table 1.

Table 1 Bias and standard deviation of parameter estimates, as a percentage of the true parameter value, from three methods of fitting the LBA for four different values of *N* (samples per condition) averaged over ten data sets. The three methods are E: Excel, R: R, and W: WinBUGS. The final column contains the average time, *t*, taken per simulation (in seconds). Times were estimated using a single core of a Pentium quad-core Q6600 2.4GHz processor.

True	v_E 0.9		v_M 0.75		v_H 0.6		*A* 300		*b* 400		T_{er} 300		*s* 0.25			
	bias	sd	bias	sd	bias	sd	bias	sd	bias	sd	bias	sd	bias	sd	*N*	*t*
E	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA
R	-11	27	-2.7	12	-3.3	6.7	6.7	33	4.3	12	-4	36	-4	20	50	48
W	-7.8	7.8	-9.3	9.3	-3.3	5	-4.3	16	-4.8	4.5	-1.3	10	-16	24		124
E	5.6	5.6	6.7	2.7	1.7	3.3	5	11	2.8	4	6	11	28	8		2
R	4.4	8.8	5.3	4	0	5	7.3	10	-1.5	7	12	15	12	20	100	48
W	-2.2	4.4	-2.7	4	0	5	1	10	-1.5	6.3	2.7	6.7	-4	8		231
E	2.2	2.2	4	2.7	1.7	1.7	-2.3	6	0.5	2.5	-0.7	2	20	4		9
R	0	6.7	1	5.3	1.7	3.3	2	10	-0.25	2.4	1.3	14	4	8	400	48
W	0	3.3	-1	2.7	0	1.7	0	4.3	3	9.5	-1.3	5.7	0	8		934
E	4.4	1.1	4	1	1.7	1.7	0.7	3	1.3	1.5	2	3	20	4		21
R	2.2	2.2	2.7	2.7	0	1.7	3	5	-0.25	1.8	5	7	8	8	1000	48
W	1.1	2.2	0	1	0	1.7	2	4	2.8	8.3	1.3	3	0	4		2308

Table 1 shows the average bias and standard deviation in parameter estimates, expressed as a percentage of the respective parameter value, for each of our three methods – the Excel sheet, the R code and WinBUGS[5]. For all methods we observe the expected pattern that as sample size decreases the bias and standard deviation of parameter estimates increase. The size and rate at which this happened varied between our methods. When sample size was only 50 observations per condition the Excel sheet failed to recover parameters. Note, however, that once sample size increased to 100 observations per condition that the parameters were recovered reasonably well even by the Excel sheet, perhaps with the exception of *s*. Note also that for the Excel sheet, although there was a reasonable reduction in both bias and standard deviation of parameter estimates when *N* increased from 100 to 400, the increase from 400 to 1000 made very little difference. For R and WinBUGS when *N* is only 50 the drift rate in high

5 We use the mean of the posterior distribution to determine bias in parameter estimates in WinBUGS. Note that we could have also used an alternate measure of central tendency such as the median.

accuracy condition is overestimated. This is reasonable because, with only 50 samples and high expected accuracy, there are very few error responses in this condition. Note that for 100 samples per condition or more that there is relatively little bias in parameter recovery for any of the techniques and the standard deviations for each of the parameters are small and decrease at a rapid rate as *N* grows.

Fixing Parameters Across Conditions

When estimating model parameters, we made the assumption that only drift rate should vary between conditions. This assumption is the one usually made when the data come from an experiment where the conditions correspond to different stimuli presented within-subjects and that vary unpredictably from trial to trial. This is because parameters such as *b*, which determines the amount of evidence required to make a response, are thought to be under the strategic control of the participant. Ratcliff (1978) argued that these participant-determined parameters can not be adjusted on a trial-by-trial basis depending on which stimulus is presented. If, however, we were to fit data with conditions that varied between blocks of trials, or between participants, then it is reasonable to expect that parameters such as *b* could vary across these conditions. For example, if participants were instructed to respond as accurately as possible in one block of trials, given a break and then told that for the next block of trials to respond with speed emphasis, then we could expect that the participant has been given enough time to adjust their cautiousness in responding by adjusting their *b* parameter.

In our simulated example, because we knew exactly what parameters generated the data, it was straightforward to decide which parameters should vary across conditions. In practice, we will not necessarily know what parameters are expected to vary across conditions. Researchers should, therefore, fit a number of different versions

of the LBA where we change which parameters are allowed to vary across conditions and then select the model which provides the best account of our data. This approach is not straightforward, however, because adding extra parameters will always give a fit that is at least as good as the less complex model, even if the extra parameters overfit (i.e., only accommodate noise in) the data. What is required, therefore, is a measure which only improves if the extra parameters provide a genuine improvement. This is usually accomplished by penalizing a model for having extra parameters. Many such measures exist, but we focus on three easily computed options, the Akaike information criterion (AIC), Bayesian information criterion (BIC) and Deviance information criterion (DIC). Each measure uses deviance (-2 times the log likelihood) as its measure of misfit but applies a different complexity penalty. BIC provides the largest penalty for having more parameters – $k\log(N)$, where k is the number of parameters in the model and N is the number of data points, AIC applies the smallest penalty – $2k$, and DIC, which can only be calculated from Bayesian outputs, applies a penalty which is often somewhere between AIC and BIC in its severity. The DIC measure is based on an estimate of a model's effective number of parameters, pD, which takes account of differences in the functional form between models (see Spiegelhalter, Best, Carlin & van der Linde, 2002 for details of the calculation of pD). For each of these measures, the model that produces the smallest value is the one that best accounts for the data, given both goodness of fit and model complexity.

To demonstrate these model selection methods we fit our example data, where we know that only drift rate varied between conditions to generate the data, with two different versions of the LBA – one with only drift rate varying between conditions and another where b, A, T_{er} and v were allowed to vary across conditions. We report the

results of using WinBUGS to estimate parameters here, however, when we used our R code we found the same pattern of estimates. The deviance for the more complex model was -293.6 compared with -294 for the model where only drift rates were allowed to vary. In other words, there was very little improvement in the quality of the fit when parameters other than drift rate were allowed to vary across conditions. After adding the various complexity penalties, all three measures of model fit were smaller for the LBA when only drift rate varied across conditions (AIC: -277.99 vs. -277.59; BIC: -238 vs. -190; DIC: -287 vs. -281). This tells us that allowing parameters other than drift rate to vary across conditions gives an increase in quality of fit which is not large enough to warrant the complexity of the extra parameters. Indeed, when we looked at the parameter values estimated in the LBA where b, A, T_{er} were also allowed to vary, we observed almost no change across difficulty conditions. The same principles can be used to try any number of other parameter constraints, such as allowing fewer parameters to change across conditions.

General Discussion

We have provided four different methods for fitting the LBA to data – one of these specifically for fitting multiple choice data. Our aim was to provide the potential user of the LBA with three separate methods for implementing estimation. We (and others; e.g. Wagenmakers et al., 2007) would argue that mathematical models of choice, such as the LBA, can provide an important tool for data analysis that can provide much more information about decision processes than the typical ANOVA method applied to RT and accuracy. We have provided three separate methods of estimation to data to ensure that the LBA is accessible to users with a range of different levels of programming and mathematical ability. The Excel spreadsheet is straightforward to

apply to new data which is fairly similar to that from our example data (i.e. a one within-subjects factor). Given R's flexibility and computational power, our R code can be extended to fit accuracy and RT data from almost any experimental set up. However, this requires some programming knowledge and changes to not only the 'fit-example.r' script, but also the 'pq-lba.r' code. We included the WinBUGS implementation of the LBA because, as Vandekerckhove et al. (submitted) argue, it offers a highly flexible framework for model fitting which is accessible to someone with relatively little computing background. One can adjust which parameters vary between conditions, regardless of the number of different conditions or variables in the simple way we previously discussed. We direct the reader interested in possible hierarchical extensions of the LBA, or diffusion model, to Vandekerckhove et al.'s discussion.

The intent of this article was to provide multiple ways to apply the LBA to data, not to compare these methods. As shown in Table 1, all methods recovered parameters quite accurately when applied to data with 100 or more observations per condition. The WinBUGS method provided parameter estimates which were generally the closest match those used to produce the data. However, the WinBUGS method took, by far, the longest (around 4 hours, with the R and Excel methods taking around 1 minute). The QMPE method used in the R code is more resilient to smaller sample sizes and outlying data points than the maximum likelihood method used in the Excel code and the multiple choice R code (Heathcote, Brown & Mewhort, 2002). In the Bayesian framework hierarchical methods, that provide parameter estimates at the population rather than individual participant level, are an effective way of dealing with small samples per participant when data from a large number of participants is available.

References

Brown, S.D. & Heathcote, A. (2003). QMLE: Fast, robust and efficient estimation of distribution functions based on quantiles. *Behavior Research Methods, Instruments, & Computers, 35*(4), 485-492.

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117-128.

Brown, S.D., & Heathcote, A.J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153-178.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences, 23,* 255-282.

Carpenter, R. H. S. (2004). Contrast, Probability, and Saccadic Latency: Evidence for Independence of Detection and Decision. *Current Biology, 14,* 1576-1580.

Calin, B.P & Chid, S (1995) Bayesian Model Choice via Marcov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 473-484

Donkin, C., Brown, S., & Heathcote, A. (2009). The over-constraint of response time models. *Psychonomic Bulletin & Review, 16,* 1129-1135.

Donkin, C., Heathcote, A., Brown, S. & Andrews, S. (2009). Non-decision time effects in the lexical decision task. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Forstmann, B.U., Dutilh, G., Brown, S.D., Neumann, J., von Cramon, D.Y., Ridderinkhof, K.R., & Wagenmakers, E.J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of*

Science, 105, 17538-17542.

Gold, J., & Shadlen, M. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Science, 5*(1), 10-16.

Hanes, D. P., & Carpenter, R. H. S. (1999). Countermanding saccades in humans. *Vision Research*, *39*, 2777-2791.

Heathcote, A., & Brown, S.D. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin and Review, 11*, 577.

Heathcote, A., Brown, S.D. & Mewhort, D.J.K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin and Review, 9*(2) 394-401.

Ho, T.C., Brown, S.D., & Serences, J.T. (submitted). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience.*

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60,* 121-123.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325--337.

Maindonald, J. (2008). *Using R for Data Analysis and Graphics - Introduction, Examples and Commentary.* Web-based article downloaded on 8/12/08 at http://cran.r-project.org/doc/contrib/usingR.pdf.

Mazurek, M., Roitman, J., Ditterich, J., & Shadlen, M. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex 13*(11), 891-898.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-163.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59 –108.

Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278-291.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two choice decisions. *Journal of Neurophysiology, 90*, 1392-1407.

Ratcliff, R. Gomez, P., & McKoon, G.. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111,* 159-182.

Ratcliff, R., & Rouder, J. N. (1998). Modelling response times for two choice decisions. *Psychological Science, 9,* 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16,* 323–341.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65,* 523–535.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating the parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481.

Reddi, B.A.J. (2001). Decision making: The two stages of neuronal judgement. *Current Biology*, *11*, 603-606.

Roitman, J. & Shadlen, M. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of*

Neuroscience, 22(21), 9475-9489.

Schall, J. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews in Neuroscience, 2*, 33-42.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44,* 408–463.

Smith, P. L., & Ratcliff, R. (2004). Psychology and Neurobiology of Simple Decisions. *Trends in Neurosciences, 27*, 161-168.

Smith, P. L., & Ratcliff, R. (in press). An integrated theory of attention and decision making in visual signal detection. *Psychological Review.*

Spiegelhalter, D. J., Best, N. G., Carlin, b. P., & van der Linde, A. (2002) Bayesian measure of model complexity (with discussion). *Journal of the Royal Statistical Society B., 64,* 583-640.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7,* 424-465.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7,* 208-256.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011-1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*, 61-72.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (submitted). Hierarchical diffusion models for two-choice response times. *Psychological Methods*.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical

discrimination. *Ergonomics, 13,* 37-58.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206-1220.

Voss, A., & Voss, J. (2007). Fast-DM: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767-775.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion-model parameters. *Journal of Mathematical Psychology, 52*, 1-9.

Wagenmakers, E.-J. (in press). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology.*

Wagenmakers, E. J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses*, pp. 181{207. Springer: New York.

Wagenmakers, E.-J., van der Maas, H.L.J., & Grasman, R.P.P.P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review 4, 3-22*.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44,* 92-107.

Appendix: Setting up WinBUGS

WinBUGS can be obtained from http://www.mrc-bsu.cam.ac.uk/bugs/. To install WinBUGS you will first need to download the install file (WinBUGS14.exe); once downloaded run the executable and it will, by default, install WinBUGS to the program files folder. Note that the install directory may be different for operating systems other than Windows XP – the reader need only take note of their WinBUGS install directory and adjust any future folder references we make. Next, you are required to fill out a short registration form that will allow a registration key to be sent to the email address you provide. The email will contain the registration key and instructions on how to register WinBUGS.

Although WinBUGS has a large number of pre-specified distributions for which it can conduct a Bayesian analysis, it does not have the appropriate *probability density function* (pdf) for the LBA. We have, therefore, provided the pdf for the LBA in the 'BugsLBA.zip' folder. Making the LBA *pdf* accessible to WinBUGS necessitates the use of two additional pieces of software: the *Black-box component Builder* and the *WinBUGS Development Interface (WBDev)*. Instructions for their installation are as follows

1. Extract the lba.odc and Distributions.odc files from the BugsLBA.zip folder.
2. Download the *WinBUGS Development Interface (WBDev)* from *http://www.winbugs-development.org.uk/*. From the home page navigate to the WinBUGS Development Interface page and download the software. The contents of the zip file should be unpacked into the WinBUGS directory. Open the .txt (wbdev_01_09_04.txt at the time of writing) file that you just extracted and follow the instructions contained in the file to install the WBDev software.

3. Download the *BlackBox Component Builder* from http://www.oberon.ch/blackbox.html. In the current paper we use version 1.5. Once downloaded, run the SetupBlackBox15.exe file which will install the BlackBox Component Builder 1.5. This will add a new folder to C:\Program Files called BlackBox Component Builder 1.5.

Once you have all the programs necessary the next step is to compile the LBA pdf into WinBUGS via Black-box. After completion of the steps below you will be able to use WinBUGS to fit the LBA to data.

1. Open the WinBUGS directory and copy the entire contents of the WinBUGS folder and paste them into the newly created BlackBox directory (C:\Program Files\ BlackBox Component Builder 1.5\ by default in Windows XP); choose yes to all the "replace existing file" requests.

2. Now copy the lba.odc file to the C:\Program Files\BlackBox Component Builder 1.5\WBDev\Mod directory.

3. Open the BlackBox Component Builder program; this should now closely resemble the usual WinBUGS environment. Use File > Open to open the lba.odc file. Use Ctrl+K to compile the lba.odc file. An "ok" message should appear in the bottom left corner.

4. Lastly, put the Distributions.odc into the C:\Program Files\BlackBox Component Builder 1.5\WBDev\Rsrc\ directory. Close down any still running BlackBox or WinBUGS windows. The next time BlackBox is run then the LBA pdf should be ready to use.

For more information on the procedure outlined above as well as the use of diffusion models in WinBUGS see Vandekerckhove, Tuerlinckx and Lee (submitted).

Non-Decision Time Effects in the Lexical Decision Task

Christopher Donkin[1], Andrew Heathcote[1], Scott D. Brown[1],
& Sally Andrews[2]

1. University of Newcastle
Newcastle, Australia

2. University of Sydney
Sydney, Australia

Abstract

It has been argued that performance in the lexical decision task (LDT) does not provide a direct measure of lexical access because of the effect of decision processes. We re-examine LDT data and fits of the diffusion decision model reported by Ratcliff, Gomez and McKoon (2004) and show that they assumed too little role for non-decision processes in explaining the word frequency effect. Our analysis supports an effect of frequency on decision *and* non-decision time.

Reading is one of the most remarkable abilities achieved by the human mind. One of the key aspects enabling reading is the ability to recognize a string of characters as being a word, a process called "lexical decision". The lexical decision task (LDT) is a paradigm for studying word identification in which participants are presented with a string of letters and they must quickly decide whether or not the letters form a word. If the letters presented do make a word, then the time taken to make a 'word' response is thought to give information about how long it took to retrieve the word from their database of words, a process referred to as lexical access.

The *word frequency* effect is one of the most robust findings from the LDT paradigm: words used less frequently in natural language take longer to indentify than higher frequency words. Historically, the word frequency effect has been reported as a difference in mean reaction time (RT) for correct responses between low and high frequency words. Mean RT from high and low frequency words usually differs by around 60-80ms. However, RT in the LDT is quite variable, typically having a standard deviation of greater than 100ms. Some of this variability is because of differences between words within a frequency class, but variability also occurs between the same word on different occasions. Variability in RT is also positively skewed, with a longer right (slow) than left (fast) tail in RT distribution, and the length of the right tail has been found to vary systematically in LDT experiments. Hence, researchers have begun to investigate differences in the entire RT distribution between high and low frequency words, rather than just the mean RT (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Plourde & Besner, 1997). More recently, there have been lexical theories proposed that account for effects on all aspects of RT distribution (Ratcliff, Gomez and McKoon, 2004; Yap, Balota, Cortese & Watson, 2006).

RT distributions have been shown to be well characterized by the ex-Gaussian distribution (Luce, 1986). The ex-Gaussian distribution is produced by convolving (i.e., adding samples from) the Gaussian and Exponential distributions. It has three parameters, the mean (μ) and standard deviation (σ) of the Gaussian component and the mean of the exponential component (τ). These parameters give information about the shape of the RT distribution. In particular, the μ parameter is affected by the speed of the fastest responses made by participants. Similarly, the τ parameter is affected by the length of the right tail of the RT distribution.

Differences in parameter estimates from fits of the ex-Gaussian to high and low frequency RT distributions indicate that there are changes in the very fastest and slowest responses made by participants. Changes in μ of approximately 20-30ms have been reported (Andrews & Heathcote, 2001; Balota & Spieler, 1999; Plourde & Besner, 1997). These changes indicate that the entire RT distribution shifts to be slower for less frequent words, independently of any changes in the shape of the distribution. In the same applications of the ex-Gaussian, changes in τ of approximately 35-45ms were observed, suggesting that the right tail is longer when the words to be identified are less frequent.

Balota and Chumbly (1984) argued that the data from LDT tasks come from a combination of the lexical process *and* the decision process. Ratcliff et al. (2004) furthered this line by arguing information about lexical access can only be obtained from RT *after* accounting for the decision process. In other words, even studying the full range of behavioral data in the LDT (i.e., accuracy and RT distributions for correct and error responses) does not by itself provide clear information about lexical access. To address this issue they fit a model of the decision process, the diffusion model, to their

LDT data and used estimates of its parameters, and the parameters of a simple characterization of non-decision processes, to examine lexical access. When Yap et al. (2006) compared the diffusion account with a hybrid two-stage model of the LDT based on Balota and Chumbly's work, they concluded in favor of the diffusion model.

The diffusion model account of RT is composed of two parts – a decision time and a non-decision time. The account of LDT starts by assuming that a stimulus is perceived and encoded. This is followed by lexical access, which gives an estimate of how much evidence the stimulus provides for each response (word and non-word in an LDT). This evidence determines the rate at which information is accumulated, called *drift rate*, and drives the decision part of the diffusion model. The time taken for the initial perceptual, encoding and lexical access processes, plus the time to execute the motor response after the decision process is completed, makes up the non-decision time. The non-decision time, T_{er} in the diffusion model determines the smallest possible RT and, therefore, changes in T_{er} shift the entire RT distribution. The ex-Gaussian evidence reviewed above might have suggested that the word frequency effect would, in part, be explained by differences in T_{er} for high and low frequency words. However, when Ratcliff et al. (2004) applied the diffusion model to data from nine LDT experiments they concluded that only drift rate differed between high and low frequency words. In other words, word frequency effects in the LDT were simply due to how 'wordlike' the string of letters was, and not caused by other aspects of the non-decision processing, such as the time required for lexical access. Ratcliff et al. claimed that the shift of the RT distribution due to word frequency is captured by the inclusion of trial-to-trial variability in T_{er} and not due to systematic differences in T_{er} determined by the frequency of the word being identified.

In the current paper we re-analyze Ratcliff et al.'s (2004) data and demonstrate that their fits of the diffusion model systematically fail to account for the word frequency effect on both fast and slow responses. We then show that the misfit is greatly reduced by allowing T_{er} to differ for words of different frequency. We finish by discussing the implications of our results and possible extensions. First, however, we begin by describing the diffusion model.

The Diffusion Model

The diffusion model with trial-to-trial variability in parameters is the most successful model of choice and reaction time for simple decisions between two alternatives (Ratcliff, 1978) and has been applied repeatedly to LDT data since Ratcliff et al.'s (2004) initial work (Gomez, Ratcliff & Perea, 2007; Ratcliff, Perea, Colangelo, & Buchanan, 2004; Wagenmakers, Ratcliff, Gomez & McKoon, 2008). The diffusion model assumes that participants sample evidence from the stimulus continuously, and this evidence stream updates an evidence total, say x, illustrated as a function of time by the irregular line in Figure 1. The accumulator begins the decision process in some intermediate state, say $x=z$. Evidence that favors the response "word" increases the value of x, and evidence that favors the other response ("non-word") decreases the value of x. The evidence accumulation process continues until sufficient evidence favors one response over the other, causing the total to reach one of its two boundaries (the horizontal lines at $x=0$ and $x=a$ in Figure 1). The choice made by the model depends on which boundary is reached (a for a "word" response or 0 for a "non-word" response) and decision time equals the accumulation time.

Depending on the stimulus, evidence tends to accumulate more towards one boundary or another, and the average rate of this accumulation is called the "drift rate",

which we will label *v*. Larger positive or negative drift rates cause faster and more accurate responses as evidence heads towards the correct boundary at a faster rate. The evidence accumulation process also varies randomly from moment-to-moment during the accumulation process, and the amount of this variability is another parameter of the model, *s*. The diffusion model used in Ratcliff et al. (2004) also includes three extra variability parameters, the distribution of drift rates is assumed to vary from trial-to-trial according to a normal distribution with mean *v* and standard deviation *η*. Start point is also assumed to vary from trial-to-trial according to a uniform distribution with centre *z* and range s_z. Finally, non-decision time is assumed to vary between trials according to a uniform distribution with centre T_{er} and range s_T. Critically, non-decision variability enables the diffusion model to better account for shifts in RT distribution between conditions that differ only in drift rate. When there is no non-decision variability a change in drift rate almost exclusively slows RT by lengthening the right tail of the distribution, with only a small effect on the fastest RTs. When non-decision variability is added the effect of a drift rate change on fast RTs is increased sufficiently so that Ratcliff et al. (2004) were satisfied with an account of the word frequency effect in terms of a pure selective influence on drift rate.

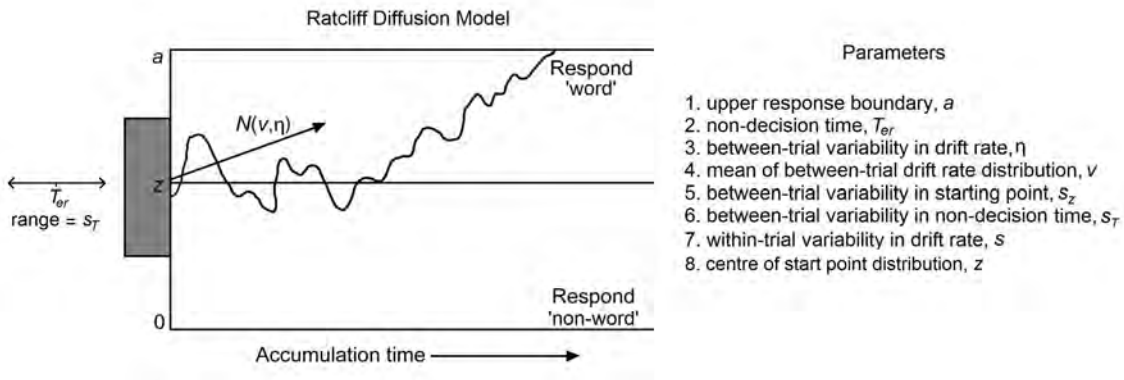

Figure 1 A graphical representation of a single diffusion model decision in an LDT task

Ratcliff et al.'s (2004) LDT Data

Fits reported in the original paper

Ratcliff et al.'s (2004) fits to all experiments were accomplished by allowing only drift rate to vary between word frequency conditions. This is common practice when applying the diffusion model. Differences in non-decision process parameters cannot be the sole account for word frequency effects, as these processes cannot influence error rates. However, although less parsimonious, there is no reason why non-decision processes might not be affected by word frequency in addition to drift rates. Indeed, Ratcliff et al.'s (2004) application of the diffusion model to the LDT was one of the first occasions on which non-decision variability was used, with most earlier applications assuming a constant non-decision time (e.g., Ratcliff, 1978).

When we looked closely at Ratcliff et al.'s (2004) published fits of the diffusion model to their LDT data averaged over participants, we found a systematic pattern of misfit that was highly consistent across all of the nine experiments which they report. In particular, despite the inclusion of between-trial variability in T_{er}, the diffusion model consistently under-predicted the magnitude of the word frequency effect on the .1 quantile results for correct responses reported by Ratcliff et al.. The .1 quantile characterizes the fastest responses from the RT distribution (i.e., it is the RT below which the fastest 10% of responses occur). Changes in the .1 quantile indicate a shift in the entire RT distribution. Averaging over their nine experiments, the .1 quantile estimate for high frequency words was 27ms and 33ms faster relative to low and very low frequency words respectively, whereas for the model it was only 16ms and 22ms faster. Although the under-prediction is relatively small (11 ms on average), it is highly consistent, occurring in every one of the 19 fits reported in their Tables 3, 7 and 9 - a

highly significant result using a binomial test ($p<.001$ for both low and very low frequency words). In contrast to results for the fast .1 quantile, the diffusion model consistently over-predicted the word frequency effect for the slow .9 quantile, for nine of ten fits comparing high and low frequency words ($p<.001$) and seven of nine fits comparing high and very low frequency words ($p<.02$).

Figure 2 is a graphical summary of these analyses of data and model fits for high and low frequency words averaged over experiments from Ratcliff et al. (2004). Though it was excluded for brevity, the plot of the difference between high and very low frequency words looks almost identical. The vertical axis shows the difference in RT between low and high frequency words. Note that the positive value of this difference means that participants were slower to respond to low frequency words – the standard word frequency effect. The horizontal axis represents the quantile values of the RT distribution. The average model predictions (shown by the solid line) for the.1 quantile fall below the observed data averaged across all experiments. Note also that the opposite is true for the .9 quantile – the average model predictions sit higher than the data in both plots. The systematic and opposite misfit for fast and slow responses resulted in over prediction of the effect of word frequency on variability (i.e., a much larger range between the 10% and 90% quantiles than observed in data).

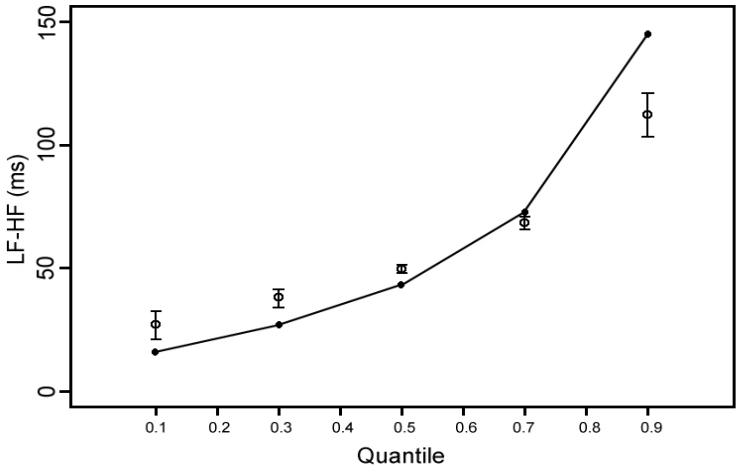

Figure 2: Word frequency effect quantile function based on responses to high frequency (HF) and low frequency (LF) words in Ratcliff et al.'s (2004) experiments 1-9. Average model fits across experiments and conditions are plot as lines, and data as symbols. Standard error bars indicate variability across experiments and condition

The diffusion model has clearly raised the bar for accounts of LDT performance by simultaneously fitting accuracy and RT distribution for both correct and error responses. Although we agree that the diffusion model provides an impressively comprehensive account of many aspects of performance in the LDT, the systematic misfit of the word frequency quantile functions indicates that there may be reason to re-examine the assumptions made by Ratcliff et al. (2004) in their application of the diffusion model.

The diffusion model appears to have misfit Ratcliff et al.'s (2004) data largely because the assumptions underlying the mapping of the diffusion model to the LDT task are too simple. Although simplicity is a virtue in quantitative modeling, identifying word frequency effects entirely with drift rate may represent an over-application of Occam's razor. Most models of reading assume that lexical access is accomplished more quickly as the frequency of a word increases (see Andrews & Heathcote, 2001, for a

discussion). In the diffusion model framework, this could be interpreted as a faster non-decision time for high than low frequency words. Allowing for such a possibility might reduce the underestimation of the word frequency effect at the .1 quantile apparent in Figure 2. In other words, perhaps the diffusion model would provide a better account of the word frequency effect in LDT data if it were to also allow for changes in T_{er} for words of different frequency. We explore this possibility in the next section.

Exploring frequency effects on non-decision time

We fit four different versions of the diffusion model to data averaged over participants from Experiments 3, 4 and 5 from Ratcliff et al. (2004). All experiments were of nearly identical procedure, with differences being in the type of words used: Experiment 3 used high frequency, low frequency and pseudo-words, Experiment 4 was identical but used random letter strings instead of pseudo-words, and Experiment 5 was the same as Experiment 3 but also included very-low frequency words. Our re-analyses was limited to these three experiments because Ratcliff et al. did not publish critical information for fitting (e.g., quantiles for error RT) for the remaining experiments. The four versions of the diffusion models differ according to how non-decision time, T_{er}, varied. There were two ways in which T_{er} was allowed to vary – randomly between trials (cf. Ratcliff et al., 2004) or systematically between word frequency conditions.

Between-trial variation was uniformly distributed with mean T_{er} and range s_T. Between-condition variation in T_{er}, like between-condition variation in drift rate, meant that each of the word conditions had its own T_{er} value. The between-trial variability in T_{er} requires one parameter, s_T, whereas between-condition variability in T_{er} requires the estimation of an additional k-1 parameters, where k is the number of word frequency conditions in the experiment being fit. The four different models were factorial

combinations of these two methods: 1) neither between-trial nor between-conditions variability in T_{er}, 2) only between-trial variability in T_{er}, 3) only between-conditions variability in T_{er}, and 4) both between-trial and between-conditions variability in T_{er}. The data to be fit were accuracy and quantile values for correct and error responses averaged over participants from each experiment. We fit the diffusion model using an adaptation of Voss and Voss's (2008) diffusion model code to use quantile maximum likelihood estimation (Heathcote, Brown & Mewhort, 2002). The Bayesian information criterion (BIC) was calculated using the BIC statistic for N observations grouped into bins:

$$BIC = -2(\sum_i Np_i \ln(\pi_i)) + M \ln(N)$$

where p_i is the proportion of observations in the i^{th} bin, and π_i is the proportion of observations in the i^{th} bin as predicted by the model. M is the number of parameters of the model used to generate predictions. The BIC is composed of two parts, the first is a measure of misfit, and a second part, $M\ln(N)$, penalizes a model for its complexity as indicated by the number of estimated parameters. When comparing two models, the model with the smaller BIC is thought to have provided a better fit after complexity has been taken into account. Best fitting parameter estimates for each of the four models to all three experiments and their respective BIC values are given in Table 1.

Despite the complexity of the analysis, the pattern of results was relatively simple. Adding between-trial variability in T_{er} always improved the BIC value, and so too did adding between-condition variability in T_{er}. In all three experiments the model with both between-trial and between-condition variability in T_{er} had the lowest BIC. This implies that the improvement in fit due to the extra free parameters outweighed the penalty for added complexity. The next best fitting model in two out of three

experiments was the model used to originally fit the data in Ratcliff et al. (2004) – the model with between-trial variability in T_{er}. In Experiment 5 not the model without between-trail variability in T_{er}, but with between-condition variability in T_{er} achieved the second best fit.

Table 1: Parameter estimates from fits of four different versions of the diffusion model to Experiments 3-5. M1 was the model with no variability in T_{er}, M2 had variability between-trials, M3 had variability between-conditions and M4 had both. In all models starting point, z, was set at $a/2$.

	Model	a	s_z	η	v_h	v_l	v_o	v_v	s_t	T_{er}	T_{er} *HF*	T_{er} *LF*	T_{er} *O*	T_{er} *VLF*	BIC
Exp3	M1	.128	.059	.037	.348	.176	-.226			.404					91887
	M2	.122	.069	.108	.446	.219	-.282		.17	.444					91126
	M3	.127	.065	.052	.335	.188	-.243				.396	.421	.422		91449
	M4	.122	.076	.113	.412	.226	-.301		.16		.428	.451	.461		**90843**
Exp4	M1	.133	.08	.089	.367	.361	-.302			.378					98571
	M2	.126	.075	.101	.381	.361	-.366		.11	.39					98415
	M3	.132	.081	.093	.37	.319	-.358				.379	.391	.375		98453
	M4	.127	.078	.011	.391	.334	-.374		.105		.392	.404	.387		**98320**
Exp5	M1	.147	.069	.069	.354	.214	-.259	.128		.409					89190
	M2	.144	.075	.01	.394	.234	-.253	.141	.139	.431					89000
	M3	.144	.074	.074	.336	.243	-.217	.132			.402	.435	.429	.425	88693
	M4	.148	.093	.124	.404	.257	-.296	.163	.125		.422	.451	.461	.454	**88546**

The model with neither between-trial nor between-condition variability in Ter consistently performed the worst of the four models. Inspection of the fits revealed that, as expected, this model predicted almost no change in the .1 quantile due to changes in word frequency. Because of this it was also unable to capture other aspects of the RT distribution. Hence, we do not consider the model without variability in T_{er} any further. Although, for brevity, we do not show the complete fits of the model to quantiles for correct and error responses for all word frequency conditions, these graphs clearly agree with our conclusions based on BIC values (they may be obtained by emailing the authors).

Our reason for investigating between-condition variability in T_{er} was based on the systematic misfit of the word frequency effect. Figure 3 shows that there is an

improvement in the account of the word frequency effect when between-condition variability in T_{er} is added to the diffusion model. The plots in Figure 3 are like those in Figure 2, but are from individual experiments rather than averaged across all nine experiments in Ratcliff et al. (2004). Each of the three plots also now contains three sets of model predictions (represented by solid lines) rather than one. The filled black dots represent the difference between RTs from high and low frequency words at each of the .1, .3, .5, .7 and .9 quantiles from the data. For all experiments we again observe that the difference between low and high frequency words is positive at all quantile values. This suggests that the RT distribution for low frequency words is shifted above that of high frequency words.

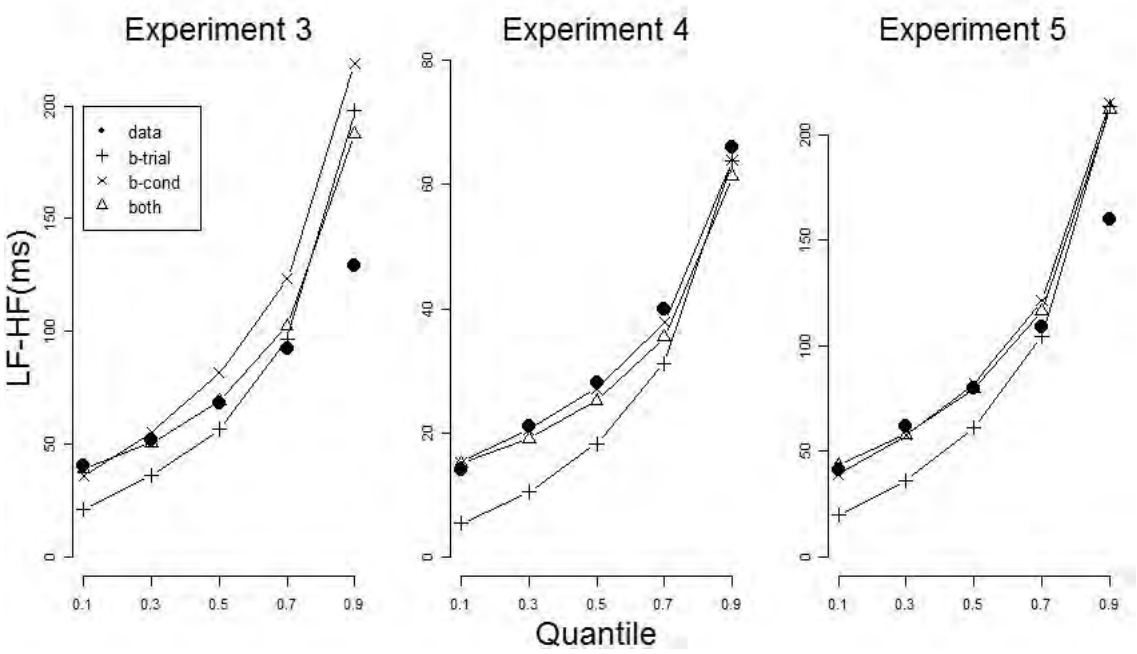

Figure 3: Word frequency effect quantile function based on responses to high frequency (HF) and low frequency (LF) words in Ratcliff et al.'s (2004) experiments 3-5. Data are shown as filled black dots and model predictions from a diffusion model with between-trial variability in T_{er}, a model with between-condition variability in T_{er} and a model with both forms of variability are shown by lines connected with a plus symbol (+), a cross (x), and a triangle, respectively.

The models with between-condition variability in T_{er} both provide a good

account of the word frequency effect, while the model with only within-condition variability in T_{er} still systematically fails to capture the effect. The lines connected by plus signs (+) are the predictions of the diffusion model with only between-trial (within-condition) variability in T_{er} (i.e. the same as the model used in Figure 2 and Ratcliff et al., 2004). Note the systematic under-prediction of the .1 quantile in all experiments, and the over-prediction of the .9 quantile in Experiments 3 and 5. The predictions of the models with between-condition variability in T_{er} or both forms of variability in T_{er} (representing in Figure 3 by lines joined by crosses and triangles, respectively) provide a much better account of the word frequency effect. Indeed, the two models produce an almost identical account of the word frequency effect in Experiments 4 and 5. In these experiments both models provide an excellent account of the difference between RTs from high and low frequency conditions at all quantiles except for the .9 quantile in Experiment 5. In Experiment 3 the model with both types of variability provides and excellent account of all but the .9 quantile, whereas the two other models also provide a less accurate account at three of the four remaining quantiles. Though we do not show it here due to space restrictions, a plot like Figure 3, but comparing high and very low frequency words from Experiment 5, showed the same pattern of results (once again this plot may be obtained by emailing the authors).

Discussion

We were prompted to fit a diffusion model which allowed mean non-decision time (T_{er}) to vary as a function of word frequency because of a) results from previous analyses of RT distribution using the Ex-Gaussian distribution, b) systematic misfit of the word frequency effect by a diffusion model which allows only drift rate to vary between frequency conditions, and c) the fact that a shift is plausible according as most

reading models, which assume that word frequency affects the time taken for lexical access. A diffusion model with both between-condition and between-trial variability provided a better fit to the data, even after accounting for this models increased parametric complexity. In particular, the model with both forms of variability provided an improved account of the word frequency effect compared to Ratcliff et al.'s (2004) original model with only between-trial variability in T_{er}, as it did not systematically under-predict the shift in the RT distribution between high and low frequency words.

A diffusion model with between-condition variability in T_{er}, but without between-trial variability in T_{er}, was also able to account for the shift effect. However, in terms of overall fit, this model did worse in two of three experiments than the Ratcliff et al. (2004) original model. A diffusion model with no variability in T_{er} either between-conditions or between-trials fit had a poor overall fit and account of the word frequency effect. These results together suggest that the addition of between-condition variability in T_{er} greatly improves the account of the shift in RT distribution due to changes in word frequency (see also Ratcliff & Tuerlinckx, 2002).

Even the diffusion model with both forms of variability in T_{er} still over-predicted the slowest differences between high and low frequency words in two of the three experiments we examined. This suggests that our current account of the word frequency effect and the LDT may not be complete. Indeed, given the intricacies of the lexicon, an even more complex model of the effects of frequency on non-decision time seems quite plausible and may account for these failings. However, it has been argued that the .9 quantile estimate is much more variable than the other quantile estimates, and most subject to the influence of slow outlier responses, so this misfit is not necessarily indicative of a failed model. An alternative possibility is raised by Donkin, Brown and

Heathcote's (submitted) recent demonstration that the moment-to-moment variability parameter has been, without justification, over-constrained in all previous applications of the diffusion model. When we let this parameter vary across frequency conditions BIC improved and excellent fits were obtained to all quantiles of the word frequency effect, and all other aspects of the data. However, due to space restrictions, details concerning these fits will be reported elsewhere.

Acknowledgments

We acknowledge support from an ARC Discovery project grant to Andrews and Heathcote.

References

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Human Perception and Performance, 27,* 514-544.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340–357.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *Journal of Experimental Psychology: General, 128,* 32-55.

Donkin, C., Brown, S., & Heathcoate, A. (submitted). The over-constraint of response time models. *Psychonomic Bulletin & Review.*

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology*: *General*, *136*, 389-413

Heathcote, A., Brown, S.D. & Mewhort, D.J.K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin and Review, 9*, 394-401

Luce, R. D. (1986) Response times: Their role in inferring elementary mental organization. NY: Oxford University Press.

Plourde, C. E., & Besner, D. (1997). On the locus of the word frequency effect in visual word recognition. *Canadian Journal of Experimental Psychology, 51*, 181-194.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 88,* 552-572.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*, 159-182.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers, *Brain and Cognition, 55*, 374-382.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438-481.

Voss, A., & Voss, J. (2008). A Fast Numerical Algorithm for the Estimation of Diffusion-Model Parameters. *Journal of Mathematical Psychology, 52*, 1-9

Wagenmakers, E-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140-159.

Yap, M.J., Balota, D.A., Cortese, M.J. & Watson, J.M. (2006). Single- versus dual-process models of lexical decision performance: Insights from response time distributional analysis, *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1324-1344.

The Over-Constraint of Response Time Models:

Rethinking the Scaling Problem

Chris Donkin, Scott D. Brown & Andrew Heathcote

The University of Newcastle,
Newcastle, Australia

Abstract

Theories of choice response time provide insight into the psychological underpinnings of simple decisions. Evidence accumulation (or sequential sampling) models are the most successful theories of choice response time. These models all have the same 'scaling' property– that a subset of their parameters can be multiplied by the same amount without changing their predictions. This property means that a single parameter must be fixed to allow estimation of the remaining parameters. We show that the traditional solution to this problem has over-constrained these models, unnecessarily restricting their ability to account for data and making implicit, and therefore unexamined, psychological assumptions. We show that versions of these models which address the scaling problem in a minimal way can provide a better description of data than their over-constrained counterparts, even when increased model complexity is taken into account.

Many psychological experiments involve a choice between two alternatives. Despite their apparent simplicity, there are many complicated empirical regularities associated with the speed and accuracy of such choices. Response time (RT) distributions take on characteristic shapes, which differ systematically depending on whether the associated response is correct or incorrect and depending on any number of experimental manipulations of stimulus properties or of instructions to the participants. A range of theories have been proposed to account for both choice probability and response time when making simple decisions (for reviews see Luce, 1986; Ratcliff & Smith, 2004). Over the past 40 years, evidence accumulation (or "sequential sampling") models have dominated the debate about the cognitive processes underlying simple decisions (e.g., Busemeyer & Townsend, 1993; Ratcliff, 1978, Ratcliff & Smith, 2004; Smith, 1995; Stone, 1960; Usher & McClelland, 2001; Van Zandt, Colonius, & Proctor, 2000).

More recently, evidence accumulation models have been applied more widely, for example, as general tools to measure cognition in the manner of psychometrics (Schmiedek, Oberauer, Wilhelm, Suss & Wittmann, 2007; Wagenmakers, van der Maas & Grasman, 2007; Vandekerckhove, Tuerlinckx, & Lee; submitted), and as models for the neurophysiology of simple decisions (e.g.: Forstmann, Dutilh, Brown, Neumann, von Cramon, Ridderinkhof, & Wagenmakers, 2008; Ho, Brown, & Serences, 2009; Smith & Ratcliff, 2004). In light of this growing influence it is especially important that users of these models are not misled by implicit - and hence unexamined - assumptions.

Evidence accumulation models all share a basic framework wherein, when making a decision, people repeatedly sample evidence from the stimulus. This evidence is accumulated until a threshold amount is reached, which triggers a decision response.

These models naturally predict the response made (depending on which response has accumulated the most evidence) and the latency of the response (depending on how long the evidence took to accumulate). We illustrate these models using the example of a lexical decision task, where a participant must decide whether a string of letters is a valid word (e.g. "DOG") or not (e.g. "DXG"). The participant samples information from the stimulus repeatedly, and finds some evidence that suggests that the stimulus is a word and other evidence to suggest that the stimulus is not a word. The participant accrues this information, waiting until there is enough evidence for one of the two options before responding. Their choice corresponds to the response with the most evidence, and the time taken for this evidence to be accumulated is the response latency.

Over the past four or five decades, dozens of evidence accumulation models have been proposed, and all of them share a mathematical "scaling property": one can multiply a subset of their parameters by an arbitrary amount, without changing any of the model's predictions. To avoid complications arising from the scaling property, just one parameter of the model must be constrained arbitrarily. We show that the conventional approaches – which have been universally applied to solve the scaling problem – have actually over-constrained the models by fixing more than one parameter. This over-constraint has been largely unrecognized by the field, and so it is equivalent to making a tacit, untested, psychological assumption. Further, we show that this tacit assumption can sometimes have important consequences: when the scaling problem is solved in a minimal way, the models can sometimes provide better account for data.

Overview of the models

There are two major classes of evidence accumulation models: single

accumulator models (Busemeyer & Townsend, 1993; Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002; Smith, 1995; Stone, 1960) and models that have one accumulator for each possible response (Brown & Heathcote, 2005, 2008; Smith & Van Zandt, 2000; Smith & Vickers, 1988; Townsend & Ashby, 1983; Usher & McClelland, 2001; Van Zandt, Colonius & Proctor, 2000; Vickers, 1970). The customary method for solving the scaling problem differs between the two classes of models, even though the principle is the same. To simplify our discussion we choose a specific model from each class: the single accumulator diffusion model (Ratcliff & Tuerlinckx, 2002) and the multiple accumulator linear ballistic model (LBA; Brown & Heathcote, 2008). We have chosen these two models largely for convenience, as both have easy-to-implement computer code that is freely available (see Donkin, Averell, Brown & Heathcote, in press, and Voss & Voss, 2007). The general point that we make, however, applies to all evidence accumulation models.

Continuing the lexical decision example, the diffusion model assumes that participants sample evidence from the stimulus continuously, and this evidence stream updates an evidence total, say x, illustrated as a function of time by the irregular line in Figure 1. The accumulator begins the decision process in some intermediate state, say $x=z$. Evidence that favors the response "word" decreases the value of x, and evidence that favors the other response ("non-word") increases the value of x. The evidence accumulation process continues until sufficient evidence favors one response over the other, causing the total (x) to reach one of its two boundaries (the horizontal lines at $x=0$ and $x=a$ in Figure 1). The choice made by the model depends on which boundary is reached (a for a "non-word" response or 0 for a "word" response) and response time equals the accumulation time plus a constant, T_{er}, that represents the time taken by non-

decision processes, such as encoding the stimulus and producing the response.

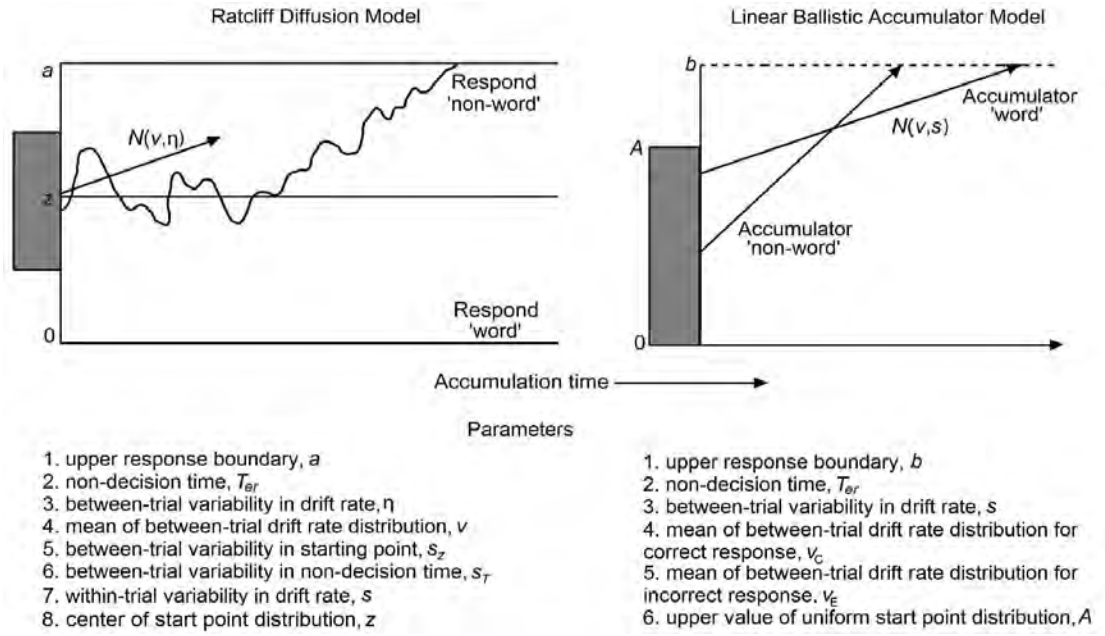

Figure 1 Graphical representations of a single decision made by the diffusion model and the LBA model.

Depending on the stimulus, evidence tends to accumulate more towards one boundary or another, and the average rate of this accumulation is called the "drift rate", which we will label *v*. The evidence accumulation process also varies randomly from moment-to-moment during the accumulation process, and the amount of this variability is another parameter of the model, *s*. Recent applications of the diffusion model include three extra parameters, but these are not important for our purposes, so we delay their introduction until later. When experimental conditions differ only in stimulus characteristics that vary randomly from trial to trial, all parameters except the drift rate are conventionally assumed constant over conditions (Ratcliff, Gomez & McKoon, 2004).

The LBA is a multiple accumulator model, meaning that it assigns a separate evidence accumulator to each possible response: for example, in lexical decision, one accumulator gathers evidence in favor of the response "word" and the other gathers

evidence for the "non-word" response, as illustrated in Figure 1. The activity level in each accumulator begins at a value that is randomly sampled (separately for each accumulator) from the interval [0,A]. Evidence accumulation is noiseless ("ballistic") and linear with a slope that we again call the "drift rate", v. When the evidence accumulated for either response reaches a threshold, b, a response is made. Like the diffusion model, the LBA assumes that non-decision processing takes fixed time, T_{er}. The drift rates are assumed to vary from trial to trial according to normal distributions with means v_W for the "word" accumulator and v_{NW} for the "non word" accumulator, and a common standard deviation, s.

Scaling Properties

Consider just one of the evidence accumulators from the LBA. The accumulator begins a trial with some activity, say x_0, between 0 and A, and increases at a rate of v units per second (v is the drift rate for this accumulator). Evidence accumulation ends when the threshold b is reached, which will take $(b-x_0)/v$ seconds. If all of these model parameters were multiplied by a common amount the predicted response time would remain unchanged; for example, if the parameters were doubled then the evidence accumulation process would travel twice as quickly, but would also have to travel twice as far. This scaling property is true of all evidence accumulation models – all parameters that affect evidence accumulation can be multiplied by any fixed amount without altering the model's predictions.

The scaling property makes it impossible to estimate unique model parameters from data unless the value of one parameter is fixed arbitrarily. In single accumulator models, including Ratcliff's diffusion, this has always been done by fixing the variability of the diffusion process at either s=0.1 or s=1 (e.g., Ratcliff, 1978; Ratcliff &

Rouder, 1998; Ratcliff & Tuerlincx, 2002; Smith & Ratcliff, 2004; Van Zandt, Colonius & Proctor, 2000; Voss, Rothermund, & Voss, 2004). In fact, the diffusion coefficient is usually referred to as *the* "scaling parameter" of the model, even though any other parameter could equally well be fixed to avoid scaling problems.

For the LBA and other multiple accumulator models, problems due to the scaling property have been avoided by fixing the sum of the drift rates for the two accumulators to a constant (Brown & Heathcote, 2005, 2008; Forstmann, et al., 2008; Ho, et al., 2009; Ratcliff & Smith, 2004; Smith & Van Zandt, 2000; Townsend & Ashby, 1983; Usher & McClelland, 2001). For example, in the word vs. non word model above, one might fix $v_W+v_{NW}=1$. Mathematically speaking, any other parameter constraint would do just as well to solve the scaling problem, for example the boundary separation, or one of the drift rates, could be fixed. It is simply a matter of convention that the field has settled on the sum-of-drift-rates constraint for multiple accumulator models, and the diffusion noise constraint for single accumulator models.

The scaling properties just described are simple and well-understood. However, the situation is complicated in practice because evidence accumulation models are almost never used to analyze just one experimental condition in isolation. Instead, data are collected from multiple experimental conditions, which are analyzed together[2]. This allows some parameters to be fixed across experimental conditions, depending on what psychological assumptions one is willing to make. For example, when experimental conditions differing only in stimulus properties are randomly ordered from trial to trial, parameters that are assumed to be under the strategic control of the participant (such as boundary separation) are often fixed. This is justified by the notion that such parameters

2 Applications of Wagenmakers et al.'s (2007) EZ estimation technique, such as by Schmiedek et al. (2007), are an exception.

take time and effort to change, and hence that such changes are unlikely to occur between stimulus onset and the response (Ratcliff, 1978).

When parameters are fixed across conditions, changes in parameters for one condition naturally alter the predictions for other conditions. So, when multiple conditions are analyzed simultaneously, the scaling properties of the models can be constrained by fixing a single parameter in *only one* of the conditions. For example, suppose one conducts an experiment with five different levels of difficulty defined by different stimuli. In this design, scaling problems are avoided if a single parameter is constrained in only one of the five conditions. If Ratcliff's diffusion model is used, one can set s=1 in just one of the five conditions, or in the LBA, one could set the sum of the drift rates equal to one in just one of the five conditions.

However, these are not the constraints that have been used in practice. In all of the studies that we have reviewed (including our own), researchers have constrained "scaling parameters" independently in *all* conditions. To continue our example, they have either fixed s for all five conditions (in single accumulator models, like the diffusion) or fixed the sum of the drift rates for all five conditions (in multiple accumulator models). Avoiding estimation problems due to scaling properties requires fixing just one parameter value, but researchers have always fixed one parameter value *per experimental condition*.

These so-called 'scaling parameters' (i.e. s or the sum of drift rates) have also come to be treated quite differently from the other model parameters, as "fixed, not free" (p. 440, Ratcliff & Tuerlinckx, 2002). Mathematically, dealing with scaling parameters entails two independent decisions – firstly, one model parameter is arbitrarily selected for constraint, and secondly that parameter can be (and always has

been) held to its fixed value across *all* experimental conditions. The first decision has no theoretical consequence, because constraints on different types of parameters are mathematically equivalent. The second decision can have theoretical consequence. Suppose, for example, that one had decided to fix the boundary separation parameter to solve the scaling problem. Keeping this parameter fixed across conditions that differ in instructions that emphasize the speed or accuracy of responses is clearly inappropriate. In practice, however, it seems that the second decision has automatically followed the first; that choosing a scaling parameter has pre-determined its role as fixed across experimental conditions.

Our central message is that the parameter type chosen as a scaling parameter *can* vary across conditions while still satisfying the scaling constraint property. As we will now demonstrate, this can be important since using minimal constraints changes two things: the predictions and the psychological interpretations of the models. By re-analyzing previously published data we show that allowing parameters previously treated in the conventional scaling manner to vary across conditions can have a substantial, and sometimes useful, effect. We show that when a minimal rather than conventional scaling solution is used, the resulting improvement in fit is sufficient to justify the extra parametric freedom, even when a very strict complexity penalty is employed. We then briefly address the psychological implications of our proposal.

Re-analysis of Gould, Wolfgang and Smith's (2007) data

Gould, Wolfgang and Smith (2007) investigated the effect of cueing and localization in a stimulus detection task. We focus on one of their cueing conditions, the one providing the greatest challenge for evidence accumulation models. Gould et al. manipulated the difficulty of stimulus detection by varying contrast over five levels.

This resulted in ten different RT distributions – one for correct responses and one for incorrect responses from each of the five contrast levels. Figure 2 summarizes these ten distributions using quantile probability (QP) plots, with data (averaged across subjects) represented as filled circles connected by solid lines. QP plots have proven very important in discriminating between models of choice RT (Brown & Heathcote, 2008; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004). The x-axis measures response probability, and the y-axis shows the latencies associated with five quantiles of the RT distributions (10%, 30%, 50% - the median, 70% and 90%).

An example may make this clearer; consider just the data from the very easiest stimulus condition. Average response accuracy for this condition was 97.3%, so the quantile estimates from the distribution of correct responses are plotted as five filled circles vertically above x=.973, and the quantile estimates from the distribution of incorrect responses are plotted above x=.027 (i.e., 1-.973). The 10% quantile estimate for the correct RT distribution was 409msec (i.e., 10% of correct responses were faster than 409msec), so the first filled circle above x=.973 is at y=409. The procedure is repeated for the remaining quantile estimates for both the correct and incorrect distributions. The resulting plot allows one to assess how the RT distribution changes with response accuracy and between correct and incorrect responses. The solid lines in Figure 2 join each of the quantiles across contrast values for correct and error responses.

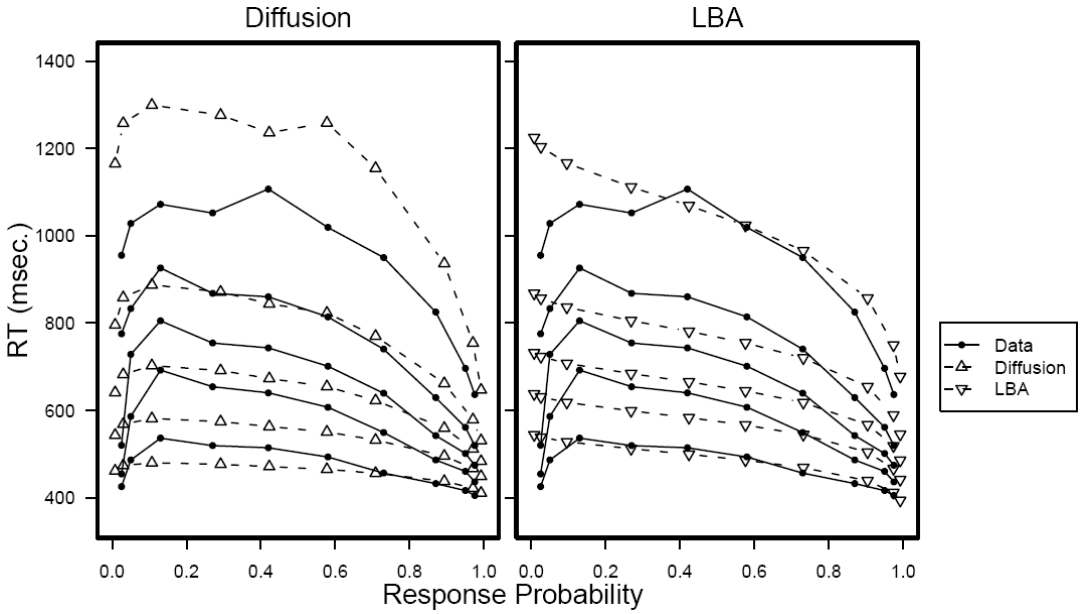

Figure 2 Quantile probability plots for Gould et al.'s (2007) data and fits (averaged across participants). The data are shown as filled circles joined by solid lines. In the left and right panels the dotted lines are the fits of the diffusion and LBA models, respectively.

As is typical, correct responses for difficult decisions were slower than easy decisions (as we move from the right to the center of the plots) at all five quantiles. Incorrect responses were slower still (on the left half of the plots) with the possible exception of errors in easiest condition (far left). What is unusual in these data is the amount of change in the fastest response times (the 10% quantile estimate). The 10% quantile estimate for correct responses was 87 ms slower in hardest condition relative to the easiest condition. Most previous studies have observed that the fastest decision times change by at most 30ms across the range of a QP plot (e.g., Ratcliff et al., 2004; Ratcliff & Smith, 2004).

Model fits – Conventional scaling

Figure 2 shows fits of the diffusion and LBA models when employing the over-constrained conventional solution to the scaling problem (i.e., fixing one parameter across all experimental conditions). To fit both models we followed the usual

convention of assuming that only drift rate varies between conditions, and that all other parameters were equal across all conditions. A constant non-decision time (T_{er}) assumes encoding and response production time does not vary across conditions. Constancy of the strategic parameters is justified because stimulus conditions varied randomly from trial to trial. We also assumed that responding was unbiased; for the diffusion model this means evidence accumulation begins half way between the bounds ($z=a/2$), and that response boundaries are equal in the LBA.

The simple diffusion model described earlier would fit these data with five free parameters. However, in practice the diffusion model uses another three free parameter types related to trial-to-trial variation: the starting point of the evidence accumulation process varies according to a uniform distribution on [z-s_z, z+s_z], the drift rate varies according to a normal distribution $N(v,\eta)$, and non-decision time (T_{er}) varies according to a uniform distribution on [T_{er}-s_T , T_{er}+s_T]. These additions make a total of ten free parameters (T_{er}, s_T, a, η, s_z, v_1, v_2, v_3, v_4, v_5), as they are assumed the same across conditions. The LBA fits also used ten free parameters for each participant (T_{er}, s_T, A, b, s, v_1, v_2, v_3, v_4, v_5). This is one more parameter than has previously been used in applications of the LBA, as non-decision time, T_{er}, is usually assumed fixed (i.e., s_T=0). We allowed non-decision time to vary here for equivalence with the diffusion model (although s_T was estimated at about zero for the LBA). The sum of average correct and incorrect drift rates in the LBA was set at 1 for all five stimulus contrast levels. The diffusion model was constrained by having the diffusion coefficient fixed at s=1 across all five conditions.

Parameters were estimated using the method of quantile maximum probabilities (Heathcote & Brown, 2004). Model predictions were evaluated using the LBA code

provided by Donkin et al. (in press) and the diffusion model code provided by Voss and Voss (2007). The Bayesian information criterion (BIC) was calculated at the best-fitting parameters for each participant: the BIC statistic for N observations grouped into bins is:

$$BIC = -2(\sum_i Np_i \ln(\pi_i)) + M \ln(N)$$

where p_i is the proportion of observations in the i^{th} bin, and π_i is the proportion of observations in the i^{th} bin as predicted by the model. M is the number of parameters of the model used to generate predictions. The BIC is composed of two parts, the first is a measure of misfit, and a second part, $M \ln(N)$, penalizes a model for its complexity as indicated by the number of estimated parameters. When comparing two models, the model with the smaller BIC is considered to have provided a better fit to the data, after complexity has been taken into account. We use BIC because it imposes a larger complexity penalty than alternatives such as the Akaike Information Criterion (AIC), and so it provides a more stringent test of whether the models benefit from the extra parameter variation allowed by imposing minimal constraints to solve the scaling problem.

Table 1 Parameter estimates and BIC from fits to average data from Gould et al.'s (2007) cued+FID condition (D = diffusion model, L= LBA model)

	$T_{er}{}^a$	$s_I{}^a$	a	η^b	s_z	$v_1{}^b$	$v_2{}^b$	$v_3{}^b$	$v_4{}^b$	$v_5{}^b$	s_2	s_3	s_4	s_5	BIC
D	.352	.083	1.46	1.76	.000	5.17	4.01	2.61	1.12	0.37	-	-	-	-	39406
	.364	.099	1.26	1.59	.046	4.78	3.68	2.36	1.12	.378	1.01	.923	.843	.788	39305
	$T_{er}{}^a$	$s_I{}^a$	b	s^b	A	$v_1{}^b$	$v_2{}^b$	$v_3{}^b$	$v_4{}^b$	$v_5{}^b$	$\sum v_2{}^b$	$\sum v_3{}^b$	$\sum v_4{}^b$	$\sum v_5{}^b$	BIC
L	.144	.000	.356	.276	.047	.974	.881	.757	.621	.538	-	-	-	-	39403
	.168	.002	.287	.219	.082	.797	.720	.609	.475	.378	.923	.800	.692	.648	39166

Note: [a] indicates parameters whose units are in 'seconds'. [b] have units 'per second', while other parameters have arbitrary units.

Parameter estimates and BIC values are shown in Table 1 (we focus on averaged data for brevity). As the right panel of Figure 2 shows, both models provided poor

accounts of the data – the diffusion under predicts the shift in RT distribution across conditions, while the LBA fails to capture the faster errors that occur in easy conditions. The conclusion we draw is that standard applications of both models fail to provide convincing accounts of these data.

Model Fits – Minimally Constrained

For the minimally constrained version of Ratcliff's diffusion model, we fixed the diffusion coefficient in the highest contrast condition at s_1=1 and freely estimated diffusion coefficients for the other four contrast conditions (s_2, s_3, s_4, s_5). For the minimally constrained version of the LBA, we fixed the sum of correct and error drift rates to be one in the easiest condition, and estimated this sum in the other four conditions (i.e. Σv_2, Σv_3, Σv_4, Σv_5). Table 1 reports the estimated parameters and the BIC values for the minimally constrained fits, shown in Figure 3. The quality of fit was greatly improved, with both models providing a much better account of the data than the conventionally constrained versions.

BIC values were better in the minimally constrained versions of both models, suggesting that the improvement in fit outweighed the cost of adding four additional parameters. Using methods outlined by Wagenmakers and Farrell (2004), BIC values can be converted to model selection probabilities (see Raftery, 1995, for a discussion of conventions for interpreting such probabilities). The BIC improvement provided very strong evidence (p>.99, Raftery, 1995) favoring both minimally constrained models over their conventionally constrained counterparts. The improvement in the diffusion model seems to have come from predicting a larger shift in RT distribution across conditions and no longer predicting such extreme skewness for difficult decisions. The minimally constrained LBA was better able to accommodate the fast errors. As before,

estimated drift rates for both models decrease in a sensible manner with decreasing stimulus contrast (Table 1). For both models, the scaling parameter also decreased with decreasing stimulus contrast (the sum of the drift rates for the correct and incorrect response accumulators in the LBA, and the diffusion variability coefficient in the diffusion model).

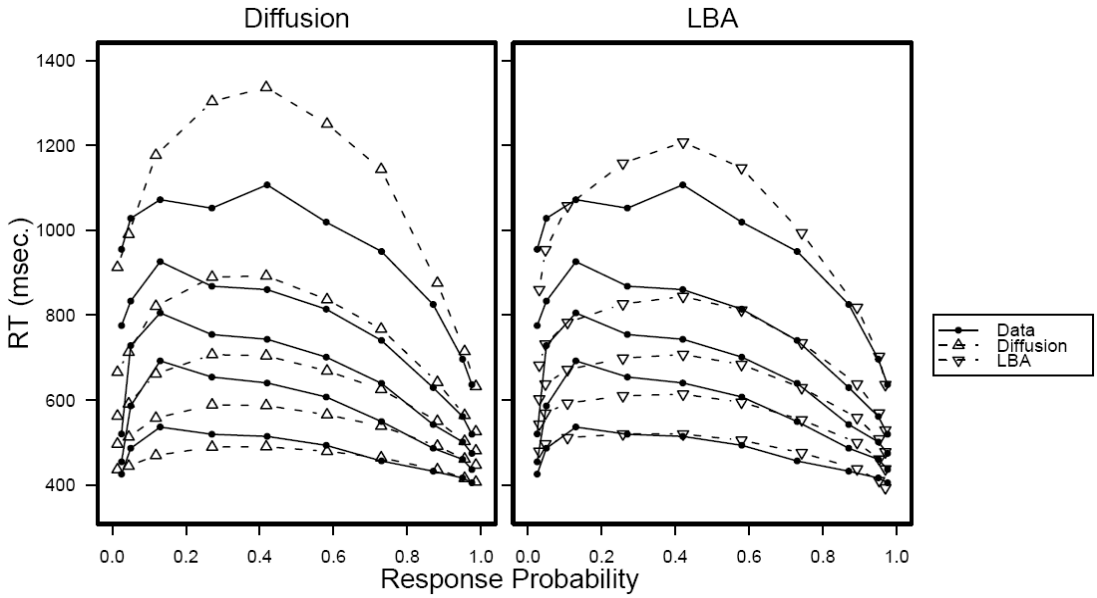

Figure 3 Quantile probability plots and fits averaged over participants for the minimally constrained versions of the diffusion model (left panel), the LBA (right panel) for Gould et al.'s (2007) data. The data are shown as filled circles and solid lines.

Discussion

All evidence accumulation models require a "scaling property" to be fixed, before parameters can be estimated. To this end, researchers must choose a parameter to constrain, but this choice is logically independent of the subsequent decision of whether to further constrain that parameter across experimental conditions. In practice, however, these two decisions have never been separated – the parameter chosen to satisfy the scaling property has always also been constrained across experimental conditions. This is a nontrivial assumption, because the scaling parameters of the models could plausibly be driven by stimulus characteristics which often differ between conditions. A re-

analysis of one such case, from Gould et al. (2007), showed that separating these two decisions was justified by improved fits to data for both the LBA and diffusion models, even allowing for a very stringent model complexity penalty.

For multiple accumulator models, such as the LBA, the assumption of a constant sum for correct and incorrect drift rates across conditions implies that increasing the stimulus evidence in favor of one response will equally increase the evidence against the other response. However, it seems reasonable that some stimulus manipulations could decrease the evidence available for *both* responses. Contrast is plausibly one such manipulation: as contrast decreases there may be less evidence supporting either response, and our parameter estimates from Gould et al.'s (2007) data were consistent with this interpretation. For single accumulator models, such as Ratcliff's diffusion, the conventional constraints imply that the variability in evidence accumulation is independent of the mean rate of accumulation. This assumption might be reasonable if, for example, the decision signal arises from one set of processes whereas *all* decision noise arises from an independent set of processes. However, our parameter estimates suggest that the diffusion model may better account for the effects of decreasing stimulus contrast by assuming some dependence between decision signal and decision noise.

Although we have focussed on the diffusion and LBA models, the same arguments apply to all evidence accumulation models. Other multiple accumulator models have been similarly over-constrained, particularly the many variants of Usher and McClelland's (2001) leaky competing accumulator model, including the racing diffusion of Ratcliff, Cherian and Segraves (2003) and the ballistic accumulator (Brown & Heathcote, 2005). The Poisson counter models (Smith & Van Zandt, 2000; Townsend

& Ashby, 1983; Ratcliff & Smith, 2004; Van Zandt, Colonius & Proctor, 2000) have all been similarly over constrained by their own conventional solutions for the scaling problem.

General Discussion

The way in which the parameters of evidence accumulation models are constrained across conditions is based on careful argument and empirical evidence. For example, Ratcliff (1978) proposed that strategic parameters (e.g., boundary separation) should not differ among conditions whose order is randomized within blocks of trials. In contrast, parameters related to the quality of evidence provided by the stimulus (e.g., drift rate) should vary whenever stimulus properties change (see also: Ratcliff & Rouder, 1998; Voss et al., 2004). However, the "scaling parameters" of evidence accumulation models have always been fixed across all conditions, even though these parameters may most naturally be interpreted as ones influenced by stimulus properties. A review of the literature reveals neither careful argument nor empirical evidence to justify this extra constraint – it appears to have been a result of a misunderstanding scaling property (this is certainly true on our own part). Our results show that this over-constraint may not have always been benign; it can restrict the models' ability to account for data, and it makes implicit psychological assumptions.

It is possible that experts in the field may have been aware of this additional assumption being made when scaling parameters were fixed across conditions. For example, more than 25 years ago, Weatherburn (1978), Pike and Dalgleish (1982) and Weatherburn and Grayson (1982) discussed whether or not scaling parameters might vary in earlier instances of multiple accumulator models. This discussion, however, did not include actually trying out such models, and our literature review suggests that the

implications of their discussion have since gone unrecognized (we could find no citations of these papers in the past 14 years). Our aim is to ensure that the ever-expanding group of researchers who use response time models are aware of the implicit assumptions made when fixing the scaling parameter constant across conditions.

References

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112*, 117-128.

Brown, S.D., & Heathcote, A.J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153-178.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432– 459.

Donkin, C., Averell, L., Brown, S.D., & Heathcote, A. (in press). Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator model. *Behavior Research Methods*.

Forstmann, B.U., Dutilh, G., Brown, S.D., Neumann, J., von Cramon, D.Y., Ridderinkhof, K.R., & Wagenmakers, E.J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Science 105,* 17538-17542.

Gould, I. C., Wolfgang, B. J., & Smith, P. L. (2007). Spatial uncertainty explains exogenous and endogenous attentional cuing effects in visual signal detection. *Journal of Vision. 7*(13):4, 1-17.

Heathcote, A. & Brown, S.D. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin & Review, 11,* 577-578.

Ho, T.C., Brown, S.D. & Serences, J.T. (2009). Domain general mechanisms of perceptual decision making in human cortext. *Journal of Neuroscience.*

Luce, R. D. (1986). *Response times.* New York: Oxford University Press.

Pike, A.R., & Dalgleish, L. (1982). Latency-probability curves for sequential decision

models: A comment on Weatherburn. *Psychological Bulletin, 91,* 384-388.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59 –108.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two choice decisions. *Journal of Neurophysiology, 90*, 1392-1407.

Ratcliff, R., Gomez, P. & McKoon, P. (2004). A diffusion model account of the lexical decision task, *Psychological Review*, *111*, 159-182.

Ratcliff, R., & Rouder, J. N. (1998). Modelling response times for two choice decisions. *Psychological Science, 9,* 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating the parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481.

Schmiedek, F., Oberauer, K., Wilhelm, O., Suss, H.M. & Wittmann, W.W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence, *Journal of Experimental Psychology: General*, *136*, 414-429.

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review, 102,* 567–591.

Smith, P.L., & Ratcliff, R. (2004). Psychology and Neurobiology of Simple Decisions. *Trends in Neurosciences, 27*, 161-168.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology, 53,* 293–315.

Smith, P.L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology, 32,* 135-168.

Stone M. (1960). Models for choice-reaction time. *Psychometrika, 25*, 251–260.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes.* Cambridge, England: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108*, 550 –592.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7,* 208 –256.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (submitted). Hierarchical diffusion models for two-choice response times. *Psychological Methods*.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13,* 37–58.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206-1220.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavioral Research Methods, 39*, 767–775.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192-196.

Wagenmakers, E.-J., van der Maas, H.L.J., & Grasman, R.P.P.P. (2007). An EZ-

diffusion model for response time and accuracy. *Psychonomic Bulletin & Review 4, 3-22.*

Weatherburn, D. (1978). Latency-Probability functions as bases for evaluating competing accounts of the sensory decision process. *Psychological Bulletin, 85,* 1344-1347.

Weatherburn, D., & Grayson, D. (1982). Latency-Probability functions: A reply to Pike and Dalgleish. *Psychological Bulletin, 91,* 389-292.

An Integrated Model of Choices and Response Times in Absolute Identification

Scott D. Brown[1], A.A.J. Marley[2], Christopher Donkin[1], & Andrew Heathcote[1]

1. University of Newcastle
Newcastle, Australia

2. University of Victoria
British Columbia, Canada

Abstract

Recent theoretical developments in the field of absolute identification have stressed differences between relative and absolute processes: that is, whether stimulus magnitudes are judged relative to a shorter-term context provided by recently presented stimuli or a longer-term context provided by the entire set of stimuli. We develop a model (SAMBA) that integrates shorter- and longer-term memory processes and accounts for both the choices made, and the associated response time distributions, including sequential effects in each. The model's predictions arise as a consequence of its architecture and require estimation of only a few parameters with values that are consistent across numerous data sets. We show that SAMBA provides a quantitative account of benchmark choice phenomena in classical absolute identification experiments and in contemporary data involving both choice and response time.

Keywords: absolute identification; absolute identification experiments; absolute identification models; response time; response time distributions; sequential effects.

Performance in absolute identification tasks has fascinated researchers for over 50 years (e.g., Garner, 1953; Miller, 1956; Pollack, 1952, 1953). Research in the past 35 years has emphasized both data and formal theories (e.g., Braida & Durlach, 1972; Durlach & Braida, 1969; Laming, 1984; Lockhead, 2004; Luce, Nosofsky, Green & Smith, 1982; Marley & Cook, 1984; Petrov & Anderson, 2005; Stewart, Brown & Chater, 2005; Treisman & Williams, 1984) and, most recently, has been concerned with both the choices made and the time it takes to make them (Kent & Lamberts, 2005; Lacouture & Marley, 1991, 1995, 2004). As Shiffrin and Nosofsky (1994) stated in an article reassessing the significance of Miller's classic paper, "absolute identification has captured the imagination…not only because the empirical results are so startling but also because [they] provide perplexing problems for classic psychophysical models". Luce (1986, Chapter 10) gives an excellent summary of data and theory to that date, and Lockhead summarizes data and theory most relevant to relative interpretations of absolute identification, where the relativity is with respect to stimuli and responses from previous trials. Stewart et al. and Petrov and Anderson provide comprehensive reviews of choice data and the related theory, with emphasis on theoretical approaches over the past 20 years.

A typical absolute identification task requires a participant to identify, on each trial, which stimulus has been presented from a relatively small pre-specified set. In general, people are unable to accurately identify more than about 8-10 stimuli that vary on a single physical dimension. For example, the stimuli might be a set of 10 lines varying only in length, with the shortest line labeled "#1" and the longest "#10". A participant previews the entire labeled set and is then shown the lines one at time, over numerous trials, and asked to identify the presented line with the appropriate response

label. Typically, a participant in this task is unable to achieve an overall accuracy above about 80%, which is surprising given that the stimuli are chosen such that *comparative* judgments of any pair of them are completely accurate (i.e., judging whether one stimulus is smaller or greater than another stimulus presented in rapid succession).

With such an extensive history, the study of absolute identification is a mature field with many well-established benchmark behavioral phenomena that describe how choices and response times are affected by stimulus manipulations and by the history of stimuli and responses. We broadly separate these phenomena into global and local effects:

1. Global effects: Stimulus range and set size. For a fixed set size (N) of stimulus-response pairs, performance – measured by the amount of information transmitted – increases quickly to an asymptotic level of 2-3 bits as the range of stimuli on the physical dimension increases (Braida & Durlach, 1972). Similarly, as set size increases the amount of information transmitted increases rapidly at first but then asymptotes at 2-3 bits (e.g., Garner, 1953; Pollack, 1952, 1953). Responses to the largest and smallest stimuli in a set are faster and more accurate than responses to the middle stimuli – the bow effect. Both accuracy and response times worsen for any given stimulus as other stimuli are introduced to the set (Kent & Lamberts, 2005; Lacouture & Marley, 1995). As shown by Luce et al.'s (1982) *d'* analysis[3], bow effects in accuracy are partly due to bow effects in sensitivity, and partly to response bias produced by the constraints on available responses for stimuli near the ends of the range. Ward (1987) showed that scaling methods that require either relative or absolute judgments reveal profound effects of stimulus-response mappings over days.

2. Local effects: Sequential effect on accuracy and errors. Previous stimuli and

3 Throughout this paper we use Luce et al.'s (1982) method of calculating *d*' to quantify sensitivity.

responses can affect the response to the current stimulus (e.g., Lacouture, 1997; Ward & Lockhead, 1970, 1971). When an incorrect response is given, it tends to be toward, rather than away from, the stimulus from the previous trial (*assimilation*). The opposite pattern occurs for longer lags (*contrast*): errors tend to be away from, rather than towards, stimuli from two or more trials previously[4]. Accuracy is improved when stimuli are constrained to be similar on successive trials (e.g., Luce et al., 1982). In particular, the difference in magnitudes between the stimuli presented on the current and previous trials influences response accuracy (e.g., Petrov & Anderson, 2005; Rouder, Morey, Cowan & Pfaltz, 2004; Stewart at al., 2005).

Previous theoretical accounts of absolute identification have attempted to show that some or all of these phenomena could be accounted for using only relative processes or only absolute processes. A model that uses only absolute processes is one where decisions are made about stimulus magnitudes based on comparisons with some longer-term referents (e.g., the context-coding component of Braida, Lim, Berliner, Durlach, Rabinowitz & Purks, 1984; Lacouture & Marley, 1995, 2004) or a longer-term frame of reference (e.g., Marley & Cook, 1984). On the other hand, a model that uses only relative processes (e.g., Laming, 1984; Lockhead, 2004; Stewart et al., 2005) posits that decisions are made using only comparisons with recent stimuli and responses. Range and set size effects have most often been attributed to absolute processes (e.g., Braida et al. and Marley & Cook). Sequential effects, particularly assimilation and contrast, have been frequently explained by shorter-term relative judgment processes. Other than the model (SAMBA) developed in this paper, there is no model, either relative or absolute, that accounts for all of the global and local benchmark phenomena

4 Since stimuli and correct responses are correlated, care must taken in interpreting these as solely stimulus (or response) effects.

described above.

Stewart et al. (2005) present a league table (their Table 2) comparing absolute identification models on their ability to account for nine benchmark phenomena, and associated classical data sets, under three broad headings: limited information transmission, bow (set size) effects and sequential effects. The most comprehensive relative theory, Stewart et al.'s Relative Judgment Model (RJM[5]), performs well on choice-related phenomena, but does not address response times (RTs). The most comprehensive absolute theory, Lacouture and Marley's mapping model (1995, 2004), performs well on global phenomena for both choice and RT, but does not address sequential effects. Stewart et al. count RT as one benchmark phenomenon, even though a wide range of benchmark RT phenomena have been identified (Lacouture, 1997; Lacouture & Marley, 1995, 2004), some paralleling those found in choices, including set size effects, stimulus magnitude effects, and sequential effects, and some specific to RT, such as distribution shapes for correct and error responses for each stimulus.

We propose a model of both choice and RT in absolute identification in which sequential effects, including assimilation and contrast, result from short-term memory effects in an absolute judgment process. Our model further develops key concepts from several previous models. Reflecting this cumulative history, the acronym for the model, SAMBA, highlights the three core elements: Selective Attention (Marley & Cook, 1984), Mapping (Lacouture & Marley, 1995, 2004), and Ballistic Accumulation (Brown & Heathcote, 2005). Stewart et al. (2005), in their league table, show that the separate components of SAMBA are, by themselves, inadequate. This gives motivation for SAMBA, which integrates and extends these components, to account for assimilation,

5 The term "relative judgment model" was earlier used by DeCarlo and Cross (1990) for their model of magnitude scaling - their Equation (16).

contrast, asymmetries in bow effects and local judgment effects. SAMBA provides a unified account of choice phenomena as well as the associated RT phenomena, as we demonstrate by fitting SAMBA to Lacouture's (1997) full range of choice and RT data, which is averaged over participants, and to Kent and Lamberts' (2005) and Lacouture and Marley's (2004) individual subject data. While SAMBA is the first model of absolute identification to provide a comprehensive account of response times and response choices that includes sequential effects, of course we do not claim that other such models cannot be developed. For example, both Lacouture and Marley's (2004) and Kent and Lamberts' (2005) models provide a good account of response times and most choice phenomena, but do not explain sequential effects, but these models may be further developed to cover such effects. Similarly, models that cover response choices but not response times (such as Stewart et al.'s RJM, or Petrov and Anderson's ANCHOR), may be developed further to include a response time mechanism. We note, however, that an account of response time added to a choice model is not guaranteed success – see, for example, Karpiuk, Lacouture and Marley (1997).

Recently, Brown, Marley and Lacouture (2007) highlighted the theoretical importance of sequential effects in accuracy (see also Petrov & Anderson, 2005; Rouder et al., 2004; Stewart et al., 2005), thus going beyond just assimilation and contrast, which describe sequential effects in errors. Brown et al. focused on Rouder et al.'s analysis of accuracy as a function of the difference between the current and previous stimulus. Analyzing data from Lacouture's (1997) line length task, they observed improved accuracy for a stimulus similar to the one before, and also for a stimulus very different from the one before. Subsequently, Stewart (2007) noted the same pattern in three other data sets (Kent & Lamberts, 2005; Neath & Brown, 2006; and Stewart et al., 2005).

Brown et al. attributed higher accuracy when successive stimuli are similar to a comparison of the present stimulus with the prior stimulus, and higher accuracy when successive stimuli are very different to comparison with an end stimulus. Stewart explained the pattern by modifying the RJM to include a memory for the stimulus *two* trials back that on some trials is used instead of the memory for the stimulus one trial back as the basis for relative judgment. Other experiments by DeCarlo and Cross (1990) and DeCarlo (1994) demonstrate that instructions significantly impact whether magnitude judgments are made relative to short- or long-term referent stimuli and responses. These findings present problems for models that rely solely on absolute or solely on relative mechanisms, requiring suitable extensions of the relative approach, as shown by Stewart, and a suitable extension of the absolute approach, as we will show for SAMBA.

SAMBA is an integrative model not only because it accounts for choice and RT data, but also because it includes relative as well as absolute processes. We show that SAMBA is able to model the complex effects shown in Brown et al.'s (2007) analysis by replacing one of the end anchors used by Marley and Cook's (1984) selective attention process with the estimated magnitude of the previous stimulus. However, this relative judgment process is not required for SAMBA's account of classic absolute identification phenomena, including assimilation and contrast, so we omit it in our fits to benchmark data sets exemplifying these phenomena. Only Lacouture's (1997) and Stewart et al.'s (2005) data show effects that are strong enough to require explicit modeling by the relative judgment process in SAMBA.

In the following sections we first provide details of SAMBA, then describe a set of benchmark empirical phenomena and associated classic data sets. Along with these

descriptions we show that SAMBA accurately fits each of these classic benchmark data sets. The benchmark sets presented here were chosen in order to provide insight into the workings of SAMBA and to illustrate its account for phenomena beyond the scope of any of the models from which is was derived, with a particular emphasis on sequential effects. Note that SAMBA also accounts for many other benchmark phenomena that we do not have space to illustrate here, such as the effects of set size on RT. Finally, we present comprehensive fits of SAMBA to two data sets: Lacouture (1997) and Stewart et al.'s (2005) Experiment 1. We used Lacouture's data to test SAMBA's ability to simultaneously account for all of the choice and RT phenomena in a complex data set using a single set of parameter values. Although Stewart et al. did not collect RTs, their data are important because of experimental manipulations which allow a strong test of SAMBA, and for comparison of its fits to those of the RJM.

The Theoretical Challenge

The paradox of absolute identification is that the task is superficially very simple, yet the performance of participants is both inaccurate and surprisingly complex. Our approach to this challenge is similar to that taken by Ratcliff's (1981) theory of perceptual matching: we explicitly model an integrated architecture for perceptual, memory and decision processes in sufficient detail to obtain predictions for the broadest possible range of observed behavior. However, our approach differs from Ratcliff's, and many theories of absolute identification, which assume that repeated presentations of any given stimulus result in a distribution of internal magnitude estimates, and that certain parameters of the distribution, often the variance, must be estimated for each experimental context in which the stimulus appears. Such approaches are analogous to signal detection theory, providing a successful description of the data without

addressing the deeper question of how these magnitude estimates arise. As well as being less intellectually satisfying, models that begin with parameterized distributions also fail to provide constrained accounts of some of the most fundamental benchmark phenomena. For example, there are powerful effects caused by very simple stimulus manipulations such as set size (*N*), stimulus spacing, and the bow effects due to stimulus magnitude. In a framework similar to Ratcliff's, these effects are modeled by changes in parameter estimates, but this approach fails to provide an explanation of *how* the changes arise.

With SAMBA we adopt a less flexible approach that provides strict constraints, and reduces the number of parameters required. We directly model the process by which an observer *produces* a magnitude estimate when confronted with a stimulus (see also Kent & Lamberts, 2005). We model this process using an extended version of Marley and Cook's (1984) selective attention theory. This theory explains how an observer can attach a numerical magnitude estimate to a stimulus, and how these magnitude estimates are distributed across repeated trials. This establishes a mechanism for *producing* the distributions, and for describing how those distributions change under experimental manipulations, without requiring arbitrary parameter changes. The remaining parts of the SAMBA architecture provide a similarly constrained explanation of the process by which a response is chosen in light of the magnitude estimate.

An even deeper philosophical question arises when considering the theoretical completeness of the simple parametric approach to stimulus representations. The participant's task in absolute identification is to attach a response label, such as #1, #2 and so on, to a stimulus, with the physical stimulus measured in, for example, decibels, Hertz, or metres. A theoretical account is incomplete if it begins by assuming that

numbers are attached to stimuli, and that these are simply transformed into response labels. Lacouture and Marley's (1995, 2004) and Stewart et al.'s (2005) models suffer from this weakness. Put another way, the central task of absolute identification is to associate a numeral with a stimulus magnitude, so it is dissatisfying to consider a theory that begins by assuming that stimulus magnitude estimates are available, with no explanation of how they arise.

SAMBA differs from the simple distributional approach in a second way that greatly reduces the number of parameters that must be estimated in fits to data. Like signal detection theory, simple distributional models typically make decisions using a set of cutpoint or referent parameters. The disadvantage of this approach is that the number of referent parameters grows linearly with the number of response alternatives and usually no explanation is provided of how participants choose appropriate values. SAMBA adopts a framework similar to that of Petrov and Anderson's (2005) ANCHOR model; we assume that participants learn average magnitude estimates corresponding to each response, and that these magnitudes act as referents. As this learning is assumed to be accurate, at least when feedback is correct, the referent values are entirely determined by the experimental design, and hence in model fitting they do not have to be estimated. Consequently, SAMBA's flexibility in fitting data is greatly curtailed. Lacouture and Marley's (1995) parameter free mapping model is used to transform referent estimates to what are, effectively, a set of tuning curves that provide input to SAMBA's decision stage. Hence, we meet the theoretical and practical challenge posed by the potentially large number of responses in absolute identification, as the number of parameters required to fit SAMBA does not change with the number of responses.

Model Overview

SAMBA integrates three successful elements from previous models. It uses the time-dependent (i.e., not the asymptotic) version of Marley and Cook's (1984) selective attention model of stimulus representation, and Brown and Heathcote's (2005) ballistic accumulator model of response selection. These model elements do not constitute a complete account because the selective attention component produces a single magnitude estimate while the ballistic accumulator component requires *N* numerical inputs, one for each possible response. We link the two components with Lacouture and Marley's (1995) mapping process. The vertical integration of model components is an important feature of SAMBA. Often in cognitive psychology, different levels of processing are considered separately, and models are developed independently for each. Greater model constraint and explanatory power can be achieved by integrating models to span several levels of explanation, from stimulus representation to response selection (see Ratcliff, 1981, for a similar approach to perceptual matching). The three elements we borrow from prior models successfully explained some aspects of absolute identification behaviour, but not others, making each incomplete. Some of these problems are naturally fixed by integration within SAMBA, but others are not. We address these remaining issues by modeling the selective attention and ballistic accumulator processes at the level of the duration of the trial; the model can be specified in real time as more temporally fine-grained data becomes available (see the ***General Discussion***).

Figure 1A illustrates the three stages of the model. SAMBA's first stage is a modified version of Marley and Cook's (1984) selective attention theory. The selective attention (SAMBA) stage maintains a representation for the context of the experiment,

and uses this context to produce estimates of sensory magnitude. The context representation is maintained by activation of a range of units that are in one-to-one correspondence with stimulus magnitudes, such as line lengths or intensity of tones of the same frequency. Input to the selective attention stage comes from a relatively accurate psychophysical representation of each stimulus. We assume that the stimuli are represented topographically at the psychophysical level and that psychophysical input causes selection of a corresponding unit. Importantly, neither the psychophysical representation nor the selected unit directly provides a numerical estimate of the stimulus magnitude. In this, we agree with Krantz' (1972, p. 175) view that "I do not see how sensations could be paired directly with numbers at all". Instead, stimulus magnitudes are estimated by the summation of activity between the unit selected by the psychophysical input and the ends of the active context.

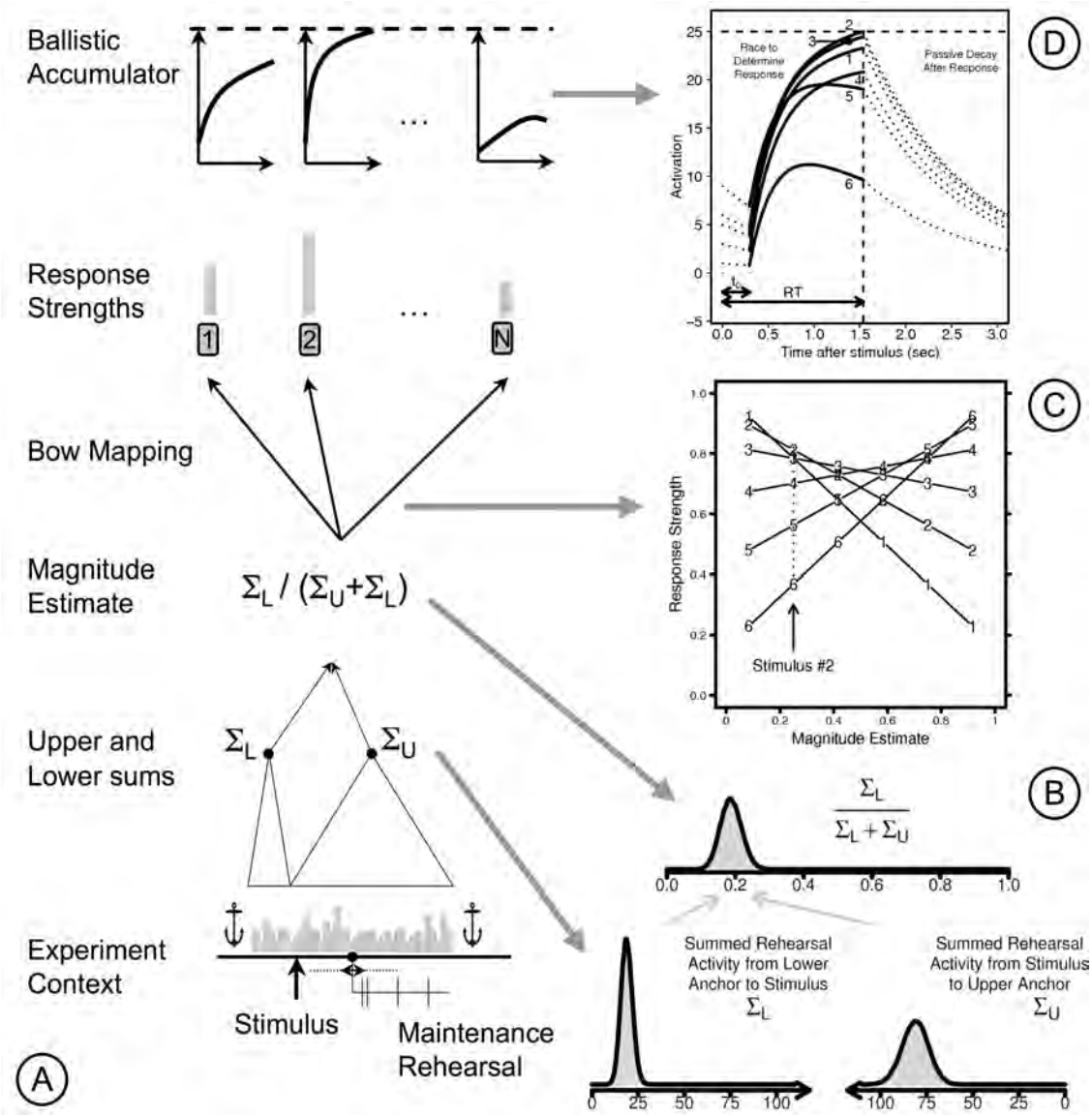

Figure 1. The SAMBA model. Panel A illustrates the model at a general level, to be read from bottom to top. Panel B shows how the magnitude estimate produced by the selective attention stage varies over trials. Summed activities from the stimulus location to the lower and upper anchors are combined to make a magnitude estimate in the interval [0,1], and this varies with changes in the activity of the Poisson accumulators. Panel C shows how the magnitude estimate is transformed into N response strengths by the mapping process. In the example, a magnitude estimate of 0.25 (corresponding to stimulus #2) is transformed into N=6 response strengths, shown by the heights of the six lines above x=.25. Panel D shows how the ballistic accumulator stage makes a response decision. The response strengths from the mapping stage drive ballistic accumulators, with the first one to reach a threshold determining the choice. Between trials, the activity in the ballistic accumulators decreases by passive leakage.

This magnitude estimate is transformed into *N* response strengths[6] – one for each

6 We use "response strengths" instead of "mapping outputs" or "drift rates" simply as a mnemonic convenience. We hope the reader will be reminded by this terminology that these (unobserved)

of the *N* possible responses – by Lacouture and Marley's (1995) mapping. The transformation works using a long-term memory for each stimulus, given by the *average* of its magnitude estimates on previous trials. The final stage of SAMBA is an elaborated version of Brown and Heathcote's (2005) ballistic accumulator. This stage takes the *N* response strengths produced by the mapping and assigns each one to a separate decision accumulator. The activations of these accumulators increase at rates determined in part by the response strengths output from the mapping, and in part by the dynamics of the accumulators, including mutual inhibition. A response is made as soon as any accumulator's activation exceeds a response threshold. RT is modeled as the time taken for that accumulator to reach threshold plus a constant amount of time taken for non-decision processes. We now describe each stage in more detail.

As input to SAMBA we assume a simple spatial psychophysical stimulus representation corresponding to topological projections in the sensory areas of the brain, such as retinatopic or tonotopic maps (e.g., Romani, Williamson & Kaufman, 1982; Wiemer & von Seelen, 2002). The physical magnitudes of stimuli are mapped quite accurately onto this psychophysical dimension, with only a small amount of variability. Most absolute identification experiments involve stimuli that are sufficiently separated so that psychophysical variability does not cause errors. Hence, variability in the psychophysical stage is neglected in the fits of SAMBA that we report, except in two cases where stimuli are closely spaced.

The selective attention stage produces a magnitude estimate from the ordinal psychophysical representation. This process, with the addition of decision cutpoints, has had considerable success in fitting classic choice data in absolute identification (Marley & Cook, 1984) and also magnitude estimation data (Marley & Cook, 1986). The

quantities drive the ballistic accumulators towards making an overt response.

selective attention model has been successful because it provides a stimulus representation that dynamically adapts to changes in the experimental design, such as stimulus spacing and choice set size. Although motivated differently, this stage performs similarly to the successful theory of sensory trace and context coding (see, e.g., Berliner, 1973, Berliner Durlach & Braida, 1977, and Braida et al., 1984) and to the attention band proposals of Weber, Green and Luce (1977), Luce et al (1982) and Nosofsky (1983).

Selective attention gives SAMBA a mechanism by which to attach context- and time-dependent numerical representations to stimuli and stimulus differences, without making arbitrary assignments. For example, suppose an experiment uses pure tone stimuli, all of the same frequency but of different intensities. Over pre-experimental training and initial trials, the participant might estimate that the stimuli range is between 50dB and 70dB, and so the context becomes the range between these values. Importantly, the selective attention stage does not assume that the psychophysical representations are addressable – that is, there are no numbers associated with the psychophysical locations. Instead, we explicitly model the process by which participants estimate stimulus magnitudes without having access to numerical tags. A finite set of leaky accumulators is put in a one-to-one ordered correspondence with the psychophysical dimension. Presentation of a test stimulus results in two effects. Firstly, a corresponding location on the topographic psychophysical representation becomes active (i.e., is selected). Secondly, the accumulator that corresponds to the psychophysical representation of the stimulus is selected. The selection of this accumulator allows the participant to identify its *ordered* location within the array of units. That is, the observer can identify which accumulators are below, or above, the one

selected by the presented stimulus.

To aid in understanding how the selective attention process helps to determine magnitude information from an ordered set of accumulators consider an analogy. Imagine a very long row of lamps on the wall of a room. A contiguous range of the lamps are lit green, but with flickering intensities. The lit lamps represent the range of stimuli in an experiment and the brightness of each lamp corresponds to the current level of activation of each accumulator. Now suppose that the only method an observer has of measuring the activity of the lamps is to gauge the total brightness of portions of the array, but the absolute position of any single lamp. Further suppose that, if any particular lamp changes to red, an observer can select it and use this location to partition the array into those lamps above and those below. With only these abilities, the observer is able to *estimate* (relative) magnitudes using the total brightness of the green lamps below the selected red lamp, and the total brightness of the green lamps above the selected red lamp. Each of these sums will be large or small depending upon the position of the currently lit red lamp, so they carry relative magnitude information. Finally, as with a flashlight shining on a surface, the average total intensity of the lit lamps is fixed, independent of the range of lit lamps.

Suppose, for example, that presentation of a #2 stimulus (a 52 dB tone in a set of range 50-70 dB) corresponds to a lamp near the lower end of the green range turning red. The separation between the lower end of the row of green lamps and the red lamp is small, so the sum of the intensity of the lamps between the two will also be small. However, the estimate of magnitude is noisy because the lamps are flickering, and so the sums of their intensity will also vary. Other measurements are also possible – for example, if two lamps were to change color, the sum of the intensity of the lamps

between them could be estimated in a similar way.

We call the first and last active accumulators on the attention dimension "anchors"; they correspond to stimulus magnitudes that span the range used in the experiment. The anchor positions are assumed to be under the direct control of the participant. The entire range of accumulators between the anchors is kept active by continual attention during an experiment, analogous to the process that keeps the lamps flickering. Activation is modeled as a Poisson process of mean rate λ events per trial, and so we will sometimes refer to the accumulators as "Poisson accumulators". Without loss of generality, we assume each Poisson event increases the activity of a randomly selected accumulator by a unit amount. This activation is combined with a passive decay process: in the absence of attention, each accumulator's activity decreases by a factor of α over the course of a trial[7]. The combination of Poisson activation and passive decay results in each accumulator having an activation value that varies from trial to trial (Marley & Cook, 1984). The average total activity in all accumulators is set by the balance between the attention and decay rates, namely $\eta=\lambda/(1-\alpha)$. This average total activity is the major determinant of the overall accuracy of responses in SAMBA.

Stimulus magnitudes are estimated by summing activity in sub-ranges of accumulators, in particular, the total activity from the upper anchor (U) and from the lower anchor (L) to the current stimulus. Taking the sums between the current stimulus and each of the anchors produces two estimates of stimulus magnitude, one relative to the lower anchor (Σ_L) and one relative to the upper anchor (Σ_U). These are combined into a single magnitude estimate by the ratio $\Sigma_L/(\Sigma_L+\Sigma_U)$, which is naturally constrained

7 The parameters λ, α and η are similar to Marley and Cook's (1984) parameters of the same names, but differ because we use discrete time (trials) and they used continuous time. If T is the duration of a trial, then our λ corresponds to their λT, our α corresponds to their $e^{-\alpha T}$, and our η is the reciprocal of theirs.

to be between zero and one. As shown in Figure 1B, this magnitude estimate varies from trial to trial, even if exactly the same stimulus is repeated, because the activities in the Poisson accumulators vary as a result of the attention process. Importantly for absolute identification, the variability of the magnitude estimate is largest in the centre of the range, causing SAMBA to predict a bow effect in response accuracy and sensitivity. This projection of psychophysical stimulus representations for all stimulus modalities onto a common bounded interval is supported by research on magnitude estimation and cross-modality matching by Krantz (1972), Teghtsoonian (1973) and Teghtsoonian and Teghtsoonian (1978, 1997).

The observer accumulates magnitude estimates $\Sigma_L/(\Sigma_L+\Sigma_U)$ from each trial and stores an average magnitude estimate for each stimulus in a long-term memory. We assume that this long-term memory becomes stable relatively quickly, particularly when accurate feedback is provided. Since we aggregate data over many trials, we approximate the learning and memory process by the assumption that a participant has available an accurate memory for the average magnitude estimate corresponding to each stimulus. These memories, combined with the end anchors, constitute the set of referents stored in long-term memory that underpin SAMBA's absolute process, similarly to the anchor values in Petrov and Anderson's (2005) ANCHOR model, or the cut points in Stewart et al.'s (2005) RJM. The learning of referents is an important issue (see, e.g., Petrov & Anderson), and is an area where SAMBA could be extended in future. When discussing a false-feedback experiment below, we explore a simple beginning to a referent learning mechanism in SAMBA.

The magnitude estimate produced by the selective attention stage must be transformed into N response strengths in order to provide inputs for the ballistic

accumulator stage. The transformation is made by Lacouture and Marley's (1995, 2004) mapping process, illustrated in Figure 1C and presented in more detail below. The mapping is error free relative to its referents: for example, if the estimate of stimulus magnitude is closest to the referent for stimulus #2, then the mapping will provide a response strength that is largest for response #2 (strengths for responses #1 and #3 will be next, and so on).

On each trial, the outputs of the mapping stage are analogous to a tuning curve over the possible responses, with the response strengths contingent on how well the current input matches the long-term referent for each stimulus. In contrast to the multiple parameters of other tuning curve models, the positions, widths and shapes of the tuning curves produced by the mapping stage are determined entirely by the values of the referents held in long-term memory. The fixed form of Lacouture and Marley's mapping provides significant constraint, greatly limiting SAMBA's flexibility for fitting choice and RT data. For instance, an alternative approach, explored by Karpiuk et al. (1997), linked the output of the selective attention stage with the input to the decision stage via parameterized tuning curves, given by Link's (1992) wave theory. This framework has great flexibility to adjust decision cutpoints, whereas the mapping solution has a limited ability to adjust such cutpoints. Our position is that, until clearly required to fit data, the more constrained model is preferable.

To illustrate how the mapping operates, and also how it naturally handles unevenly-spaced stimuli, consider an experiment from Lockhead and Hinson (1986). In one part of this experiment, the stimuli were tones of intensity 58dB, 60dB and 66dB. Suppose that a participant placed the lower anchor 2dB below the lowest stimulus (i.e., at L=56dB) and 4dB above the highest (i.e., at U=70dB). With this setup, regardless of

any parameter settings, the selective attention stage will produce stimulus magnitude estimates for the three stimuli with average values of {1/7, 2/7, 5/7}. Note that these magnitude estimates naturally reflect the unequal spacing of the stimuli – three times the spacing between the upper two stimuli than between the lower two stimuli, and this property holds regardless of the locations that the observer selects for the anchor values (*L* and *U*), provided the stimulus locations lie between them..

Continuing this example, the mapping transforms a magnitude estimate into three response strengths, one for each of the three possible responses (#1, #2, or #3). The computations of the mapping stage are specified entirely by the long-term average stimulus magnitude estimates. A magnitude estimate, say *z*, is linearly transformed into a response strength for each and every response j=1…N according to $(2Y_j-1)z-Y_j^2+1$, where Y_j is the average magnitude estimate for the jth stimulus[8]. For the Lockhead and Hinson (1986) example, suppose the 60dB stimulus was presented. On this particular trial, the selective attention stage might produce a magnitude estimate of .3, which is quite close to the long-term average value of 2/7 for this stimulus. The mapping stage transforms this value into three response strengths: the strength for response #1 is $(2\times\frac{1}{7}-1)\times 0.3-(\frac{1}{7})^2+1 = 0.765$; for response #2 it is $(2\times\frac{2}{7}-1)\times 0.3-(\frac{2}{7})^2+1 = 0.790$; and for response #3 it is $(2\times\frac{5}{7}-1)\times 0.3-(\frac{5}{7})^2+1 = 0.618$. Notice that the strength is greatest for the correct response (#2).

Figure 1C illustrates the mapping solution for a more standard experiment, with six evenly-spaced stimuli, and six lines corresponding to the response strengths for the six possible responses. In the example, stimulus #2 is presented so the selective

8 Lacouture and Marley (1995) motivate this parameter free solution by requiring a solution that: a) makes use of all of the bounded input and bounded output range in every experiment and b) has cutpoints that are midway between the mean activities for adjacent stimuli.

attention stage will – on average – produce a magnitude estimate of .25, shown by the vertical arrow in Figure 1C. From this magnitude estimate, the mapping produces six response strengths, shown by the heights of the six lines immediately above the arrow. As required for a correct response, the strength corresponding to response #2 is greatest, with responses #1 and #3 next, response #4 after that and so on.

The mapping stage brings many useful properties to SAMBA. Firstly, it provides a powerful way to simplify the model and greatly reduce the number of free parameters required to produce response time predictions. Models of response time distributions for N-alternative tasks generally have some multiple of N^2 parameters (e.g., see Busemeyer & Townsend, 1992, 1993). These parameters specify response strengths or drift rates for each response, contingent on each stimulus. Lacouture and Marley's mapping replaces this entire set of $O(N^2)$ parameters with a process for *producing* drift rates given a stimulus magnitude estimate as input.

The outputs of the mapping component provide the input to the final response selection or decision stage of SAMBA, which is based on Brown and Heathcote's (2005) ballistic accumulation model. The ballistic accumulator model brings three important properties to the modeling absolute of identification data. First, the use of an accumulator process allows SAMBA to make very detailed predictions about response time. Second, the process of passive leakage in ballistic accumulator activities provides a natural explanation of very short-term sequential effects without the need to posit extra processes. Third, the ballistic accumulator includes competition - also called lateral inhibition - between accumulators, and this naturally means that responses become slower as the number of alternatives (set size, N) increases.

The decision stage associates each of the N responses with a competitive

ballistic accumulator (Brown & Heathcote, 2005; see also Usher & McClelland, 2001). The inputs to the decision accumulators are the outputs of the mapping stage, with a common (single) sample of zero-mean Gaussian noise added to every response strength, with the standard deviation (σ_M) estimated from the data. The noise is analogous to the "drift variance" included in almost all models of choice RT, as first introduced by Ratcliff (1978, his parameter η). In our applications, the estimated standard deviation of the noise (σ_M) was sufficiently small so that at least one decision unit always received a positive input. As the same sample is added to all inputs, σ_M models stimulus independent variability due to global factors such as fluctuations in arousal. The value of σ_M has little influence on accuracy, because it does not alter differences between inputs. Hence, σ_M has a selective influence on variability in RT, which is also determined by sequential effects in the decision stage discussed in detail later[9].

Figure 1D illustrates an example decision process. Each of the *N* response accumulators begins the decision stage with a starting activation level determined by previous inputs. The starting levels are different for each accumulator, and the activation levels increase deterministically at rates dictated by the response strengths from the mapping stage and competition amongst accumulators. Activation x_j in accumulator j=1…*N* changes according to a linear, first order system of differential equations:

$x'_j(t) = I_j - b \sum_{p \neq j} x_p(t)$. Here, I_j represents the response strength from the *j*th unit of the mapping stage, plus the Gaussian noise sample with standard deviation σ_M. The

9 A reviewer questioned whether the σ_M parameter might be replaced by within trial variability in the decision stage, or perhaps another source of variability in the magnitude estimates. Other sources of variability affect error rates, which σ_M does not, and would thus increase model flexibility, by providing another mechanism for modelling error responses. Our decision not to include this extra flexibility was based on parsimony, as the data we examined did not seem to require it. However, we have no grounds to rule out other types of variability.

parameter $\beta>0$ represents lateral inhibition, and causes the increase in the activation to be nonlinear. The system of coupled differential equations describing accumulation during the decision stage can be solved analytically by matrix algebra, for any response set size *N* (see Brown & Heathcote, submitted).

A response is chosen corresponding to the first accumulator to reach a threshold (*C*, the same value for all accumulators) and the response time is given by the time taken to reach that criterion, plus a constant time for non-decision processes, t_0. This system of accumulators ensures that a finite response threshold will *always* be reached in a finite time. This occurs because at least one input is positive and because we use a simplified version of the ballistic accumulator model with no passive leakage within a trial (see Brown & Heathcote, 2005, for further details). The example activation trajectories in Figure 1D correspond to the mapping example illustrated in Figure 1C. The trajectory corresponding to the correct response (#2) increases fastest, and reaches the threshold first, so a correct response would be made in this example with a response time of just over 1.5s. If the response threshold were set lower, say at C=20 instead of C=25, the response time would be faster, around 0.9s, but an error would be made: the model would give response #3, instead of the correct response (#2). The error occurs because the ballistic accumulator for response #3 began the decision stage with an advantage over the accumulator for response #2, and it takes some time for this advantage to be overcome.

The decision stage of SAMBA is more constrained than Brown and Heathcote's (2005) ballistic accumulator model, and most other models of two-choice RT (see Ratcliff & Smith, 2004, for a recent summary), in its assumptions about the starting points of the evidence accumulation processes. These models assume that accumulation

for each response unit starts from a random value. In SAMBA, the accumulator start points are completely determined by passive decay from their values at the end of the previous trial (also see Laming, 1968). This mechanism not only specifies previously unspecified details of the genesis of start point variability in theories of choice RT, it also enables the model to explain short-term sequential phenomena in absolute identification, including assimilation and response repetition effects. For typical parameter values, about one quarter of incipient choices generated by the selective attention and mapping stages of SAMBA are changed by the ballistic accumulators, due to differences in the starting points of the evidence accumulation processes, left over from the previous trial.

Brown and Heathcote's (2005) ballistic accumulator model, used in SAMBA's decision stage, has been simplified further by Brown and Heathcote (submitted) by omitting the lateral inhibition term (i.e., by setting $\beta=0$), making a *linear* ballistic accumulator model (LBA). The LBA model has computational and analytic advantages over the ballistic accumulator model, but nevertheless we choose not to use it in SAMBA. The lateral inhibition element in the ballistic accumulator makes SAMBA predict increased response time with increased set size (N), without the need for any parameter changes with set size. As discussed by Brown and Heathcote (submitted), if LBA were used in SAMBA then an extra assumption would be required to fit the increase in response time with set size. For example, the LBA would predict increased RT with set size if the outputs of the bow mapping stage decreased with increased set size (e.g., by fixing the sum of the response strengths across set size).

An Absolute Account of Sequential Effects on Errors

SAMBA makes strong predictions about sequential effects in errors and RT,

because it unambiguously attributes contrast effects to processes producing a context-dependent stimulus representation, and repetition and assimilation effects to the processes producing a response. As in other models of choice RT, Brown and Heathcote (2005) assumed that the start points for the ballistic accumulators (x_0) were independent random samples from a common uniform distribution. SAMBA replaces this assumption with a deterministic mechanism based on passive leakage. This passive leakage is illustrated in Figure 1D, where dotted lines show how accumulated evidence decays during non-decision and between-trial times. Computationally, after each decision is completed, the lateral inhibition and stimulus input processes ($I_j - \beta \sum_{p \neq j} x_p$) are replaced by a passive decay process. Decay returns each accumulator's activation exponentially back towards zero at a constant rate: $x'_j(t) = -\gamma x_j(t)$. Therefore, after a constant inter-trial interval (ITI) each accumulator begins the next decision process with activation Dz, where z was the activation level for that accumulator when the previous response was made, and D is given by $exp(-\gamma.\text{ITI})$. In our model fits, we make the approximation that ITI is constant throughout the experiment, and estimate only the value D. At the beginning of each decision process the accumulators corresponding to the previous response and those for nearby responses have an advantage, as illustrated in Figure 1D. This results in response assimilation and in RT and accuracy advantages for repeated stimuli.

Contrast effects arise through reallocation of activity in the selective attention stage. In particular we assume that the participant has some control over the Poisson process, and that this control is used to preferentially attend to locations selected by recently presented stimuli. Reallocation is modelled by a probability (M) that, each time

an accumulator is to be incremented by the Poisson process, the increment is directed to the accumulator selected by the most recently presented stimulus. Consequently, the activity of accumulators selected by recently presented stimuli is increased, which in turn makes magnitude estimates near those values larger. These locally larger values cause a contrast effect that decreases with time as activity in the accumulators decays.

To keep the model as simple as possible, we made the default assumption that participants re-direct activity only towards the accumulator selected by the most recently presented stimulus. This assumption causes contrast to have its greatest magnitude at lag=2, after the (stronger) assimilation effect has passed. This pattern matches most absolute identification data, but for data from Lacouture (1997) we observed the peak contrast effect later, at lag=3 or 4. We modeled those data by assuming that the observer *persists* in re-directing activity for *K* trials. As before, each Poisson event has a probability *M* of being redirected, but the redirection is to one of the locations selected by the *K* most recently viewed stimuli.

Integrating Absolute and Relative Judgment

Research into absolute identification has become focused on the distinction between absolute vs. relative interpretations, both for empirical phenomena and theoretical accounts. We think this distinction is not as useful, nor as clear-cut, as others believe. Even the most relative models (such as Stewart et al.'s, 2005, RJM) employ absolute knowledge about the global nature of the experiment: for example, when the RJM is used to model unequally-spaced stimuli, such as in Lockhead and Hinson (1986), the (global, absolute) scaled magnitudes of the stimulus differences must be taken into account in setting the spacing of cutpoints, ensuring that the global stimulus setup is captured in the model.

Our description of SAMBA has, so far, been entirely in terms of absolute processes in the sense that the present stimulus magnitude is evaluated against long-term referents such as the anchors. However, Rouder et al. (2004) observed an accuracy bonus for repeated stimuli, and for stimuli that are very similar to the preceding stimulus, that cannot be accommodated within SAMBA without including a partially relative process. Below, we extend Rouder et al.'s analysis and demonstrate that the accuracy bonus also translates to an RT bonus. The absolute version of SAMBA we have outlined so far accommodates this phenomenon in a qualitative sense; it predicts that responses to repeated stimuli are faster and more accurate. However, it fails to quantitatively fit these data, as the predicted accuracy and RT bonuses are *smaller* than observed.

The RJM also matches the accuracy effects qualitatively, but fails quantitatively by predicting too *large* an accuracy bonus for several data sets (see Brown et al., 2007). Stewart (2007) showed how the RJM could provide a better quantitative fit when modified to be slightly more absolute, by extending the model's memory so that the stimulus two trials back is, on certain theoretically specified trials, used as the basis for judgments in place of the memory for the stimulus one trial back. Similarly, we have found that SAMBA can provide a better quantitative fit when modified to be slightly more relative – confirming that Rouder et al.'s (2004) analysis, and our extension, suggests the need for *both* absolute and relative processes in models such as RJM and SAMBA.

Working from the assumption that absolute identification must include both relative and absolute processes, we have developed SAMBA to provide the first theoretical account to integrate these processes in a consistent manner. The key to the

integration is the use of anchors in SAMBA. In the absolute version of SAMBA, the anchors bracket the smallest and largest stimuli that are important in the experimental context, and incoming stimuli are judged against these anchors, via the ratio $\Sigma_L/(\Sigma_L+\Sigma_U)$. This ratio estimates the magnitude of the current stimulus within the range defined by the two anchors, [*L*,*U*]. On the other hand, in the relative version of SAMBA, a stimulus is judged within a smaller interval, with one of the anchors replaced by the Poisson accumulator selected by the previous stimulus. If the previous stimulus is larger than the current one, the current stimulus is judged relative to the interval between the lower anchor and the previous stimulus; if the previous stimulus is smaller, the current stimulus is judged relative to the interval from the previous stimulus to the upper anchor. These assumptions are similar to assumptions made by RJM's relative judgment process, although there are also important differences (e.g., the RJM uses a separate mechanism when successive stimuli are judged to be equal).

To illustrate SAMBA's relative process consider an example. Suppose stimulus #4 was presented on the previous trial. If stimulus #7 is presented next, it would be judged in the interval [4', *U*], where 4' indicates the Poisson accumulator associated with the previous presentation of stimulus #4. After being restricted to these subintervals, SAMBA works as before[10]. That is, the current stimulus magnitude is estimated by summing up the activation between the current stimulus and each end of the subinterval, i.e., 4' and U, and these two sums are combined into the ratio of the form $\Sigma_{4'}/(\Sigma_{4'}+\Sigma_U)$. Then, the mapping stage operates in the usual manner to transform the magnitude estimate into a set of response strengths, except that the mapping is

10 We have not directly modeled the process by which the observer decides whether the current magnitude estimate is smaller or larger than the previous one, but it would be trivial to do so. Since the numerical sizes of these magnitude estimates are available to the observer, a simple accumulator mechanism could instantiate our assumption of very fast and accurate larger versus smaller decisions.

restricted to the responses commensurate with the subinterval. In particular, zero strengths are given to responses that are smaller than the previous response, if the current stimulus is larger than the previous stimulus, and vice versa. Finally, the ballistic accumulator stage proceeds as normal to select a response, with an associated response time.

This mechanism is partially relative because it uses a memory of the previous stimulus, and it is partially absolute because it still requires one of the anchors from long-term memory, and employs the mapping solution, which is based on long-term memories for stimulus magnitude estimates. In SAMBA, the relative judgment process is under strategic control, so that a participant must choose to use relative judgment. In fitting data from the comprehensive data set of Lacouture (1997), we obtain the best fit by assuming partial use of the relative mapping process. We model this with a parameter *P* indicating the proportion of trials and/or participants that use the relative process – there were insufficient data at the individual subject level to separate these interpretations. Only the analysis developed by Rouder et al. (2004) provides clear evidence for the use of the relative process through accuracy and RT bonuses for repeat and near-repeat stimuli (where a "near-repeat" stimulus is one that is close, in the rank order of the stimuli, to the stimulus presented on the previous trial). We modeled the data from Stewart et al.'s (2005) Experiment 1 with the simpler assumption that all participants used a relative process on all trials. All other analyses, including our fits to assimilation and contrast phenomena, do not depend on the relative process.

Parameters of the Model

Table 1 shows SAMBA's 13 parameters, along with the data characteristics that they affect. A key feature of SAMBA is that the parameters' effects are defined by the

architecture of the model and can be interpreted in terms of psychological processes. This provides strong predictions and allows for tests of selective influence – certain parameters must only be altered by particular experimental manipulations, and those manipulations must affect *only* the parameters in question. In this section we outline some of these parameter constraints. The first stage of SAMBA, our simple psychophysical stimulus representation, has just one parameter (σ_P). Each of the remaining stages are affected by two or more parameters, with one extra parameter (t_0) to account for the sum of the times taken to complete processing in non-decision stages and to make a response after the decision is made.

The first parameter (σ_P) determines the standard deviation of noise in the psychophysical stimulus representation. It only has an appreciable effect on SAMBA's predictions when stimuli are close enough in magnitude to cause errors in comparative judgment (i.e., judgments about stimuli presented simultaneously or in a rapid sequence). With few exceptions the experiments we model use adjacent stimuli that are sufficiently widely spaced so that we can fix σ_P at zero. Four of the next six selective attention stage parameters determine the distribution of Poisson activity in that stage, namely: the mean number of pulses in the Poisson process over one trial (λ), the rate of decay of each accumulator (α), the proportion (M) of activity directed towards the unit selected by the previous stimulus (or stimuli), and the duration in trial units (K) of that direction. Activity is otherwise assumed to be distributed with equal probability across locations. The remaining two selective attention stage parameters, the positions of each end anchor (L and U), determine the range of the accumulators, and allow SAMBA to accommodate asymmetries in data (i.e., more accurate, and slower, responses to larger than smaller stimuli, or vice versa).

Table 1. Model elements.

Stage Affected	*Symbol*	*Description*	*Principle Effect*
Psychophysical	σ_P	Psychophysical noise	Overall accuracy for small stimulus range
Selective Attention		Accumulator Update Equation: $x_{n+1} = \alpha x_n$ + Poisson(λ)	
	η [$\lambda/(1-\alpha)$]	Ratio of the mean number of pulses in the Poisson process to the accumulator decay rate	Overall accuracy and bow in accuracy
	α	Rate of decay for accumulators	Overall accuracy and contrast time course
	M,K	Mean proportion (*M*) and duration (*K*) of activity directed to prior stimulus location	Contrast
	L,U	Position of the lower and upper anchors	Size and symmetry of bow effects
Selective Attention and Mapping	*P*	Probability of using relative process	Sequential effects on accuracy and RT
Mapping		Output for response *j* given magnitude estimate *z:* $(2Y_j\text{-}1)z\text{-}Y_j^2\text{+}1\text{+}N(0,\sigma_M)$	
	σ_M	Standard deviation of noise added to outputs of the mapping	Variability of RT distributions
Decision		Activation change $x'_j(t) = \begin{cases} I_j - \beta \sum_{p \neq j} x_p(t) & pre-decision \\ -\gamma x_j(t) & post-decision \end{cases}$	
	β	Rate of lateral inhibition	Size of bow and set size effects in RT
	C	Decision criterion	Overall RT and accuracy
	D	Rate of decay of decision unit activation during inter-trial time $D=exp(-\gamma.ITI)$.	Assimilation and shape of bow effect
Non-decision	t_0	Non-decision component of reaction times	Overall RT

One parameter - the probability (*P*) of using the relative judgment process - affects both the selective attention and mapping stages, because the relative process replaces one of the anchor values and scales the output of the mapping stage. Another parameter - the standard deviation of noise (σ_M) added to mapping outputs - affects only the mapping stage. Finally, there are three decision stage parameters which have the same value for all decision units: the evidence accumulation threshold (*C*), the strength of competition between the decision units (β), and the decay rate (*D*) of decision unit

activation.

Benchmark Phenomena

There are a plethora of benchmark phenomena in absolute identification, concerned with both the choices made by participants and with the distribution of the times (RT) to make such choices. The parameters we use for the benchmark phenomena are based on the parameter values estimated from Lacouture's comprehensive data set (see Table 3). Only one or two parameters needed to be adjusted from this baseline in order to fit each benchmark experiment. The σ_P (psychophysical noise) parameter was fixed at zero for all fits, except those for Braida and Durlach's (1972) study, and Stewart et al.'s (2005) Experiment 1, since each of those designs contained some conditions with closely spaced stimuli. Where sequential effects were not at issue the local judgment process was not used (i.e., P=0), and only one parameter was estimated for the Poisson process in the selective attention stage: the ratio η, which largely determines overall accuracy. In those cases, effectively only eight parameters were required to fit the data, and where only choices (not RT) were considered this number dropped to six (without t_0 and σ_M). Hence, the model fits are parsimonious and use parameter values that are consistent across several data sets from different paradigms.

Stewart et al. (2005) provide a comprehensive review of benchmark phenomena concerning choices, but not RT, and demonstrate that their RJM accommodates the former. SAMBA accommodates all of the benchmark phenomena listed by Stewart et al. (2005), and others related to RT, as illustrated by our fits to comprehensive data sets below. Absolute models, from which SAMBA derives, have often been unable to account for sequential phenomena, such as assimilation and contrast, so SAMBA's

accounts of those phenomena are covered in detail below. The following section on critical tests addresses data patterns that have previously been thought to rule out absolute models.

Stimulus Spacing Effects

We fit SAMBA to two data sets showing stimulus spacing effects, Braida and Durlach (1972) and Lockhead and Hinson (1986). SAMBA accommodates the effects of unequal spacing as a natural consequence of its architecture, with the physical values of the unequal stimulus spacings providing direct, parameter free, constraints on the psychophysical front end. Braida and Durlach examined the effect of changing the physical range of the stimuli. They performed eight absolute identification experiments, each with 10 pure tones equally spaced in intensity (dB), with a different stimulus range, and hence spacing, in each experiment. The first experiment used a very small range, with the smallest and largest tones differing by only 2.3dB, so adjacent stimuli were sufficiently closely spaced to bring psychophysical noise into play. The remaining experiments steadily increased the range up to 54dB (see Figure 2). Two important effects were observed in these data. First, the overall accuracy, as measured by the amount of information transmitted from stimulus to response, increased as the range increased, but quickly reached an asymptote. Second, stimulus sensitivity, as measured by *d'* per bel (B^{-1}), showed the standard bow effect for large stimulus ranges, but almost no bow effect for small ranges.

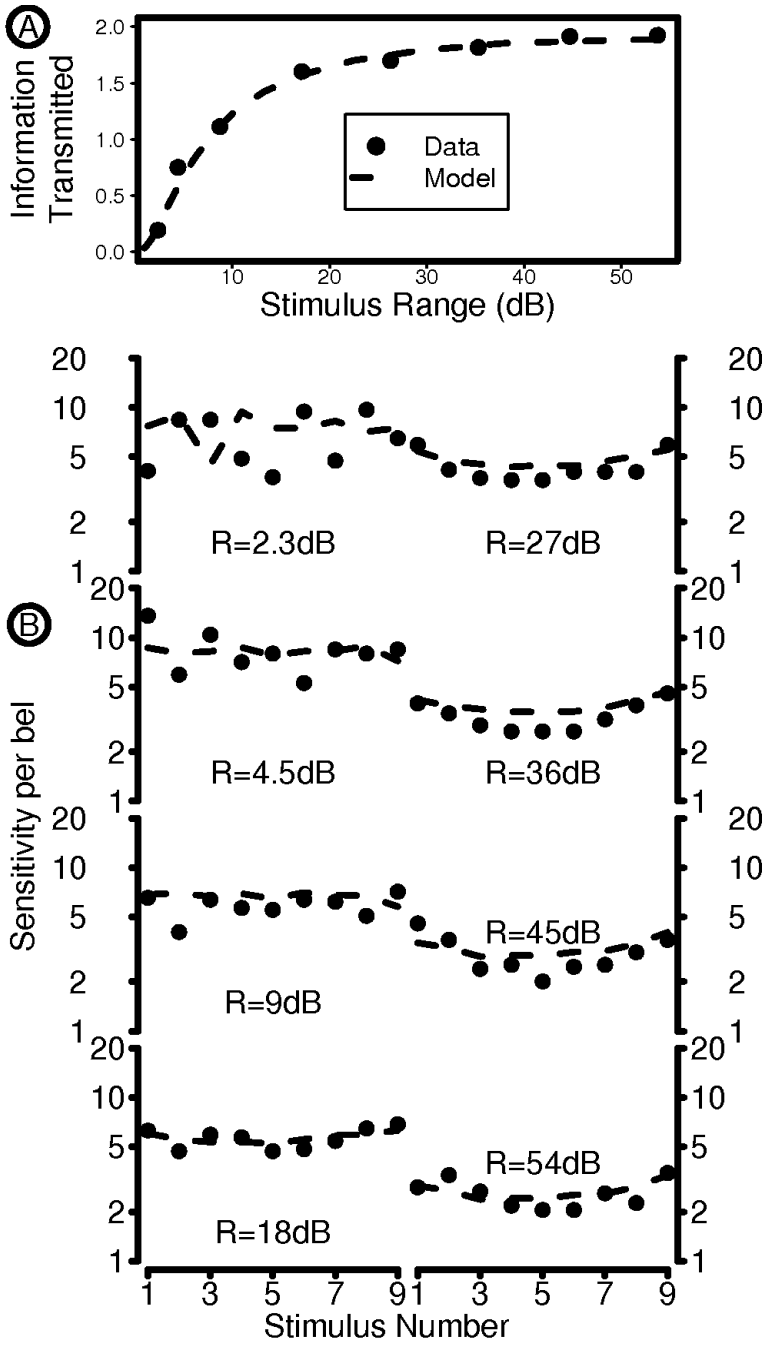

Figure 2. Data from Braida & Durlach's (1972) study (symbols) along with model fit (lines). Panel A shows how information transmitted increases to an asymptote as the stimulus range increases. Panel B shows how sensitivity per bel (B^{-1}) decreases as range increases, with a corresponding increase in the depth of the bow effect.

The upper panel of Figure 2 shows how information transmitted increases from almost zero at the smallest range to an asymptote of about 1.9 bits at the largest ranges. The lower panels show how sensitivity per bel decreases with increasing range, while

simultaneously the depth of the bow effect increases. As illustrated by the dashed lines, SAMBA provides an accurate account of both phenomena. As for all the benchmark phenomena, the model's fits were parsimonious, adjusting only some of a fixed set of reference parameter values used in fitting Lacouture's (1997) data (see Table 3). For Braida and Durlach's data, we increased the selective attention ratio to η=30, and included a psychophysical noise parameter, σ_P=0.96dB. The ratio η was increased to match the overall accuracy level of Braida and Durlach's participants, and the psychophysical noise parameter was changed as it was fixed at zero for Lacouture's (large-range) data. We are re-assured about the interpretability of model parameters by comparison with the estimates given by Marley and Cook (1984). Using the same data set, but using an asymptotic approximation of the selective attention model and decision cutpoints, Marley and Cook estimated σ_P=0.9dB and η=26.

Importantly, no parameters were adjusted between the various stimulus ranges to achieve these fits. Instead, SAMBA captures the effects of increased stimulus range solely through the action of psychophysical noise (σ_P). When stimulus separations are small, psychophysical noise causes confusion between adjacent stimuli, decreasing performance in the small-range conditions. As the stimulus range increases, the separation between adjacent stimuli rapidly grows larger than the imprecision introduced by psychophysical noise. Once the stimulus separation is greater than ~3dB, psychophysical variability (with a standard deviation of 0.96dB) becomes unimportant, and performance asymptotes. The reason that SAMBA predicts almost flat bow effects for small stimulus ranges is that the very high error rates for small stimulus separations is almost exclusively due to psychophysical variability, which is constant across the stimulus range.

A similar limit on absolute identification performance relates to increases in the number of stimuli (*N*) rather than stimulus range. Pollack (1952) and Garner (1953) measured performance in terms of the amount of transmitted information (in bits) when the number of stimuli increased. Their most important observation was a limit – no more than about 2.8 bits of information were transmitted through responses, no matter how large the stimulus set size became. This limit is fundamental to absolute identification, and is incorporated at an appropriately fundamental level in SAMBA. The selective attention process has a limited capacity (the total average activity in all accumulators, η).

This capacity is independent of manipulations such as stimulus set size (*N*) or stimulus range (as in Braida & Durlach, 1972), and it determines the variability of the stimulus magnitude estimates.

Lockhead and Hinson (1986) performed another benchmark experiment that manipulated stimulus spacing using three tones that differed in intensity. In the "equally spaced" condition, adjacent tones were separated by 2dB (at 58dB, 60dB and 62dB). In this condition, confusion matrices (i.e., the probability of each response conditional on each stimulus) were typical of other absolute identification data sets (see Figure 3, middle panel). Lockhead and Hinson created two other conditions by manipulating the spacing of the end stimuli. Compared with the equally-spaced condition, in the "low-spread" condition the lowest stimulus was made much lower (54dB), and in the "high-spread" condition the highest stimulus was made much higher (66dB). The confusion matrices for these two conditions are shown in the left and right panels of Figure 3 (cf. Stewart et al., 2005, Figure 7). The important effect is that the unchanged stimuli (the upper two stimuli in the low spread condition and the lower two stimuli in the high

spread condition) were more often confused in the low spread and high spread conditions than in the equally spaced condition. This poses a theoretical challenge since the relevant pair of stimuli do not differ physically between pairs of conditions, yet they are more often confused when the third stimulus is far away than when it is near. The dashed lines in Figure 3 show that SAMBA can explain these data more parsimoniously than previous accounts – with no parameter changes between conditions. The different predictions in the three conditions arise solely from differences in the stimulus spacing, set directly by the experimental design.

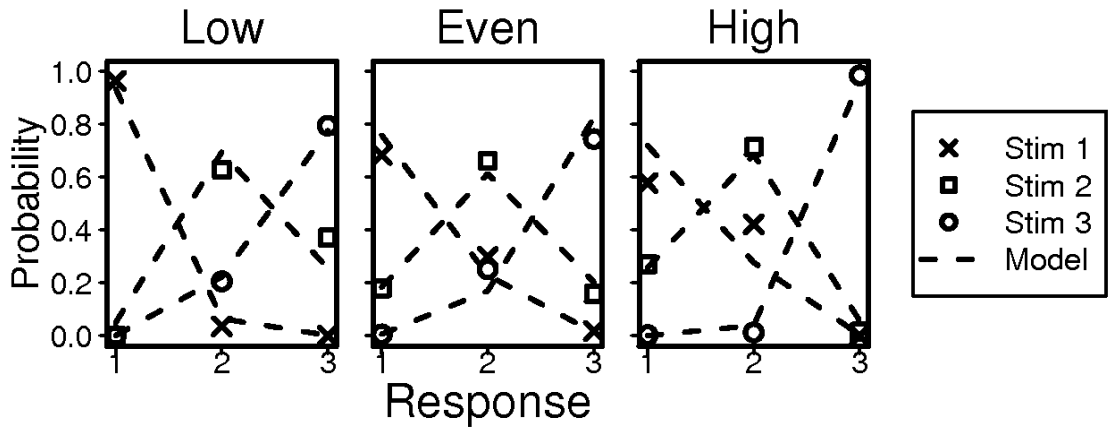

Figure 3. Data and fits for each of the three conditions in Lockhead and Hinson (1986). Each panel refers to a different experimental condition, from left to right: the lower stimulus is much lower ("low spread"); the stimuli are evenly spaced; and the upper stimulus is much higher ("high spread"). Each graph shows the probability of each response (abscissa) conditional on each stimulus (separate lines, see legend).

Only one parameter was changed from our fits to Lacouture's (1997) data, η=10.5, in order to fit the overall accuracy of Lockhead and Hinson's (1986) participants. SAMBA's differential predictions for the three conditions are a natural consequence of the geometry of the model's mapping solution. When the stimuli are equally spaced, the standard map applies, as described in Lacouture and Marley's (1995) original formulation. However, when the stimuli are unequally spaced, the section of the mapping around the closely-spaced stimuli becomes compressed as

consequence of the mean activities in the selection attention stage being close, leading to poorer performance.

Sequential Effects on Errors

The two most studied sequential effects in absolute identification are assimilation and contrast (see: Ward & Lockhead, 1970; Holland & Lockhead, 1968; Mori & Ward, 1995; and Lacouture, 1997). Assimilation and contrast are sequential effects concerned with the distribution of incorrect responses among the possible responses (as opposed to effects concerned with the overall *number* of incorrect responses, which we examine later). Panel A of Figure 4 shows assimilation effects in detail for one benchmark data set, and panels B and C show both assimilation and contrast for that data set and another benchmark data set.

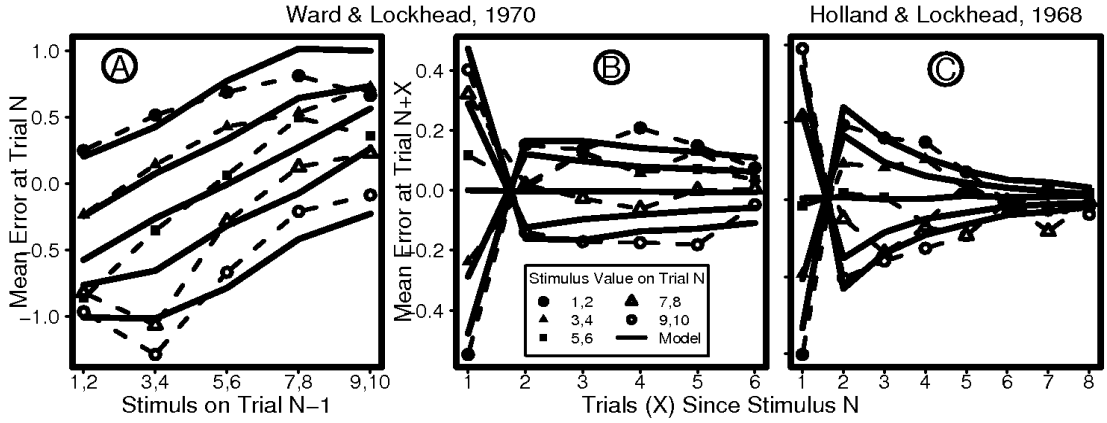

Figure 4. Panel A shows assimilation effects in data from Ward and Lockhead (1970): the average error on trial N was positive when the stimulus on the previous trial (N-1) was large, and vice versa. Panels B and C show both assimilation (at X=1) and contrast (at X>1) in Ward and Lockhead's and Holland and Lockhead's (1968) data. When X=1, assimilation is shown by negative average errors when the previous stimulus was small (filled symbols) and positive average errors when the previous stimulus was large (unfilled symbols). The opposite pattern at longer lags (X>1) is the contrast effect. Solid lines are predictions from SAMBA.

Assimilation means that errors tend to be made *toward* rather than *away from* the previous stimulus. In Figure 4, this is shown using the average error, which is the average difference between the correct response and the actual response. For example, if the correct response is #3 and the subject responds #5, the error is +2. Assimilation is evident at lag 1 (i.e., X=1 in B and C) – that is, average errors are positive (respectively, negative) when the preceding stimulus is large (respectively, small). Contrast is evident in the opposite pattern for longer lags (i.e., X=2-8 in B and C) – that is, the average errors are positive (respectively, negative) when the previous stimulus is small (respectively, large). In SAMBA assimilation is due to prior *responses*, through the starting point of the decision accumulators on the next trial, and contrast is due to prior *stimuli,* through attention allocated to the units selected by previous stimuli.

Creating a theory that simultaneously produces assimilation at short lags and contrast at longer lags is challenging: it must predict that responses are biased towards the previous response, but away from responses prior to that one. SAMBA quantitatively describes assimilation and contrast quite well, as shown by the solid lines in Figure 4. Also, in agreement with data SAMBA could never make the prediction that assimilation and contrast occur at the opposite time scales (i.e., lag=1 for contrast and lag>1 for assimilation). In general, models that set parameters separately for assimilation and contrast *could* make such a prediction, which is problematic (Roberts & Pashler, 2000).

SAMBA cannot make this counterfactual prediction because it is constrained by the nature of its processing architecture. Assimilation is naturally predicted by passive decay in the ballistic accumulators. Between each trial, the activation values of the accumulators decay smoothly back towards a baseline level. This means that the

accumulator corresponding to the response made on the previous trial will begin the next trial with an advantage, and that advantage will also extend to nearby responses, as they typically have activations close to that of the winning response. The rate of decay in accumulator values during the inter-trial interval (parameter *D*) governs the size of assimilation effects, but the effects must always be assimilative, never contrastive. The model's competitive response selection stage also restricts these effects to the previous trial only, never to earlier trials.

Similarly, SAMBA predicts both the direction and time course of contrast due to preferential treatment for recent stimuli in the selective attention process. The Poisson process that activates accumulators is biased towards incrementing the accumulator selected by previous stimuli, normally the most recent. The magnitude of the bias is set by parameter *M*, representing the proportion of activity redirected this way, and the duration of the bias is set by parameter *K*. The contrast mechanism directs extra activity to accumulators selected by recently presented stimuli, and this activity causes the expansion of magnitude estimates which include those locations. For example, suppose the previous stimulus was #2, and the current stimulus is larger, #3. The magnitude of the current stimulus is estimated by summing activity between stimulus #3 and the upper and lower anchors (giving the values Σ_L and Σ_U). The value Σ_L includes *extra* activity in the accumulator selected by the previous stimulus (#2), and so the estimated magnitude of the current stimulus will be larger than it otherwise might be. A parallel argument shows that if the current stimulus is smaller than the previous stimulus, it is judged to be smaller than it otherwise might be. The resulting effect is contrast: stimuli are judged to be further away from recently seen stimuli, and this is the case whatever the relative locations of the current and previous stimuli. Such context effects make

adaptive sense for a participant tracking stimulus distributions that vary over time (Petrov & Anderson, 2005; Ward & Lockhead, 1970).

For the fits in panels A, B and C of Figure 4, we began with parameter values estimated from Lacouture's (1997) data. To fit Ward and Lockhead's (1970) data in Figure 4A and 4B, we changed one parameter that affects assimilation (decay in the decision stage, D=.2), one parameter that affects the magnitude of contrast effects (M=.75) and two that govern its time course (α=.9 and K=2). To fit Holland and Lockhead's (1968) data in Figure 4C, we changed just the one parameter that affects assimilation (D=.17).

Response Time Distributions

Even when response times are collected in absolute identification experiments, they are rarely subjected to the detailed analysis given to response choices. Previous research has identified several effects in mean response times analogous to effects in choice. These include bow effects, in which responses to extreme stimuli are faster than those to middle stimuli (e.g., Lacouture & Marley, 1995, 2004; Lacouture, 1997), and set size effects, where response times slow down as set sizes increase (e.g., Kent & Lamberts, 2005; Lacouture & Marley, 1995, 2004). Sequential effects on mean response times have also been observed due to response repetition (Petrov & Anderson, 2005) and assimilation (Lacouture, 1997). We illustrate SAMBA's ability to accommodate these phenomena in mean RT later, in our fits to Lacouture's data.

An even more stringent model test is provided by fitting full response time distributions. This has rarely been attempted in absolute identification, with a few exceptions, notably Kent and Lamberts (2005) and Lacouture and Marley (2004). We have taken data from all three of those studies, and fit them with SAMBA. We present

the fits to Kent and Lamberts' data and Lacouture and Marley's data here, and those to Lacouture's data in our fits to comprehensive data sets below. Our analyses of the data from Kent and Lamberts' Experiment 1 and Lacouture and Marley are particularly important, as we fit full response time distributions for individual subjects, without averaging. For Lacouture and Marley's data we went one step further, and separately analysed data and model predictions for the RT distributions of both correct and incorrect responses (data for incorrect responses were not available for Kent and Lamberts).

Table 2. Parameter estimates for Kent and Lamberts (2005, Experiment 1). For definitions and explanations of the parameters, see text and Table 1. For units, see below the Table.

Stage	**Selective Attention**	**Mapping**			**Decision**			
Parameter	η	L^a	U^a	σ_M	β^b	D	C	$t_0{}^c$
Subject 1	75.0	80	330	.227	.053	.0044	806	.076
Subject 2	33.3	80	372	.298	.067	.021	730	.175
Subject 3	46.7	80	330	.231	.094	.010	772	.051

Units: a. pixels; b. per second; c. seconds.

In their Experiment 1, Kent and Lamberts (2005) analysed full response time distributions for three individual subjects, which we summarize using quantiles. For each of the 30 distributions (three subjects by ten stimuli), we calculated the 10%, 30%, 50%, 70% and 90% quantiles: that is, the response time below which 10%, 30%, 50% (i.e., the median), 70% and 90% of the data fall. These quantiles are shown along the bottom row of Figure 5, using three panels, one for each participant. The x-axis measures stimulus magnitude (from 1..10) and the five solid lines on each plot show quantiles calculated from the data. The upper and middle rows of Figure 5 show response accuracy and mean RT, respectively, also as functions of stimulus magnitude. The data show several standard effects. Firstly, there are clear bow effects, where responses to middle stimuli are slower and less accurate than those to edge stimuli. For the RT distributions, these bow effects are greatest in the slow tails (the 90% quantile).

The RT distributions are also positively skewed, with greater distances between the 70% and 90% quantiles than between the 10% and 30% quantiles. SAMBA captures all of these effects very well, and provides a good quantitative fit to the data. The parameters used to fit the model are shown in Table 2. The RT quantile and accuracy data allowed us to estimate the decision stage parameters, the overall accuracy parameter (η), and the anchor parameters (*L* and *U*). The published data contain no information on sequential effects, so we kept those parameters fixed at values estimated from Lacouture's (1997) data (see Table 3).

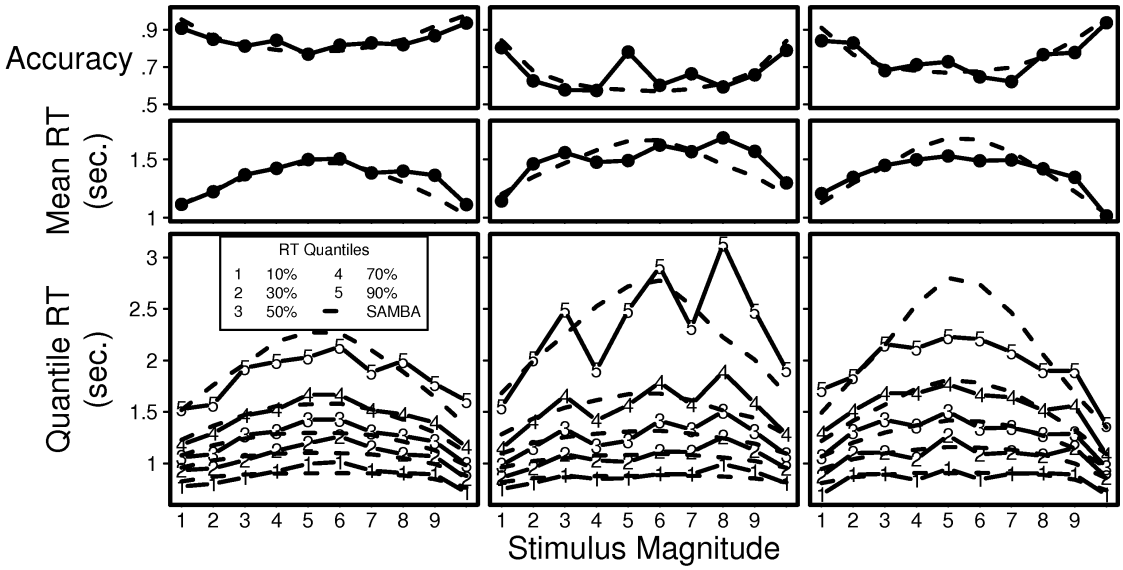

Figure 5. Top panels show accuracy, middle panels show mean RT, and bottom panels show RT quantiles for correct responses for each of three participants (columns) from Kent and Lamberts (2005, Experiment 1). Solid lines correspond to data and dashed lines to predictions from SAMBA. The five numbered lines in the lower panels correspond to the 10%, 30%, 50% (median), 70% and 90% quantiles of the RT distributions for the ten stimulus magnitudes. The reader may note that our mean RT scales are more compressed than in Kent and Lamberts' figures. This compression was required to show the very longest RT quantiles in the figure – Kent and Lamberts do not plot the tails of the distributions in their bow effect plots.

Kent and Lamberts (2005) were unable to fit their model (ECGM-RT) directly to RT distribution data because "Although it might be possible in principle to estimate the properties of the residual-time distribution, the number of simulated trials needed to

produce consistent estimates proved prohibitively large." (p. 297). Instead they fit only mean RT and accuracy. Kent and Lamberts' then used representative parameter values from these fits to generate illustrative RT distribution predictions. The predictions had qualitative trends that matched their RT data, although they were not intended to provide close quantitative fits of the sort we provide in Figure 5. Kent and Lamberts also did not report any data, or model predictions, for RT distributions associated with incorrect responses. Hence, a complete comparison of SAMBA and ECGM-RT will have to await the development of suitable estimation techniques for EGGM-RT. However, a comparison of the top two rows of Figure 5 with the corresponding data in Figure 1 of Kent and Lamberts show that SAMBA's fits just to mean RT and accuracy are comparable to those of the ECGM-RT, even though we did not directly optimize on those quantities. For comparison against possible future models for Kent and Lamberts' data, we note that the χ^2 values for SAMBA's fits, summed across the ten stimuli separately for each participant, are: 291, 235 and 397 for the left, middle and right panels of Figure 5. These are similar to those obtained by fits of other RT models to data from binary choice tasks (e.g., Ratcliff et al., 2004; Ratcliff & Rouder, 1998). These χ^2 values were calculated in the usual manner for RT fits, based on quantiles estimated from the data, making them inappropriate for comparison with theoretical χ^2 distributions.

We were able to access the complete sequence of raw data for the single participant in Lacouture and Marley's (2004) Experiment 2. We analysed the data from the standard absolute identification section of that experiment in which manual responses were used. We fit both the choice and RT data comprehensively with SAMBA, providing accurate fits of all benchmark phenomena including bow effects

and sequential effects. However, for brevity, we present only the analyses of RT distributions. There were 3103 correct responses and 574 incorrect responses that were either +1 or -1 response away from correct. Figure 6 shows the same five quantile estimates as used for Kent and Lamberts' (2005) data, graphed against stimulus magnitude: the left panel show model predictions and data for the RTs of correct responses, the right panel for the RTs for the +/-1 errors. The parameters used to generate these fits are shown in Table 3.

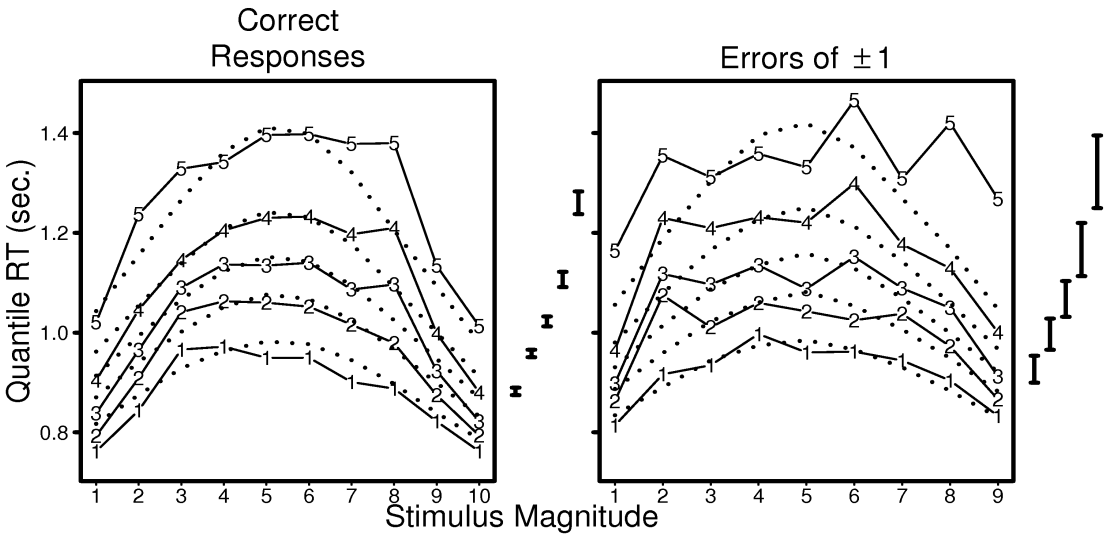

Figure 6. RT quantiles for Lacouture and Marley (2004, Experiment 2). The five numbered lines correspond to the 10%, 30%, 50% (median), 70% and 90% quantiles of the RT distributions for all ten stimulus magnitudes. The left panel shows data and model predictions for correct responses, the right panel for errors where responses were either +1 or -1 away from correct (the undershoot errors have been flipped along the x-axis before averaging with the overshoot errors, to preserve direction). The error bars beside each plot show average standard errors for each quantile, calculated by bootstrap, as in Ratcliff, Gomez and McKoon (2004).

SAMBA fits the shape of the RT distributions for both correct and incorrect responses, as shown by the relative spacing of the quantiles, and it accommodates changes in the shape and variance of the distribution with stimulus magnitude. The most serious misfit is to the slowest quantiles (90%), especially for incorrect responses to the largest (#8, #9) and smallest (#1, #2) stimuli; however, the data were quite noisy for the

incorrect responses especially in the slow tails of the distributions. The data reveal that the standard bow effect – longer RTs for middle than end stimuli – is evident in *all* quantiles, at least for correct responses. That is, both the fastest and the slowest parts of each RT distribution are slower for middle than end stimuli. In binary choice RT modeling (N=2), researchers have sometimes observed quite small effects in the 10% quantile, in the order of tens of milliseconds (e.g., Ratcliff, et al., 2004), and even these have proven theoretically challenging when attributed only to changes in the input to the decision process. The data from Lacouture and Marley (2004) and Kent and Lamberts (2005) show vastly larger bows in the 10% quantile, over 200msec. in magnitude.

Table 3. Parameter estimates for Lacouture (1997), Experiment 1 from Stewart et al. (2005), and Lacouture & Marley (2004). For definitions and explanations of the parameters, see text and Table 1. For units, see below the Table.

Stage	**Selective Attention**					**Mapping**				**Decision**			
Parameter	σ_P	α	λ^b	*M*	*K*	*L*	*U*	*P*	σ_M	β	*D*	*C*	t_0[f]
Lacouture (1997)	-	.75	6 (η=24)	.14	4	91[c]	420[c]	.625	.22	.0307[e]	.07	878	.265
Stewart et al. (2005)	1.9[a]	.92	0.87 (η=7)	.36	3	10[d]		(1)	-	-	.11	-	-
Lacouture & Marley (2004)	-	.80	19 (η=95)	.06	3	86[c]	384[c]	-	.10	0.08[e]	.02	535	.223

Units: a. Percentage of stimulus frequency; b. events per trial; c. pixels; d. percentage of stimulus range; e. per second; f. seconds.

SAMBA is able to account for large effects on the 10% quantile due only to changes in inputs (response strengths) to the ballistic accumulator stage. The ballistic accumulator stage is very tightly constrained – it has only three parameters that are free to vary when fitting response times (t_0, C, and β) without constraint from other aspects of the data (e.g., D is fixed by assimilation). Even with this constraint, SAMBA accounts for the large bow in the 10% quantile, and importantly none of these parameters vary between different stimulus conditions. Further, the bow effect is predicted not by arbitrarily estimating response strengths for each stimulus, but by the way the response strengths are produced by the earlier stages of SAMBA. The original

ballistic accumulator model (Brown & Heathcote, 2005) is unable to accommodate such bows in leading edge RT without making post-hoc assumptions about input strength parameters, assumptions that are superceded by SAMBA's architectural constraints.

Critical Tests of Absolute vs. Relative Models

In this section, we review some empirical results that have been construed as critical tests – that is, as providing qualitative evidence against absolute theories of absolute identification, or in favor of relative theories. Sequential effects on response accuracy have been claimed to provide critical evidence supporting relative accounts of absolute identification, and refuting absolute accounts. We examine several of these effects and illustrate SAMBA's predictions. Each of the effects has a common theme, comparing responses made to stimuli that were preceded by very dissimilar stimuli with responses made to stimuli that were preceded by similar stimuli. The first effect we discuss was identified by Luce et al. (1982), and the second by Rouder et al. (2004). These findings suggested related effects, which we later examine further using data from Lacouture (1997) and Stewart et al.'s (2005) Experiment 1. First, however, we examine the effects of false feedback, which was claimed by Stewart et al. to provide a critical test between the class of relative judgment theories and the class of absolute judgment theories.

False Feedback

We first describe the fit of the revised SAMBA, then the nature of the revision – namely, the addition of a referent learning mechanism. Stewart et al.'s (2005) Experiment 2 involved a standard absolute identification task with equal-loudness tones of different frequencies. On just a few trials in each block, stimulus #3 was presented

but – after making a response – the participant was told that it was stimulus #4. This misleading feedback caused the participants to overestimate the magnitude of the stimulus that followed, as shown in the left panel of Figure 7 by the large positive error for trials following misleading feedback. Note that Stewart et al. reported their data broken down by whether the response on the previous trial was incorrect or correct. However, both RJM and SAMBA predict the same effect whether the previous response is correct or incorrect, and the effect of that variable in that data was small – so Figure 7 presents the results only for trials following a correct response.

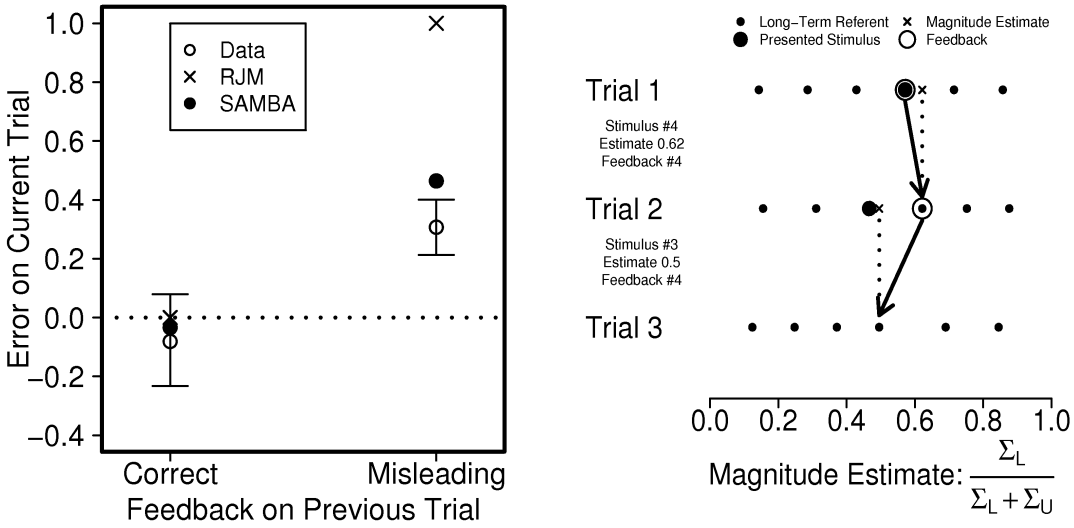

Figure 7. The left panel shows the effect of misleading feedback, with data taken from Stewart et al.'s (2005) Experiment 2. Average error was close to zero when correct feedback was given on the previous trial – subjects performed the task properly. After misleading feedback, participants overestimated the magnitude of the current stimulus. Vertical bars show standard errors. The right panel illustrates the feedback mechanism in SAMBA. The three rows of dots show the positions of the long-term referents for each stimulus, for three trials. See text for explanation.

Stewart et al. (2005) considered this experiment to be a critical test of absolute vs. relative theories. As explained in their Table 8, relative theories (including RJM) predict an average error of +1 following the misleading feedback in Stewart et al.'s

experiment, and zero average error following standard feedback. In contrast, Stewart et al. concluded that existing absolute ("mapping") theories predict a very different pattern: +1 errors for correct feedback following an error response, or for misleading feedback following a correct response, and zero error otherwise. These predictions clearly do not accord with the data. However, they also do not accord with the predictions from the SAMBA model, which fit the data quite well. Figure 7 shows the predictions from the RJM as cross symbols and the predictions of the SAMBA model as solid circles. SAMBA's predictions were obtained using the same parameter values we used in fitting the data of Stewart et al.'s (2005) Experiment 1 (see Table 3). RJM's predictions match the data in a qualitative sense, with zero average error following correct feedback and large positive errors following the misleading feedback, but they substantially overestimate the magnitude of the misleading feedback effect. SAMBA's predictions also match the data in the qualitative sense. Moreover, their quantitative agreement with the data is much better than that of RJM. Importantly, the predictions from SAMBA are unaffected by the inclusion or exclusion of the relative mapping process. These results indicate that a *purely absolute* model can account for the effects of misleading feedback at least as well as a relative model.

We incorporated the effects of feedback into SAMBA in a simple and constrained manner, by partially developing a referent learning algorithm of the sort discussed in detail by Petrov and Anderson (2005). On each trial, the SAMBA model produces an estimate of the stimulus magnitude – the ratio $\Sigma_L/(\Sigma_L+\Sigma_U)$. The mapping stage of the model operates using long-term memories for the mean values of these magnitude estimates for each stimulus, with a value of zero given to the lower anchor (*L*) and a value of one to the upper anchor (*U*). We assumed that feedback helps the

observer maintain these long term referents, as illustrated in the right-hand panel of Figure 7. We provide a simple physical analogy to a spring system, as an intuitive description of the system; mathematically the adjustment is made in a simple linear fashion[11].

The top row of dots in the figure show the long term stimulus referents for a hypothetical experiment with N=6 stimuli. Suppose stimulus #4 is presented, shown by the larger filled circle in the top row, and this presentation produces a magnitude estimate of .62, shown by the small cross. The observer is then provided with correct feedback, shown by the ring around stimulus #4. Feedback allows the referents to be updated for the following trial (second row) so that the referent for stimulus #4 is moved to match its observed magnitude estimate (.62). The long-term referents for the all the stimuli move as if they are locations on a linear spring, whose ends are fixed at zero and one. The point on the spring corresponding to the last magnitude estimate (.62) is deflected so that it aligns with the expected magnitude estimate corresponding to the feedback (stimulus #4). As in a spring system, this deflection causes compression of locations on one side and expansion on the other, with the ends remaining fixed. On the second trial (second row) stimulus #3 is presented, producing a magnitude estimate of .5. *False* feedback now suggests that this was actually stimulus #4, so the long-term referents are adjusted to make the referent for stimulus #4 match the observed estimate (.5), shown by the arrow. In this case, this causes significant compression on the left side and significant expansion on the right side, again with the ends remaining fixed.

The example from the right panel of Figure 7 makes clear the reason for SAMBA's good fit to the data from the left panel of Figure 7. First, adjustments to the

11 If we let Y_i be the long term referent for stimulus i, then given a magnitude estimate of z and feedback indicating that this was stimulus s, the referents for $i \leq s$ are adjusted according to $Y_i \rightarrow zY_i/Y_s$, and for $i > s$ according to $Y_i \rightarrow 1-[(1-z)(1-Y_i)/(1-Y_s)]$.

long-term referents are, on average, smaller when they follow correct feedback than when they follow false feedback. This is because, when there are very few misleading feedback trials, the stimulus magnitude estimate is typically close to the "correct" value stored as the long-term referent, so only small adjustments are required when correct feedback is given. On average, when false feedback is given, much larger adjustments are required, causing greater subsequent errors (and corrections once correct feedback is given again). A referee wondered whether the feedback mechanism makes SAMBA into a more relative (than absolute) model. It is true that the mechanism for updating long term referents implies that information from recently-presented stimuli is used in each judgment, but this does not make the model "relative", once again illustrating the problems with separating absolute from relative models; all contemporary models include elements of both. Even with the referent learning mechanism, judgments in SAMBA are fundamentally *absolute*, as each stimulus is judged against a set of long-term referents and anchors.

One assumption of the referent learning mechanism that may require development concerns the magnitude of updates. We assumed that the long-term referent identified by feedback is moved all the way to the location of the stimulus magnitude estimate. A more realistic implementation may be softer, with the long term referents moving only some fraction of the way. We did not adopt this approach here because it would have required estimation of a learning rate parameter that specifies the relative weights given to the value of the long term referent and the magnitude estimate in the update rule. Such extra flexibility was not required by SAMBA to obtain a reasonable fit to the data.

Further experimental investigation of false feedback effects would also be useful

as SAMBA and RJM make quite different detailed predictions. The RJM predicts the same effect of misleading feedback across the entire range – when false feedback of #4 is given to stimulus #3, the RJM predicts a +1 average error on the following trial, no matter whether stimulus, say, #1 or #10 is presented (see, e.g., Stewart et al.'s Table 8). Also, RJM predicts that misleading feedback has no effect on trials after the immediately-following one. On the other hand, SAMBA predicts the effect of misleading feedback will *decrease* with the difference between the feedback and the magnitude of the next stimulus – because adjustments to the long term referents are smaller away from the feedback than near. SAMBA also predicts small effects of misleading feedback on trials *after* the trial that immediately follows misleading feedback, because the long-term referents remain affected by feedback until corrected by further trials. However, at the level of analysis permitted by Stewart et al. (2005)'s Experiment 2, the effect of false feedback is clearly not a critical test of the class of relative vs. absolute judgment theories, and quantitatively it favors SAMBA over the RJM.

Sequential Effects

Luce et al. (1982) manipulated the difference between stimuli presented on successive trials ("step size") in four conditions (Figure 8). One condition was a conventional 11-stimulus absolute identification design where the stimulus sequence was random – any stimulus could follow any other, with equal probability. This "random step" condition resulted in typical absolute identification data patterns, shown by the accuracy and *d'* graphs in the third column of Figure 8. Luce et al. also used two "small step" conditions, with constrained differences in the magnitude of successive stimuli. In the "small step 3" condition successive stimuli were always very similar: for

example, when stimulus #3 was presented, the next stimulus could only be #2, #3 or #4. In the "small step 5" condition successive stimuli were moderately similar: for example, if stimulus #3 was presented the next stimulus was constrained to be one of #1, #2, #3, #4, or #5. Data from the two "small step" conditions are shown in the left two columns of Figure 8. Finally, Luce et al. used a "large step" condition in which stimuli were always followed by very dissimilar stimuli. For example, when stimulus #3 was presented, the next stimulus could only be one of stimuli #7, #8 or #9. These data are shown on the right of Figure 8.

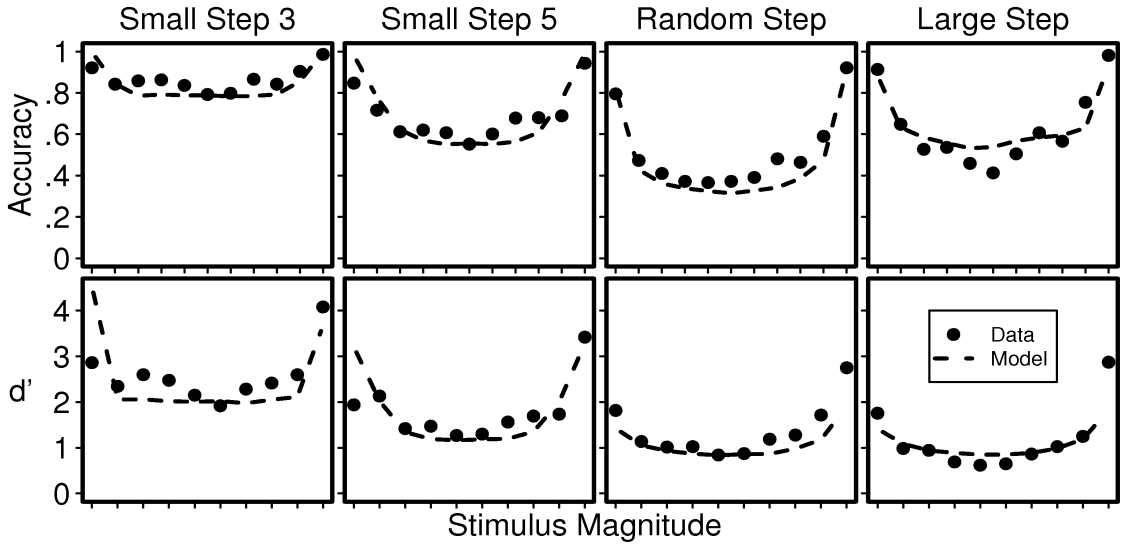

Figure 8. Response accuracy (top row) and sensitivity (bottom) from Luce et al. (1982), with fits of SAMBA (shown by dashed lines). The top left columns show conditions constrained to have small differences between successive stimuli. The third column shows a standard condition (random differences). The right column has data from a condition where differences between successive stimuli were constrained to be large.

The manipulation of step size had large effects on response accuracy, shown in the top row of Figure 8. The random step (standard) experiment resulted in the poorest performance, while the small step 3 condition resulted in the best performance. The small step 5 condition gave better performance than did the large step condition. Some of these effects could be a result of manipulating the number of possible responses,

which may have affected accuracy only via response biases. To check this, Luce et al. (1982) also examined their data using a sensitivity measure (*d'*, bottom row of Figure 8) designed to take into account response bias effects. The *d'* analysis showed that the small step conditions produce greater sensitivity than the large step and random step conditions, but also showed that sensitivity in the large step and random step conditions was quite similar, indicating that the accuracy difference between these conditions was mostly due to response bias.

These data may initially be imagined to refute absolute theories, because they appear to implicate the previous stimulus in the decision process. Nevertheless, SAMBA accurately captures the patterns of Luce et al.'s (1982) data, and does so without inclusion of the relative component. SAMBA correctly predicts the ordering of the conditions in both response accuracy and sensitivity (*d'*) and also provides a good quantitative description of the data. To fit the data from Luce et al., we added two structural assumptions that reflect the nature of the experiment, and we varied three parameters. Importantly, however, we did not vary any parameters across the four experimental conditions, so different predictions for the different conditions represent purely structural effects. We began with parameters estimated from Lacouture's (1997) data and changed the two anchor locations (L=35.5dB and U=90.5dB) and the passive decay parameter (D=.01) to capture the asymmetry and shape of the observed bow effects. We also increased the rate of decay for accumulators (α=.56) in the selective attention stage to match the sequential effects and overall accuracy (η=12.8) for Luce et al.'s data.

The two passive decay parameters (α and D) were estimated to have much smaller values in Luce et al.'s data than in our fits to other comprehensive data sets,

presented later. This difference raises an interesting speculation, and illustrates the testable nature of SAMBA's architectural assumptions. We have specified the decay parameters in units of *trials*, rather than real time – each parameter describes how much activation remains after passive decay for the period of one trial. When adjusting parameters from the fits to Lacouture's data to fit Luce et al's data, each decay parameter was reduced by approximately squaring it: from α=.75 for Lacouture, and α=$.75^2$=.56 for Luce et al.'s data; and D=.07 for Lacouture, and D=$.07^2$=.005 for Luce et al's data. This large change can be parsimoniously interpreted as if the rate of passive decay is *constant in real time*, if the trial-to-trial interval were twice as long in Luce et al.'s study as in Lacouture's. This hypothesis receives some support from the methodological details of each study: the duration of Lacouture's trials was 1.1sec, plus response time; the duration of Luce et al.'s trials was 1.5-2sec, plus the slowest of the three subjects' response times.

Two structural assumptions were made to accommodate the stimulus sequences in Luce et al.'s (1982) experiment. Firstly, we assumed that participants limited their response set; ballistic accumulators corresponding to any responses other than the allowed responses received no input. For example, suppose stimulus #4 was presented on the previous trial. In the small-step-3 condition, the only allowed responses are #3, #4 and #5, so all other response accumulators were given zero input. The second structural modification instantiated a change in the frame of reference for absolute judgments, via a change in the distribution of activation in the Poisson process. In the small-step-3 and small-step-5 conditions, we assumed that attention was only directed at that range of the accumulators corresponding to allowed responses on the next trial. That is, we assumed that participants were able to focus in their attention on the

appropriate sub-range of stimuli on the next trial.

The second structural assumption parallels suggestions made by Weber et al. (1977), Luce et al. (1982) and Nosofsky (1983) that a roving attention band is moved to focus on appropriate ranges, but that this movement is sluggish. The sluggishness of the attention band is manifested in participants' inability to refocus their attention on units corresponding to possible stimuli, particularly in the large-step condition. This can be interpreted in SAMBA as due to the relatively slow decay of activity in the selective attention stage, which has a half-life of one trial (around five seconds) in these fits. Our explanation of these data also fits with data from Nosofsky's Experiment 1. Nosofsky ran a similar experiment to Luce et al., but included a discrimination condition in which all 11 stimuli were presented in random order, as in the random condition of Luce et al., but the participants' task was simplified. They were only required to judge whether the current stimulus was the same as, smaller than, or larger than, the prior stimulus. Nosofsky observed similar performance in both the discrimination condition and the standard (random) condition for the subset of trials that met the small-step-3 constraints, with that performance poorer than in the actual small-step-3 condition. These results support the idea that improved performance in the small-step conditions is due to more than just the constrained response set in those condition.

A second sequential effect that appears to support relative accounts over absolute accounts was reported by Rouder et al. (2004, see also Brown et al., 2007, and Stewart, 2007). Rouder et al. graphed the probability of a correct response conditional on the signed difference between the current and previous stimulus. The filled circles in the left panel of Figure 9 show the results of Rouder et al.'s analysis applied to data from Lacouture (1997). The data show high accuracy for repeated stimuli at the centre

of the plot, corresponding to zero difference between successive stimuli. The accuracy bonus falls away as the difference between successive stimuli increases, then rises again at the edges of the plot. The rise at the edge of the plot corresponds to improved accuracy for an extremely small stimulus that was preceded by an extremely large stimulus, or vice versa.

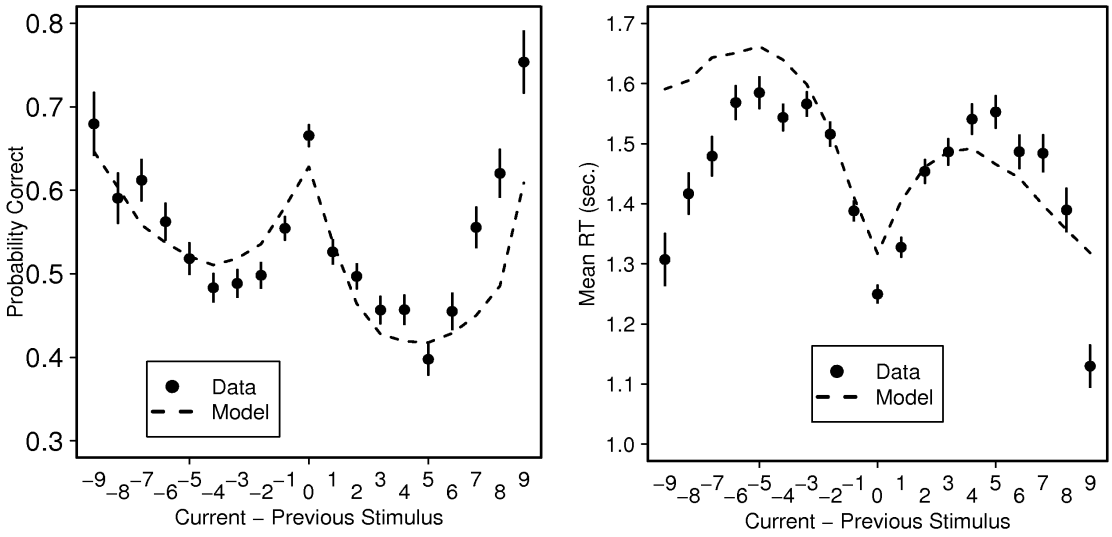

Figure 9. The filled symbols show response accuracy (left panel) and mean RT (right panel) for data from Lacouture (1997) as a function of the difference between stimuli. The graph for Prob. Correct. has a local peak at and near the center, signifying an advantage when successive stimuli are similar. The extremes of each graph also show advantages. The graph for Mean RT has a parallel dip. Error bars show normal standard errors based on the SD of each point over participants, and dashed lines show predictions from SAMBA.

The dashed line in the left panel of Figure 9 shows that SAMBA provides a reasonable fit to the data. The model captures the accuracy advantage for repeat and near-repeat stimuli (graph center) as well as the increased accuracy for very large stimulus changes (graph ends). The extreme ends of the figure are influenced only by responses for the very smallest and very largest stimuli, which are accurate in SAMBA because they are near the ends of the selective attention range. Increased accuracy for repeated stimuli (center of the graph) is caused by two mechanisms in SAMBA. Firstly,

the assimilation mechanism causes repeated responses to have a slight advantage, as the ballistic accumulator associated with the previous response begins the next trial with a higher activation. Secondly, when fitting Lacouture's (1997) data we made partial use of the relative mapping process, which operated on P=62.5% of decisions. The relative mapping process operates by judging the presented stimulus relative to the magnitude estimate of the previous stimulus and the relevant one of the two long-term anchors. The result is increased accuracy for repeat and near-repeat stimuli, as they become effectively judged against a very nearby anchor associated with the prior stimulus. Our model fits to Lacouture's data set do not distinguish whether individual participants used the relative process for 62.5% of their decisions, or 30 out of 48 of participants always used the relative process, and 18 out of 48 never used it, or some combination of the two. Individual data sets were too small for us to carry out individual fits that would reliably differentiate these possibilities.

The right panel of Figure 9 shows that SAMBA captures a similar, but inverted, effect in mean response time. SAMBA's predictions (dashed lines) capture the M-shaped quantitative trend, and match the direction of the asymmetry in the data (faster responses on the right than left). However, this fit fails to predict sufficiently fast RTs near the extreme ends of the graph. The failure is mostly due to unusually fast responses at the ends of the range, especially to the largest stimulus in the set (#10). These data points are discussed in detail later.

Figure 9 shows that repeated stimuli enjoy large advantages, and that near-repeats (such as +/-1 rank-order differences) enjoy smaller advantages. A related question concerns the duration of the advantage for repeated stimuli. Figure 10 shows mean accuracy and RT as functions of how many trials have elapsed since the current

stimulus was last presented. A stimulus repeat corresponds to *no* intervening stimuli (i.e., x=0). As before, repeated stimuli generate more accurate and faster responses. Figure 10 shows that these advantages do not extend further than immediate repeats. With just one intervening stimulus (e.g., the stimulus sequence …. #3, #4, #3, ….) response accuracy and latency are at baseline levels. The dashed lines in Figure 10 show that SAMBA provides a quantitatively accurate account of both the advantage for repeated stimuli (lag 0), and the lack of advantage for other lags. The locus of this account in the model is as described for Figure 9 – a combination of passive decay in the ballistic accumulator stage and partial use of the relative mapping process. Note that SAMBA overestimates total RT in the left panel of Figure 10 – once again, this was caused by unusually fast responses in the data to just one stimulus (#10), discussed later.

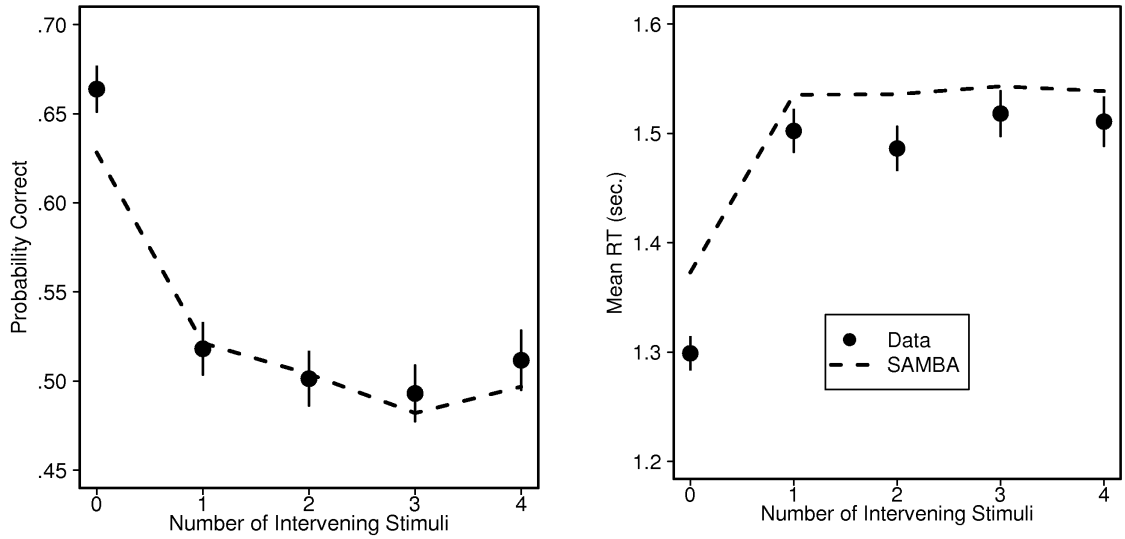

Figure 10. The symbols show response accuracy (left panel) and mean RT (right panel) for data from Lacouture (1997). X-axis shows the number of trials that have intervened since the current stimulus was last presented. Error bars show normal standard errors based on the SD of each point over participants. The dashed lines show predictions from SAMBA.

An effect similar to those described by Luce et al. (1982) and Rouder et al. (2004) was also observed by Stewart et al. (2005). Stewart et al.'s Experiment 1 was a standard absolute identification task using approximately equal-loudness tones of

different frequencies, while also manipulating set size and stimulus spacing (we analyze these data in detail below). Performance was much better for stimuli that were close to the stimulus presented on the previous trial. Figure 11 reproduces Stewart et al.'s figure 26, and graphs accuracy against stimulus magnitude. The graph has separate lines for those stimuli that were preceded by a close stimulus (either an identical stimulus, or +/- 1 rank order difference) and by a far stimulus (all others). This analysis is similar to Luce et al.'s "small-step-3" condition, and also is equivalent to taking the central three points in Rouder et al.'s analyses, or in our Figure 9.

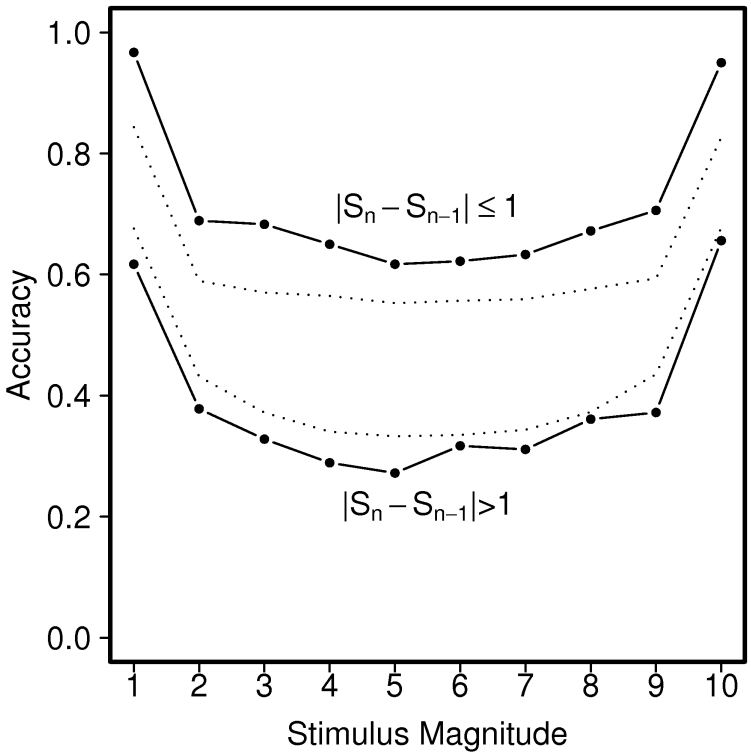

Figure 11. Accuracy graphed separately for stimuli that were preceded by "near" stimuli (either repeats, or +/-1 rank order difference) vs. stimuli that were preceded by "far" stimuli (more than one rank different). The solid lines with filled circles are data from Stewart et al.'s (2005) Experiment 1. The dashed lines are predictions from SAMBA, using parameters discussed below.

In contrast to the results of Rouder et al. (2004), Luce et al. (1982), and

Lacouture (1997), Stewart et al. (2005) observed a very large performance bonus for stimuli that were close to the preceding stimulus –accuracy nearly doubled, from 39% to 72%, and response sensitivity (*d'*, not shown) more than tripled, from 0.78 to 2.7. By contrast, Luce et al. found a *d'* advantage of about 1.1 units, Lacouture found a *d*' advantage of only 0.25 units, and Rouder et al. observed an increase in accuracy of only about 12%, after their participants were well practiced. Further, in an analysis almost identical to Stewart et al.'s, Purks, Callahan, Braida and Durlach (1980) found no significant difference in *d'*. The predictions from SAMBA are shown in Figure 11 by dotted lines. SAMBA provides a reasonable account of the advantage for repeat and near-repeat stimuli, but cannot quite predict the very large bonus observed in the data: SAMBA predicts that response to repeated and near-repeated stimuli are about 1.5 times as accurate as other stimuli, whereas the data show an effect of nearly double accuracy. It seems that SAMBA is sufficiently constrained that it cannot quite predict the extraordinarily large bonus for repeated stimuli observed in Stewart et al.'s Experiment 1. In particular, SAMBA does not include a process specific to the identification of repeated stimuli, which RJM does.

Our preceding analyses suggest that the effects of stimulus repetitions require further investigation, both theoretically and empirically. On the empirical side, previous research has observed a very wide range of effect sizes, from no significant difference (e.g., Purks et al., 1980), to extremely large effects on response accuracy and *d*' (Stewart et al., 2005). Numerous data sets seem to show small but reliable, effects, for example: Kent and Lamberts (2005); Luce et al. (1982); Rouder et al. (2004); Petrov and Anderson (2005); and our analyses of Lacouture's (1997) data in Figures 9 and 10. Two data sets show large effects: Neath and Brown (2006), and Stewart et al. (2005). The

causes of such wide variability in effect size observations are unclear, but may be due to the stimuli chosen in each experiment. The experiments that demonstrated very large effects of stimulus repetition were the only ones to use equal-loudness tones of differing frequency. The identification and discrimination of tone frequency may be rather different than for other stimuli due to the existence of critical bands (Green & Swets, 1966, Table 10.1, p.280). These critical bands may allow participants to perform very accurately on stimulus repetitions, effectively using a more powerful than usual sensory memory. Green and Swets further indicate that the width of the critical bands for frequencies around those used by Stewart et al. are in the same range as some of the frequency separations in Stewart et al.'s (2005) experiment.

Our review of the critical tests of absolute vs. relative theories of absolute identification has shown that almost all of the data can be accommodated by a purely absolute version of SAMBA. The only phenomenon that implicates a relative judgment mechanism relates to improved accuracy for repeat and near-repeat stimuli, observed in analyses of Lacouture's (1997) data and Stewart et al.'s (2005) Experiment 1. These analyses suggest that the detection of repeated stimuli requires further study, both empirical and theoretical. On the empirical side, it appears that the detection of repeated stimuli is particularly privileged when stimuli are defined by frequency, as opposed to loudness or line length, for example. As for theory, SAMBA's account of response repetition is clearly incomplete, and deserves further development. Preliminary work suggests that expanding SAMBA to include a more detailed account of learning, by continual adjustment of the long-term referents held in memory, helps to accommodate the effects of stimulus repetition without assuming that judgments are relative. We return to this point in the General Discussion.

Comprehensive Data Sets

In this section we fit two data sets – from Lacouture (1997) and Stewart et al. (2005) – in great detail. Together, these two data sets exhibit almost all of the benchmark findings in absolute identification. Hence, our analyses of these data sets test whether SAMBA is able to simultaneously accommodate the various patterns in absolute identification data with fixed parameter values. We use two data sets because they provide complimentary coverage of the domain. Lacouture's data (Session 1 from his Experiments 2..5) were collected using N=10 lines as stimuli, and include accurate measurements of response time for each decision. Stewart et al.'s data, taken from their Experiment 1, were collected using equally-loud tones of differing frequencies as stimuli. This experiment did not include measurement of response times, but did have two important stimulus manipulations: the number of tones in the stimulus set was varied *(*N=6, 8, or 10), and the spacing of the tones was either "wide" (each tone was 12% higher in frequency than the one below) or "narrow" (6% spacing).

In both Lacouture's (1997) and Stewart et al.'s (2005) experiments there were many participants, each of whom contributed around 800 data points each. Our modeling of each data set stressed parsimony, with each fit using a fixed set of parameters (different for the two data sets) to generate all model predictions for data averaged across participants. This approach provides stringent model tests. For example, in Stewart's data, set size and stimulus spacing effects must be generated by the model's architecture, rather than by different parameter settings. Our fits to these data sets include the effects of stimulus magnitude, set size, and stimulus spacing on response measures including choice probabilities, sensitivity (d') and full RT distributions. We also examine sequential effects on both choices and RT. This is the

first time that an absolute identification model has had the explanatory range to be tested so comprehensively.

Parameter estimates for the two data sets are shown in Table 3. For Stewart et al.'s (2005) experiment no parameters were changed to model the different set size conditions or the different stimulus spacings. Instead, the predictions of SAMBA change with set size and stimulus spacing simply because the stimulus representations reflect the physical elements of the experimental design. Since Stewart et al.'s data did not contain RT measurements, the parameters of the ballistic accumulator stage of SAMBA were not estimated (except for the assimilation parameter, *D*, which can be estimated from choice data). As a result, there were seven model parameters for Stewart et al.'s data and eleven for Lacouture's (1997) data. We do not claim that these parameter estimates are optimal. No formal optimization was performed and only a relatively course grid of parameter estimates were tried. Better fitting solutions likely exist, but the values displayed in Table 3 provide a sufficiently good account.

Data from Lacouture (1997)

Several phenomena observed in Lacouture's (1997) data have already been discussed in the earlier section on "Critical Tests" because they featured in our attempts to distinguish absolute from relative models of absolute identification. These phenomena included sequential effects on response accuracy, decision sensitivity, and mean RT, as well as the effect of response repetition on accuracy and mean RT. Five more phenomena are discussed in this section, including: the effects of stimulus magnitude on response accuracy, mean RT and RT distributions, and assimilation and contrast. For all graphs, we display both data (using symbols) and model fits (using dashed lines).

Figure 12 shows response accuracy as a function of stimulus magnitude. This is the standard bow effect plot, and SAMBA accounts for the bow shape and the asymmetry evident in the data, except for the largest stimulus. SAMBA posits that asymmetry is due to unequal placing of the stimulus anchors, with the lower anchor (*L*) placed much closer to the smallest stimulus than the upper anchor (*U*) is to the largest stimulus. When fitting Lacouture's (1997) data, we estimated the lower anchor for the selective attention process at 91 pixels – just one pixel smaller than the smallest stimulus – and the upper anchor at 420 pixels – 100 pixels larger than the largest stimulus. The asymmetry in the anchor placements (both in pixel and log units) accounts for the strong asymmetry in the data: responses to smaller lines were faster and more accurate than responses to larger lines. SAMBA is the first absolute identification model that accounts for asymmetry, which is often observed in data but has been almost uniformly ignored. However, SAMBA's account is still quite constrained and cannot accommodate one unusual aspect of the asymmetry in Lacouture's data – namely, responses to line #10 are both very accurate and *extremely* fast relative to other responses.

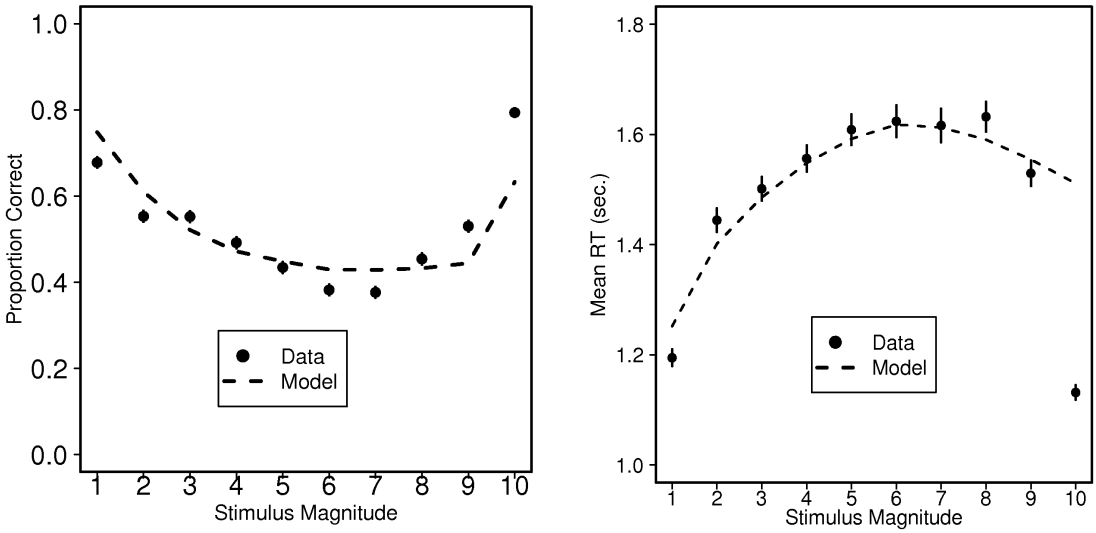

Figure 12. Accuracy and mean RT as functions of the ordinal stimulus magnitude (x-axis) for data from Lacouture (1997) and SAMBA model fits. The vertical lines on

each data symbol show normal standard errors based on the SD of each point over participants.

Figure 12 shows that SAMBA provides a good account of differences in mean RT for the different stimulus magnitudes in Lacouture's (1997) data. Figure 13 extends this analysis in two ways – using full RT distributions, rather than just means, and showing both correct and incorrect responses. Figure 13 uses a similar format to Figure 5 (for Kent & Lamberts', 2005, data). The left panel shows correct responses (7324 data points) and the right panel shows data for errors of +1 response (2072 data points) and -1 response (3407 data points). Data for +1 and -1 errors were too noisy to present separately so we have averaged them together, after flipping response magnitudes 1..10 for the -1 errors.

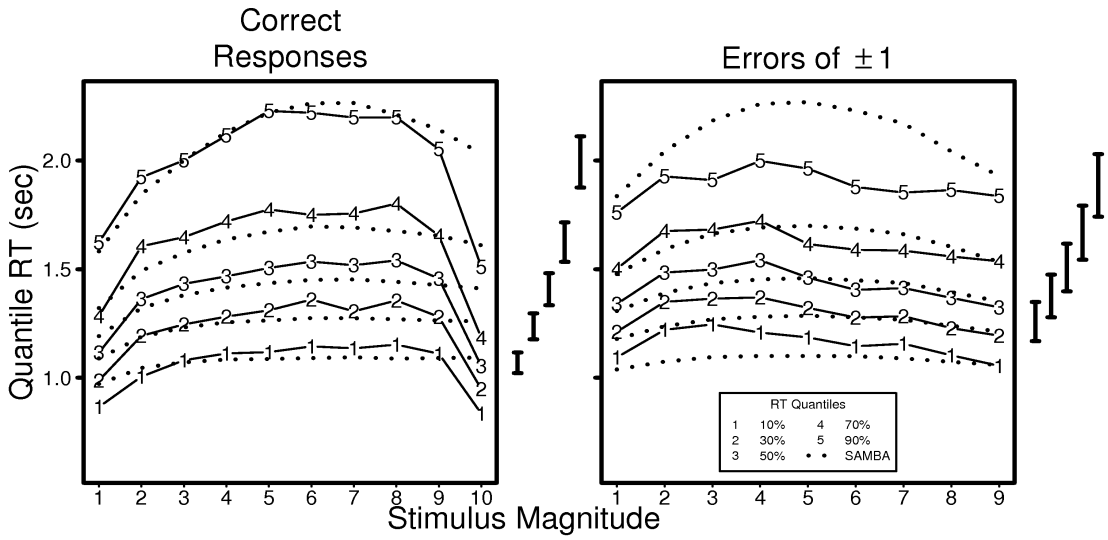

Figure 13. Response time distributions from Lacouture (1997), for correct responses (left panel) and for errors of +1 or -1 response category (right panel). The lines numbered 1..5 in each panel show 10%, 30%, 50%, 70% and 90% quantiles estimated from the data, for each of the ten stimulus magnitudes (x-axis). Dashed lines show predicted quantiles from SAMBA. Bars to the left of each plot show average standard errors for each quantile, calculated by bootstrap from the raw data (see e.g., Ratcliff, Gomez & McKoon, 2004). Note that standard errors are larger for the longer quantiles, and for the incorrect response data.

SAMBA accurately predicts RT distributions for the correct responses apart

from overestimating RT for the largest stimulus (as discussed regarding mean RT). The model captures the shape of these RT distributions, as illustrated by the relative spacing of the quantiles, and the variance of the distributions, as illustrated by the absolute spacing of the quantiles. SAMBA also captures the changes in shape and variability across the range of stimulus magnitudes. For the incorrect responses, SAMBA provides a qualitatively reasonable account, but shows systematic quantitative errors. For example, SAMBA accurately matches the entire RT distributions for errors on the smallest and largest stimuli and for all stimuli for the middle quantiles. However, it predicts too large a spread in the tails (i.e., 10% and 90 % quantiles) of incorrect RT distributions for the middle range of stimuli.

Figure 14 shows assimilation and contrast effects in Lacouture's (1997) data, along with model fits. For this figure, we use the same layout as Figure 4 (see also Ward & Lockhead, 1970, and Stewart et al., 2005), and show SAMBA's predictions using solid lines, rather than the usual dashes. The plot shows response biases, measured as average errors. For example, if stimulus #5 is presented, the error will be zero if the participant responds correctly with response #5, but will be +2 if the participant gives response #7, and -1 if the participant gives response #4. The graph shows average errors on Trial N+X, for X=1,2,…,8, conditional on the magnitude of the stimulus presented on Trial N. Filled symbols show average error when the preceding stimulus was small, open symbols show the error when the preceding stimulus was large.

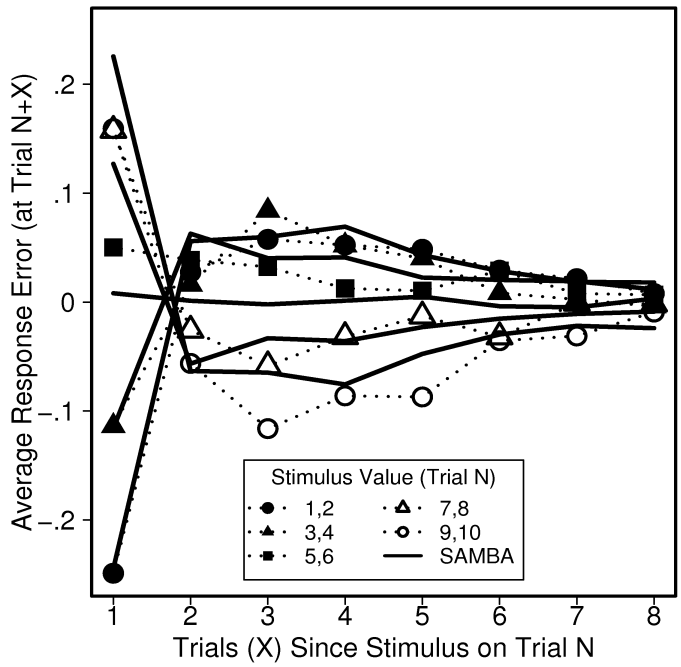

Figure 14. Average error as a function of preceding stimulus (different symbols – see legend) and number of trials since that stimulus (x-axis) from Lacouture (1997). Assimilation occurs to stimuli presented one trial previously: Errors are positive at x=1 for large previous stimuli (filled symbols), and negative for small previous stimuli (open symbols). At longer lags (x=2..5) contrast is observed, i.e., the opposite pattern.

The characteristic pattern of assimilation at lag=1 and contrast at lag>1 was observed in these data. That is, when a large stimulus was presented just one trial previously errors were positive, and vice versa for small stimuli. When a large stimulus was presented several trials previously (X=2..8), average error was negative, and vice versa for small stimuli. SAMBA successfully fits the qualitative pattern of assimilation at short lags and contrast at longer lags. In Lacouture's (1997) data, contrast is strongest at lag=3, while in many other data sets (including Ward & Lockhead, 1970, and Holland & Lockhead, 1968) contrast is strongest at lag=2 and decays monotonically thereafter. To fit the peak at lag=3 observed in Lacouture's data, we chose K=4, so the re-direction of attention in the selective attention stage lasts for four trials.

Data from Stewart et al.'s (2005) Experiment 1

Stewart et al.'s (2005) Experiment 1 provides a valuable resource as it allows a direct comparison between the goodness-of-fit to choices of two competing models. Stewart et al. provide graphs of many observed data patterns, and simultaneously provide predictions from their RJM. By fitting SAMBA to the same data patterns, we can compare the two models, although we are unable to compare them on aspects of absolute identification that the RJM does not cover (such as RT). Comparing goodness-of-fit is complicated by the varied nature of the phenomena we examine, including response probabilities, average biases, and sensitivity measures.

In the absence of an agreed statistical model, we can do no better than compare how closely the predicted values from each model (RJM and SAMBA) match the observed values. To do this, we use root mean squared error (RMSE). The magnitude of RMSE is, of course, without statistical meaning. Nevertheless, the relative size of RMSE for the two models is informative. The most intractable problem with this approach is model complexity – it is possible for a false model to more accurately fit observed data than a true model, if the false model is the more complex of the two. Stewart et al. used eight free parameters to fit the RJM to the data from their Experiment 1, whereas the fits of SAMBA have only seven, suggesting that RJM is more complex. Although the number of parameters does not give a complete measure of model complexity, this difference indicates that complexity is unlikely to explain cases where SAMBA provides an equal or better fit than RJM[12]. To foreshadow our results, we found that both SAMBA and the RJM provide quite good fits to the data, which does

12 We calculated RMSE values by reading data from the published graphs in Stewart et al. (2005), and separately comparing the predictions of the RJM to the data from the wide and narrow stimulus spacing conditions. We took the same approach for calculating RMSE values from SAMBA's predictions. Slight differences in the RMSE values may arise if a different approach were used (e.g., if one compared with the average of the wide and narrow data conditions, or if one differentially weighted the different set sizes).

not include response times.

The stimuli for Stewart et al.'s (2005) Experiment 1 were equal-loudness tones of different frequencies. In the "narrow" condition, the tones were separated by 6% (e.g., the lowest tone was 768.7Hz, the next tone was 6% higher at 814.8Hz, and so on), while in the "wide" condition tones were separated by 12%. Set size was manipulated independently of stimulus spacing, by using either all N=10 tones, or just the central N=8 or N=6. Both set size and stimulus spacing were manipulated between subjects. The data do not include response times, so several of SAMBA's parameters can be omitted. Below, we compare SAMBA's predictions with predictions from Stewart et al.'s RJM. As presented in Stewart et al. (2005), the RJM does not accommodate asymmetry in the data (i.e., more accurate responses to very small than very large stimuli, or vice versa), although SAMBA handles such data naturally, as in our fits to Lacouture's (1997) experiment. However, to keep the flexibility of SAMBA similar to that of the RJM, we constrained our fits to be symmetric (i.e., we used L=U).

Stewart et al. (2005) also did not model the effects of the "wide" and "narrow" stimulus conditions separately, even though the RJM has a perceptual noise parameter that might be able to account for this manipulation. Certainly, SAMBA's perceptual noise process (σ_P) can account for differences between the narrow and wide conditions, and so we estimated it fits to this data. We could have ignored the differences in the data from the wide and narrow spacing conditions, and therefore dropped the perceptual noise parameter from our fits. This would have had the advantage of allowing SAMBA and RJM's predictions to be compared more easily, but the disadvantage of forcing a greater disparity in model complexity: without the stimulus noise parameter, SAMBA would use just six parameters compared with RJM's eight.

We assumed that frequency is represented on a simple logarithmic scale, and that the magnitude of psychophysical noise was σ_P=1.9% of stimulus magnitude. The psychophysical noise was more influential in the narrow than the wide condition, because the stimuli were only 6% apart in the narrow condition, so the 1.9% standard deviation more often resulted in confusion between adjacent stimuli. For parsimony, we also treated the narrow and wide conditions equally: the lower anchor was always 10% of the total stimulus range lower than the smallest stimulus and the upper anchor was always 10% of the stimulus range above the largest stimulus. To provide even greater model constraint, we assumed that this anchor placement was constant across set sizes (*N*=6, 8 and 10) as well as the stimulus spacing conditions. For example, in physical terms for the narrow stimulus spacing condition with set size *N*=10, the lower anchor was set at *L*=773Hz and the upper anchor at *U*=1369Hz. Finally, we also assumed that participants in Stewart et al.'s (2005) experiment *always* used the locally relative judgment mechanism in SAMBA. This assumption follows from the extraordinarily large accuracy bonus previously observed for repeat and near-repeat stimuli in these data.

In what follows, we illustrate the effects of set size and stimulus spacing on response frequency, accuracy, sensitivity (*d*') and average bias (contrast and assimilation). Note that we have already discussed the effects of stimulus repetition and near-repetition on accuracy, in the "Critical Tests" section. All of the following different data fits, and the fits to repetition effects already reported, were modeled in SAMBA using a single set of parameters (shown in Table 3).

The first new analysis examines response frequency. Figure 15 shows how often each response was given, separately for the wide and narrow conditions, and for the set

sizes *N*=6, 8 and 10. There is little difference in the shape of the response probability curves between the wide and narrow conditions. SAMBA captures the changes in response probability with set size, and the lack of change with stimulus spacing. Both SAMBA and RJM (see Stewart et al.'s figure 20) fit these data very well, and RJM performs slightly better, with an RMSE of 0.0062 compared to 0.0079 for SAMBA.

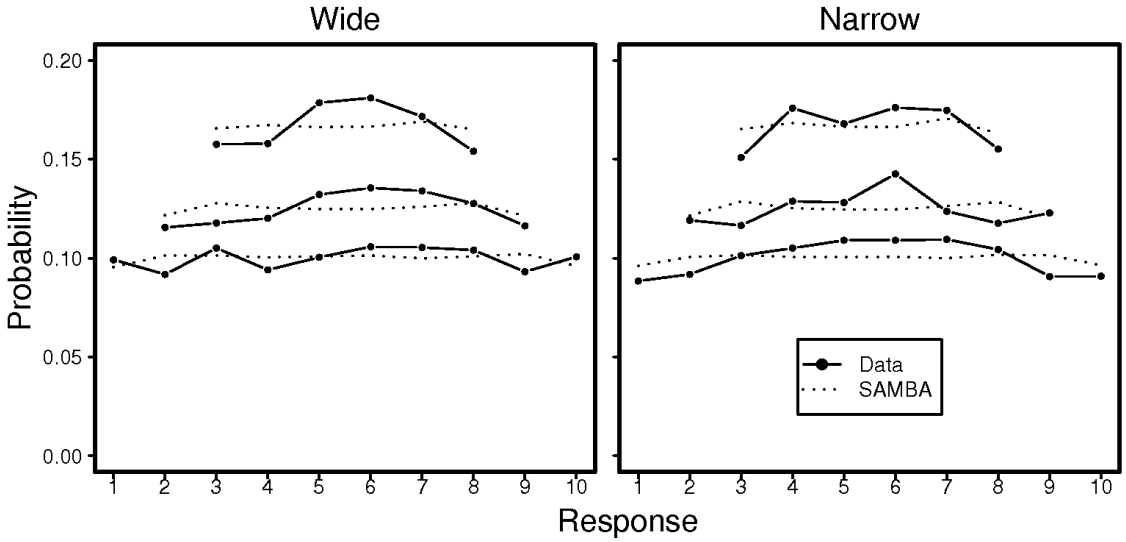

Figure 15. The probability of using each response, for narrow and wide conditions, and set size *N*=6, 8 and 10 (top to bottom, respectively). Central responses were used more frequently than extreme responses.

Figure 16 shows response accuracy for wide and narrow conditions, separately for set sizes *N*=6, 8 and 10. Note that the bow effect is apparent at every set size, and increases in depth with increasing set size. Performance was also worse for narrow than widely spaced stimuli. SAMBA accommodates the bow effect, the changes in accuracy with set size, and even the improved accuracy for wide over narrow spaced stimuli. The only apparent misfit is to the central four stimuli from the set size *N*=6 data under the wide spacing condition, where SAMBA underestimates performance. Quantitatively, SAMBA fits the data better than the RJM (RMSEs of .035 and .05, respectively).

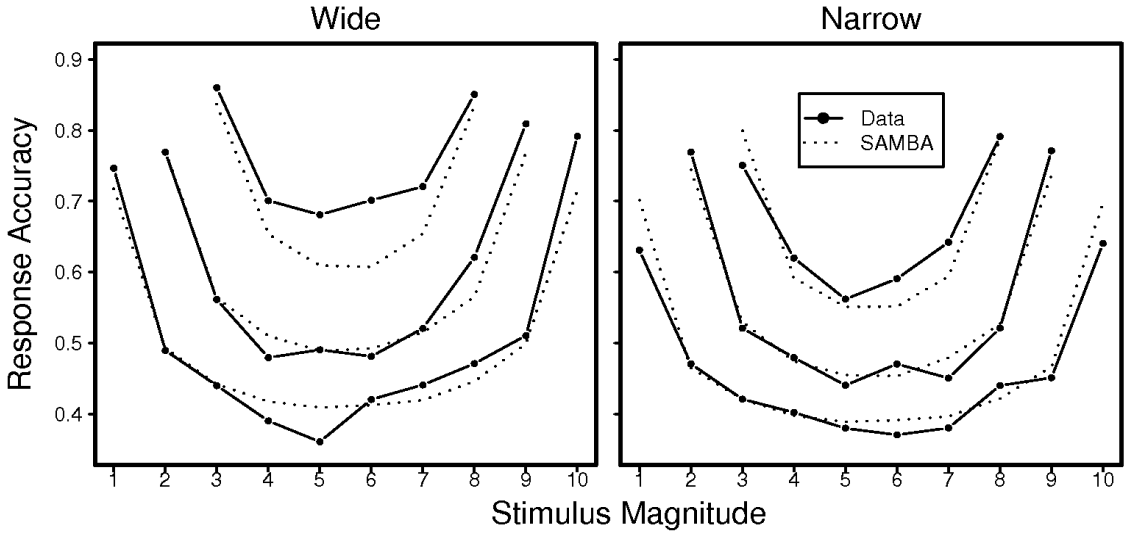

Figure 16. Response accuracy as a function of set size and stimulus spacing and set sizes *N*=6, 8 and 10 (top to bottom, respectively). The dotted lines show the predictions of SAMBA, lines with filled circles are data from Stewart et al. (2005).

Figure 17 shows the accuracy data transformed to response sensitivity (*d*'). As for accuracy, SAMBA under-predicted performance on the smallest set size, and this effect is greatly exaggerated in the *d'* data due to the nonlinear stretching effect of the inverse normal transformation for probabilities close to one. Indeed, the depth of the *d'* bow for the data for set size *N*=6 (and, to a lesser extent *N*=8) is remarkably large in Figure 17, and not well-fit by our model. The better account of stimulus spacing effects and the *d*' bow result in SAMBA fitting the data somewhat better than RJM (RMSEs of 0.39 and 0.43 respectively).

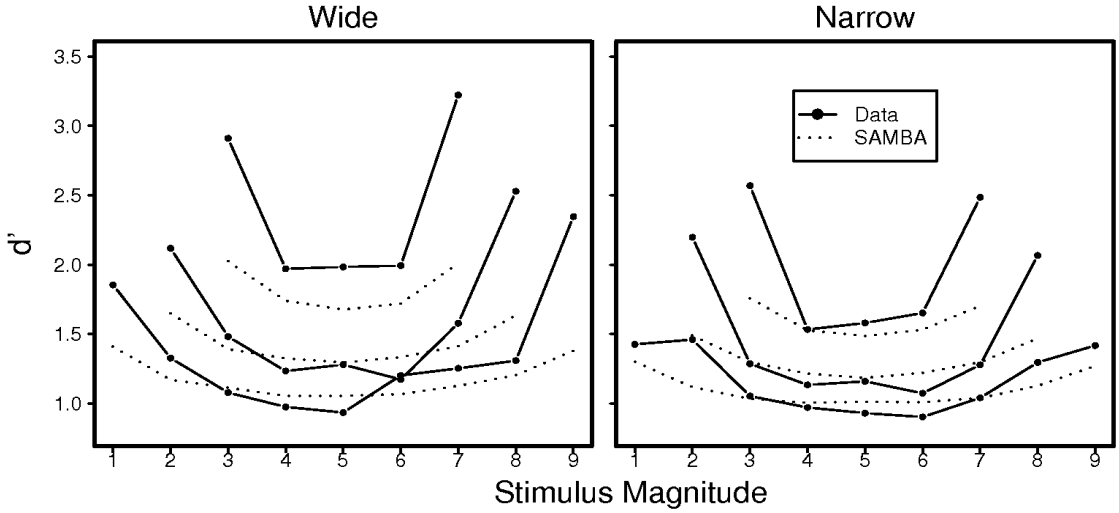

Figure 17. Response sensitivity (d') as a function of set size and stimulus spacing and set sizes *N*=6, 8 and 10 (top to bottom, respectively). The dotted lines show the predictions of SAMBA, lines with filled circles are data from Stewart et al. (2005).

The response accuracy data (Figure 16) describe only correct responses, while the sensitivity data (Figure 17) use all responses, but collapse them into a single summary statistic. Figure 18 shows the probability of all responses to all stimuli – that is, full confusion matrices – along with SAMBA's predicted values. As observed in Figures 15 and 16, the probability of a correct response (the highest peak for each line in each graph) decreases with increasing set size, and in the narrow compared to the wide stimulus spacing conditions. SAMBA captures the complete distribution of error and correct responses, including the effects of stimulus magnitude and narrow vs. wide stimulus spacing. However, for set size *N*=10 the model over-predicts the proportion of +/-1 responses to stimuli #2-#5, and for set size *N*=6 it under-predicts overall accuracy in the wide spacing condition. These same effects were apparent in the accuracy graph (Figure 16). RMSE for the confusion matrices favors RJM very slightly over SAMBA (0.038 vs. 0.041).

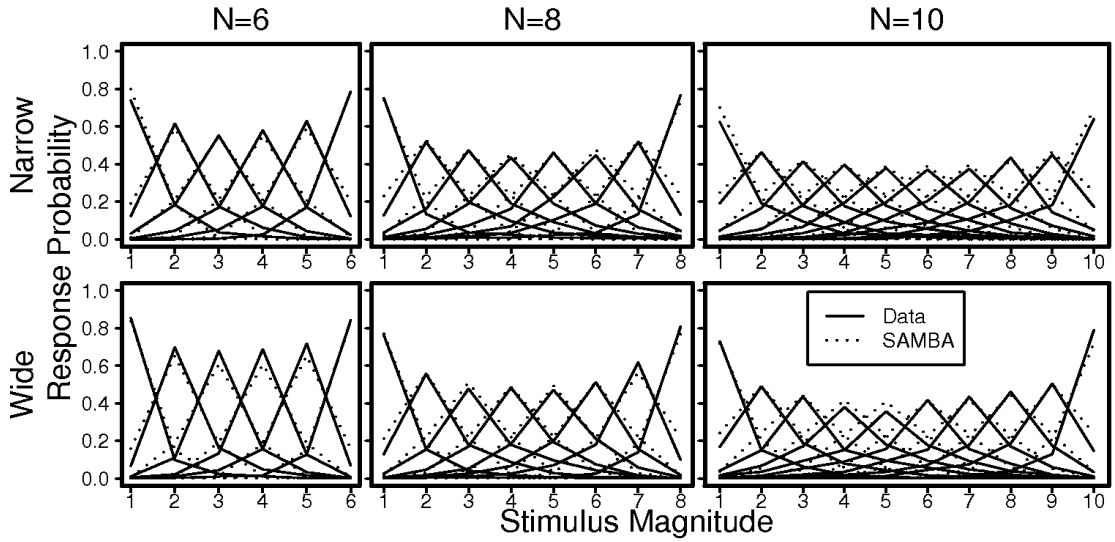

Figure 18. Response matrices for the three different set sizes and the wide/narrow stimulus spacing conditions from Stewart et al. (2005). Each line represents the probability of a particular response, conditioned on the various stimuli.

Finally, we turn to the sequential effects of assimilation and contrast. Figure 19 shows assimilation and contrast effects in Stewart et al.'s (2005) data, using the format introduced for Figures 4 and 14. Each graph plots the average error on the current trial as a function of the size of a preceding stimulus (separate lines). The x-axis shows the number of trials that have elapsed since that preceding stimulus was presented ("lag"). As usual, there is an assimilation effect at lag=1, with responses being biased *towards* the previous stimulus: if the previous stimulus was large, average error is positive, and vice versa. At longer lags (2+) contrast is observed, with responses biased *away* from the previously-seen stimuli. Most trends in the sequential data are captured well. SAMBA predicts assimilation at lag=1 followed by contrast at the longer lags. It also predicts larger effects in the narrow than the wide stimulus spacing, and larger effects with increasing set size. SAMBA's predictions match the data just as well as the RJM's (both RMSEs of 0.044).

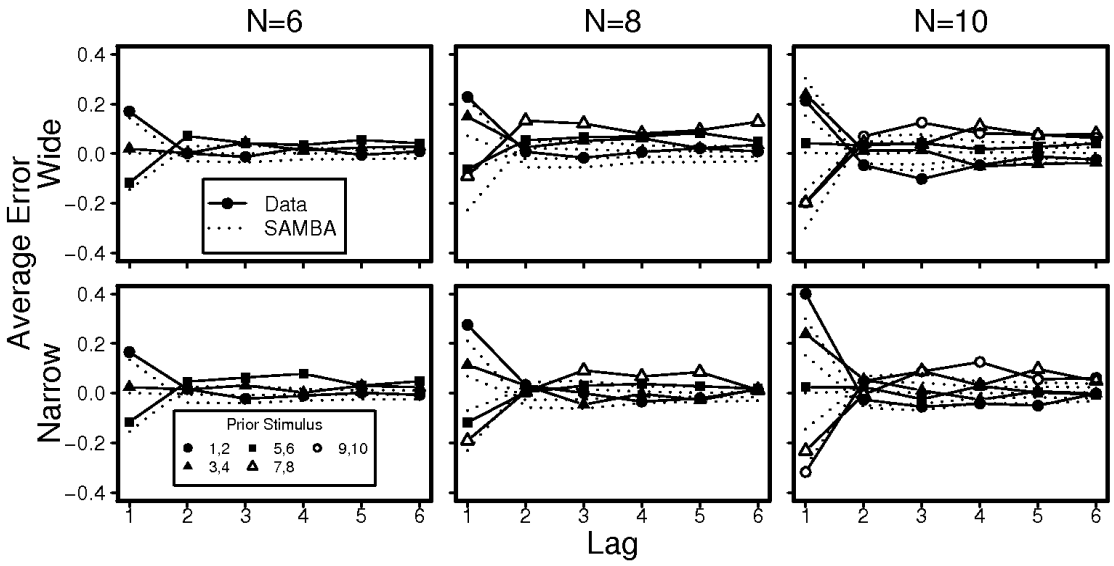

Figure 19. Assimilation and contrast effects in data from Stewart et al.'s (2005) data, separately for the three set sizes *N*=6, 8 and 10 and for wide and narrow stimulus spacing conditions. The lines in each plot show different magnitudes for preceding stimuli (see legends).

General Discussion

In interpreting data from absolute identification a distinction has been proposed between local and global phenomena. Local phenomena are those with short temporal duration, particularly effects of recent stimuli and responses on current decisions. Global phenomena are relatively stable over time, such as the effects of stimulus ranges and set sizes. Some theoretical accounts have focused on local processes (Holland & Lockhead, 1968; Lockhead & King, 1983; Laming, 1968, 1984), while others have focused on global processes (e.g., Braida et al, 1984; Marley & Cook, 1984; Lacouture & Marley, 1991, 1995, 2004). Several more recent models have incorporated both local and global processes (e.g., Nosofsky, 1997; Nosofsky & Palmeri, 1997; Petrov & Anderson, 2005). Recently, Stewart et al. (2005) took a more extreme position against global processing, affirming that absolute identification is based *only* on local processes of a particular type, namely, relative judgment, with no absolute or global processing

whatsoever (see also Laming, 1984; Lockhead, 2004).

We have described an extension of global, restricted-capacity models, developed in various papers by Cook, Karpiuk, Lacouture and Marley, to include local processes. SAMBA provides a comprehensive account of absolute identification because it predicts not only choices, but also the time taken to make them. Also, SAMBA includes an account of multiple sources of variability affecting decision processes, including sequential effects, and its predictions for response time (RT) are not restricted to just mean RT bow effects. Our analyses of data from Kent and Lamberts (2005), Lacouture (1997) and Lacouture and Marley (2004) show that SAMBA provides an accurate account of the entire distribution of response times as a function of stimulus magnitude. The model also provides an accurate account of asymmetries and sequential effects on RT distributions and response choices. The choice and RT effects are predicted by the same set of parameters, providing a more stringent test than fitting choice or RT data alone.

Although SAMBA provides a close quantitative fit to dozens of phenomena in absolute identification, there are some places where it fails to fit the data, underlining the point that SAMBA is sufficiently constrained in its predictions to be falsifiable. One of the failures, over prediction of Stewart et al.'s (2005) false feedback effect, was relatively small and SAMBA still performed better than the only other model tested against this effect. Given that a natural extension of SAMBA's learning mechanism (*partial* correction for feedback) predicts an appropriately smaller effect, this failure is not troubling. Two other small failures were associated with under prediction of bow effects involving the very largest stimulus for some conditions in Luce et al.'s (1982) and Lacouture's (1997) data. The cause of such failures is unclear, although the

limitation to one stimulus in each case suggests some idiosyncratic factor may be in play, with increased performance perhaps related to external referents that were most salient for the largest stimuli.

Two larger and more systematic failures concern responses to repeated stimuli in Stewart et al.'s (2005) Experiment 1 (see Figure 11) and the distribution of RT for incorrect responses in Lacouture's (1997) data (see Figure 13). The latter failure is striking because our RT distribution fits for correct responses are of almost the same quality as achieved by the leading accounts of two-choice RT (e.g., Ratcliff & Smith, 2004). One explanation for our poorer fit to incorrect responses is that two-choice RT models are typically not constrained to fit as wide a range of phenomena as were our fits to Lacouture's data (e.g., sequential effects). A second explanation concerns our use of quantiles averaged over participants, necessitated by small sample sizes per participant. Averaging can spread quantiles when individual differences are present, which might particularly be a factor for incorrect responses as participants can vary markedly in their speed-accuracy tradeoff setting (see, e.g., Brown & Heathcote, 2003). Different settings are associated with systematic changes in the speed of error responses. In support of this explanation, we note that our fits to the error RT distributions for the individual participant in Lacouture and Marley's (2004) Experiment 2 were quite good.

SAMBA performs well in fitting the accuracy and RT advantages for repeated stimuli in many data sets (e.g., Lacouture, 1997, and Luce et al., 1982). In Stewart et al.'s (2005) data, however, repetitions have a much larger advantage than SAMBA can predict. We noted earlier that SAMBA has no special mechanism for repeated stimuli. Given the difference between Stewart et al.'s results and others in the literature, this failure may not indicate a direct falsification. A possible explanation is that the stimuli

in Stewart et al.'s experiment differed in tone frequency, whereas Lacouture used line lengths and Luce et al. used tone intensity (loudness). It is possible that stimulus repetitions are more easily detected in frequency than in other continua; this suggestion is compatible with the existence of "critical bands" in the perception of frequency (Green & Swets, 1966). Further experimentation is required to investigate the effects of response repetitions on accuracy, in a range of different stimulus modalities.

SAMBA accounts for all standard sequence effects, global effects, and the effect of misleading feedback *without* any relative judgment process. Short-term memory in the decision stage predicts repetition and assimilation effects, whereas the short-term memory in the selective attention stage predicts contrast effects. A combination of these processes, along with restrictions on the available responses, explains the effects of manipulating the stimulus sequences found by Luce et al. (1982). Hence, SAMBA demonstrates that none of these effects necessarily implicate relative judgment. There is just one data pattern that *does* implicate relative judgment in SAMBA: the accuracy bonus sometimes observed for repeat and near-repeat stimuli. Although just a small part of the data, this effect has proven theoretically important, and not just for SAMBA. The same analyses proved critical for the RJM, prompting Stewart (2007) to modify the RJM to allow the stimulus two trials back to be used as the basis for relative judgment on some trials, instead of the memory for the stimulus one trial back. This modification is one step towards an extension where "any previous stimulus could be used [which would] introduce long-term representation of magnitudes into the model" (Stewart, 2007, p. 536).

Future Developments

There are several aspects of SAMBA that warrant further development. That

development will be aided by the clear and testable predictions that arise from several of SAMBA's assumptions about the cause and locus of certain phenomena. We will not provide an exhaustive discussion on these topics here, but instead restrict ourselves to a three aspects requiring further development, and to three novel predictions.

1. The selective attention mechanism. Like its precursors, SAMBA is based on the notion of global processing – the central tenet of the model is that stimuli are judged relative to a context that changes slowly over time. The precise nature of this selective attention context requires further study. For example, more experiments are required to determine which experimental factors cause participants to change their anchor positions, and how quickly these changes occur. Other experiments are required to investigate what factors affect the size, duration, and peak timing of contrast effects; these results will illuminate what factors affect the Poisson process. For example, it is an open question whether the dynamic aspects of that process operate in units of real time or trial time. If contrast effects operate on a time (rather than trial) basis, the peak magnitude for contrast effects could be manipulated by changing the response-to-stimulus delay interval (RSI). If the RSI is made very short, the number of trials before peak contrast is reached will increase, shifting the location of peak contrast further to the right (i.e., lag=3 or lag=4). We cannot tell from the published details the precise RSI values used in previous data sets, so we are currently pursuing new experiments in which the intertrial interval is manipulated, either within or across experiments.

2. Learning the stimulus representations. The mechanics of the mapping stage are completely determined by long term representations of average stimulus magnitude estimates. For example, if there are three evenly spaced stimuli, with anchors fixed at the lower and upper magnitudes, the average magnitude estimates are {0,½,1}.

Throughout the paper, except when addressing misleading feedback, we have assumed that these values are unvarying and accurate estimates of the true expected values. This assumption is clearly too strong, suggesting the need to develop a model of how the magnitude estimates are learned and maintained. We developed the beginnings of this model when addressing the false-feedback data from Stewart et al.'s (2005) Experiment 2. For those data, we suggested a mechanism that adjusts magnitude estimates to align with feedback. This mechanism is similar to one of two mechanisms proposed by Treisman and Williams (1984). In future work, we will examine the addition of Treisman and Williams' second mechanism (an assimilative process) into SAMBA. Together, these two processes provide a testable system for accommodating the effects of correct feedback, false feedback, and the absence of feedback (see also Mori & Ward's, 1995, discussion of a similar use of Treisman & Williams' mechanisms). Our preliminary investigations have offered the intriguing prospect that the use of Treisman and Williams' adjustment (learning) mechanisms may eliminate altogether the need for the relative judgment process in SAMBA. This suggests that those effects previously considered indicative of local relative judgment (particularly the accuracy advantage for repeat and near-repeat stimuli), may be alternatively considered as evidence of a learning process that maintains and adjusts long term referents for magnitude estimates.

3. The causes of incorrect responses. We assume that the mapping stage is error free. The decision stage can produce errors, due to the influence of responses on previous trials on the starting points of the ballistic accumulators, but our estimated parameters suggest that this effect is smaller than that of the selective attention stage. For example, in fits to Lacouture's (1997) data, only 23% of incipient choices were changed by the action of the decision phase (i.e., in 23% of cases, the final response generated by the

decision stage was different from the response corresponding to the maximum output of the selective attention and mapping stages). These assumptions stand in contrast to several prior models, which have attributed errors more directly to processing *after* a magnitude estimate is produced, such as in a decision mechanism (e.g., Kent & Lamberts, 2005; Lacouture & Marley, 1995, 2004; Nosofsky & Palmeri, 1997; Treisman & Williams, 1984).

Predictions

The current version of SAMBA leads to some novel empirical predictions, of which we present three.

1. Unequal stimulus presentation frequency (a test of the selective attention stage). The contrast mechanism associated with the selective attention stage makes the, perhaps surprising, prediction that if stimuli within a sub-interval of the range are presented more frequently, then those stimuli will eventually be identified more accurately at the expense of the remaining stimuli outside the sub-interval. We are aware of two experiments that tested the above prediction, one of which found changes in response bias, but not in *d'* (Chase, Bugnacki, Braida, & Durlach, 1983) whereas the other found a small, but significant, effect on *d'* (Nosofsky, 1983). We are currently working on sharper experimental tests of this prediction.

2. Non-uniform stimulus spacing (a test of the mapping stage). The mapping stage of SAMBA predicts that stimuli that are spaced closer together, relative to other stimuli, will be more poorly identified. This prediction sounds trivial, but is not because it holds even when the closely-spaced stimuli are still well above threshold in comparative judgment tasks. We have interpreted Lockhead and Hinson's (1986) data as supporting this prediction, although other interpretations have been advanced, e.g., the use of

appropriately selected cutpoints on a suitably constructed response variable (Stewart et al., 2005). Data from Lacouture (1977) illustrate that stimuli which are more widely spaced, relative to other stimuli, are more accurately identified.

3. The effect of inter-trial interval on assimilation (a test of the decision stage). Our assumption that assimilation is caused by passive decay in the ballistic accumulators leads to the prediction that assimilation effects will be smaller when the inter-trial interval is larger. To our knowledge, relevant absolute identification experiments have not been performed (but see DeCarlo, 1992, for related supporting theoretical and empirical work in magnitude estimation). Many absolute identification models would be able to accommodate this prediction, for example by simply adjusting parameter values associated with assimilation. However, SAMBA is currently the only model of absolute identification that makes a quantitative prediction for this effect as an *a priori* consequence its architecture.

These three predictions follow from the current version of the SAMBA model, in which we have assumed that the distribution of selective attention is mostly uncontrolled by the participant. Instead, accumulators are incremented randomly, except for the re-direction that results in contrast. With these assumptions, SAMBA makes starkly different predictions for the effect of unequal stimulus presentations, in point 1 above, and for non-uniform stimulus spacing, in point 2. That is, unequal stimulus presentations can improve discrimination for the more frequently presented stimuli, whereas non-uniform spacing is likely to decrease discrimination for the stimuli that have a decreased relative spacing in the non-uniform case. However, if attention were under sufficient strategic control this differential prediction for the two tasks would be reduced or eliminated. Therefore, stronger tests of the degree to which selective

attention is under strategic control are required, and are being pursued.

Concluding Remarks

In closing we note that our results have implications beyond absolute identification. Although strong sequential effects are present in choice and RT data from other paradigms (e.g., Gilden, 1997, 2001; Gilden, Thorton & Mallon, 1995; Kelly, Heathcote, Heath & Longstaff, 2001; Laming, 1968; Wagenmakers, Farrell & Ratcliff, 2004, 2005), these sequential effects are not accommodated by existing choice RT models (but see Wagenmakers et al., 2004, for some possible directions). We have suggested a mechanism for assimilation effects that could be as easily implemented in any number of sequential sampling models of choice RT (such as those in Ratcliff & Smith, 2004) as it can be in Brown and Heathcote's (2005) ballistic accumulation model – namely, by making the starting points of accumulation dependent on the response made on the previous trial. This mechanism moves choice RT models towards an *explanation* rather than an *assumption* of variability (i.e., start points are usually assumed to be variable with no testable explanation of the source of that variability).

We have also suggested that longer-term sequential effects (e.g., contrast) can arise as a consequence of changes over trials in the inputs to the decision stage, and have provided a mechanism that predicts the magnitudes and variability of these inputs, at least in the context of absolute identification. At present, most choice RT models (see Smith, 1995; Smith, Ratcliff, & Wolfgang, 2004, for exceptions) simply assume the appropriate magnitudes and variability of inputs to fit data, with little motivation for how these inputs arise. The latter approach, although perhaps viable for two-choice data and a limited number of stimulus conditions, breaks down as the number of stimuli and responses increase, because of parameter proliferation. The SAMBA absolute

identification model, in contrast, requires no extra parameters to accommodate an increase in the number of stimuli and responses. Similar approaches to modeling inputs to decision processes in other paradigms would greatly simplify and expand the power of models of choice response time.

Acknowledgments

This research has been supported by: Australian Research Council Linkage International Project LX0348125 to Vickers, Marley, Heathcote, Smith and Lee; Natural Science and Engineering Research Council Discovery Grant 8124-98 to the Marley; and Australian Research Council Discovery Project DP0881244 to Brown and Heathcote. Tragically, Doug Vickers passed away at the beginning of this project; we dedicate this paper to his memory.

References

Berliner, J. E. (1973). Intensity perception in audition. *Unpublished PhD thesis,* Department of Electrical Engineering, MIT.

Berliner, J. E., Durlach, N. I., & Braida, L. D. (1977). Intensity perception. VII. Further data on roving-level discrimination and the resolution and bias edge effects. *Journal of the Acoustical Society of America, 61*(6), 1577-1585.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception. II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America, 51,* 483-502.

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity Perception. XIII. Perceptual anchor model of context-coding. *Journal of the Acoustical Society of America, 76,* 722-731.

Brown, S., & Heathcote, A. (2003) Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers, 35*, 11-21

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112*(1), 117-128.

Brown, S.D., & Heathcote, A.J. (submitted). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology.*

Brown, S. & Heathcote, A. (submitted). The Ballistic Accumulator: Efficient computation for any number of choices. *Behaviour Research Methods.*

Brown, S., Marley, A. A. J., & Lacouture, Y. (2007). Is absolute identification always relative? *Psychological Review,* 114, 533-538.

Busemeyer, J. R. & Townsend, J. T. (1992). Fundamental derivations for decision field theory. *Mathematical Social Sciences,* 23, 255-282.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision field theory: A dynamic cognition

approach to decision theory. *Psychological Review,* 100, 432-459.

Chase, S., Bugnacki, P., Braida, L. D., & Durlach, N. I. (1983). Intensity perception. XII. Effect of presentation probability on absolute identification. *Journal of the Acoustical Society of America, 73,* 279-284.

DeCarlo, L. T. (1992). Intertrial interval and sequential effects in magnitude scaling. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 1080-1088.

DeCarlo, L. T. (1994). A dynamic theory of proportional judgment: Context and judgment of length, heaviness, and roughness. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 372-381.

DeCarlo, L. T., & Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General, 119,* 375-396.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46,* 372-383.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46,* 373-380.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decision making. *Psychological Science*, 8, 296-301.

Gilden, D. L. (2001). Cognitive Emissions of 1/f Noise. *Psychological Review, 108,* 33-56.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f Noise in Human Cognition. *Science*, *267*, 1837-1839.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception and Psychophysics, 3,* 409-414.

Karpiuk, P., Jr., Lacouture, Y., & Marley, A. A. J. (1997). A limited capacity, wave equality, random walk model of absolute identification. In A. A. J. Marley (Ed.), *Choice, decision and measurement: Essays in honour of R. Duncan Luce* (pp. 279–299). Mahwah, NJ: Erlbaum.

Kelly, A., Heathcote, A., Heath, R. A. & Longstaff, M. (2001). Response time dynamics: Evidence for linear and low-dimensional nonlinear structure in human choice sequences. *Quarterly Journal of Experimental Psychology, 54*, 805-840.

Kent, C., & Lamberts, L. (2005). An exemplar account of the bow and set size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 289-305.

Krantz, D.H. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology, 9,* 168-199.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60,* 121-133.

Lacouture, Y., & Marley, A. A. J. (1991). A connectionist model of choice and reaction time in absolute identification. *Connection Science, 3,* 401-433.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39,* 383-395.

Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception & Psychophysics, 66, 1206-1226.*

Laming, D.R.J. (1968). *Information theory of choice-reaction times*. UK: Academic Press.

Laming, D. R. J. (1984). The relativity of "absolute" judgements. *British Journal of Mathematical and Statistical Psychology, 37,* 152-183.

Link, S. W. (1992). *The Wave Theory of Difference and Similarity.* Hillsdale, NJ: Erlbaum.

Lockhead, G. R. (2004). Absolute judgments are relative: A re-interpretation of some psychophysical ideas. *Review of General Psychology.*

Lockhead, G.R., & Hinson, J. (1986). Range and sequence effects in judgment. *Perception & Psychophysics, 40*, 53-61.

Lockhead, G.R., & King, M. C. (1983). A memory model of sequential effects in memory tasks. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 461-473.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics, 32,* 397-408.

Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology, 37,* 136-151.

Marley, A. A. J., & Cook, V. T. (1986). A limited capacity rehearsal model for psychological judgments applied to magnitude estimation. *Journal of Mathematical Psychology, 30,* 339-390.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our

capacity for information processing. *Psychological Review, 63,* 81-97.

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics, 57,* 1065-1079.

Neath, I. & Brown, G. D. A. (2006). Further applications of a local distinctiveness model of memory. *Psychology of Learning and Memory, 46,* 201-243.

Nosofsky, R.M. (1983). Shifts of attention in the identification and discrimination of intensity. *Perception and Psychophysics, 33(2),* 103-112.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. In A. A. J. Marley (Ed.), *Choice, Decision, and Measurement* (pp. 347-365). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266-300.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112,* 383-416.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America, 24,* 745-749.

Pollack, I. (1953). The information of elementary auditory displays. II. *Journal of the Acoustical Society of America, 25,* 765-769.

Purks, S. R., Callahan, D. J., Braida, L. D., & Durlach, N. I. (1980). Intensity perception. X. Effect of preceding stimulus on identification performance. *Journal of the Acoustical Society of America, 67*, 634-637.

Ratcliff, R. (1978) A theory of memory retrieval. *Psychological Review*, 85, 59-108

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological*

Review, *88*, 552-572.

Ratcliff, R., Gomez, P. & McKoon, G. (2004).A diffusion model account of the lexical decision task *Psychological Review*, *111*, 159-182.

Ratcliff, R. & Rouder, J.N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347-356.

Ratcliff, R. & Smith, P.L. (2004). A comparison of sequential sampling models for two-choice reaction time, *Psychological Review*, *111*, 333–367.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358-367.

Romani, G. L., Williamson, S. J., & Kaufman, L. (1982). Tonotopic representation of the human auditory cortex. *Science*, 216, 1339-1340.

Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11,* 938-944.

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101,* 357-361.

Smith. P.L. (1995). Psychophysically principled models of simple visual reaction time. *Psychological Review, 102,* 567-593.

Smith, P. L., Ratcliff, R. & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays, *Vision Research, 44*, 1297-1320.

Stewart, N. (2007). Absolute identification is relative. *Psychological Review, 114*, 533-538.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative

judgment. *Psychological Review, 112*, 881-911.

Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Steven's law. *American Journal of Psychology, 86,* 3-27.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude estimation. *Perception and Psychophysics, 24,* 305-314.

Teghtsoonian, R. & Teghtsoonian, M. (1997). Range of acceptable stimulus intensities: An estimator of dynamic range for intensive perceptual continua. *Perception and Psychophysics, 59,* 721-728.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68-111.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550-592.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review, 11,* 579-615.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and self-organized criticality. *Journal of Experimental Psychology: General,* 134, 108-116.

Ward, L. M., & Lockhead, G. R. (1970). Sequential effect and memory in category judgment. *Journal of Experimental Psychology, 84,* 27-34.

Ward, L. M (1987). Remembrance of sounds past: Memory and psychophysical scaling. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 216-227.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9,* 73-78.

Weber, D. L., Green, D. M., and Luce, R. D. (1977) Effects of practice and distribution of auditory signals on absolute identification. *Perception & Psychophysics*, *22*, 223-231

Wiemer, J. C., & von Seelen, W. (2002). Topography from time-to-space transformations. *Neurocomputing*, 44-46, 1017-1022.

Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing.

Christopher Donkin[1], Scott D. Brown[1], Andrew Heathcote[1]
& A. A. J. Marley[2]

1. University of Newcastle
Newcastle, Australia

2. University of Victoria
British Columbia, Canada

Abstract

Identification accuracy for sets of perceptually discriminable stimuli ordered on a single dimension (e.g., line length) is remarkably low, indicating a fundamental limit on information processing capacity. This surprising limit has naturally led to a focus on measuring and modeling choice probability in absolute identification research. We show that choice response time (RT) results can enrich our understanding of absolute identification by investigating a dissociation between RT and accuracy as a function of stimulus spacing. The dissociation is predicted by the SAMBA model of absolute identification (Brown et al., 2008), but cannot easily be accommodated by other theories. We show that SAMBA provides an accurate, parameter free, account of the dissociation that emerges from the architecture of the model and the physical attributes of the stimuli, rather than through numerical adjustment. This violation of the pervasive monotonic relationship between RT and accuracy has implications for model development, which are discussed.

In many choice paradigms, more accurate responses are associated with faster response times, and vice versa: e.g., in Stroop-interference tasks (e.g. Kane & Engle, 2003; Wuehr & Frings, 2008), various naming tasks (Duyck, Lagrou, Gevers & Fias, 2008; Roelofs, 2006) and absolute identification (Kent & Lamberts, 2005; Lacouture & Marley, 1995, 2004; Petrov & Anderson, 2005). As a result, it might be thought that models predicting choice probability can also account for RT through a simple monotonic transformation (e.g., inversion). In this paper, we focus on the relationship between RT and accuracy in the absolute identification of unidimensional stimuli, where, on each trial, participants identify a randomly chosen stimulus from a set of stimuli varying on only one dimension. For example, the stimuli might be a set of 10 lines of varying lengths which are given the labels #1 through #10 from shortest to longest.

We expand upon a previous finding that RT is not always a simple monotonic function of accuracy in absolute identification, when stimulus spacing is manipulated (Lacouture, 1997). We demonstrate the reliability of this result, and show that it provides a powerful test of different theoretical accounts of absolute identification. This dissociation is predicted by the SAMBA theory (Brown, Marley, Donkin & Heathcote, 2008), which models both choice probability and choice response time. SAMBA predicts the violation of the pervasive negative correlation between RT and accuracy because of one of its components – the mapping model developed by Lacouture and Marley (1995). We show that SAMBA accounts for the dissociation without parameter adjustment, because its account emerges from the architecture of the model and the physical attributes of the stimuli.

There are numerous benchmark phenomena for the absolute identification of

unidimensional stimuli, when the stimuli are equally spaced. For instance, when mean RT and accuracy are plotted as functions of each stimulus' ordinal position within the set, one observes the ubiquitous “bow effect” – a U-shape for accuracy and an inverted U for RT. In such plots, stimuli associated with shorter RTs are always associated with higher accuracy, and vice versa. The inverse relationship between accuracy and RT is also observed in many other kinds of plots of absolute identification data, for example, if accuracy and mean RT are plotted as functions of: the difference between successive stimuli; the number of trials that have intervened since the current stimulus was last presented; or the number of stimuli within the set (Brown et al., 2008; Kent & Lamberts, 2005; Lacouture & Marley, 1995, 2004).

Brown et al. (2008) developed SAMBA to account for the benchmark empirical choice and RT phenomena. SAMBA was intended to be a complete account of absolute identification, one that included all stages from a psychophysical stimulus representation through to response selection, and one that modeled all of the important benchmark phenomena from the field. Brown et al. also pointed out that SAMBA makes some surprising and testable predictions, including that increasing the space between two adjacent stimuli will result in increased accuracy for those two stimuli, but will have little impact on RT. In the following sections we first provide a brief overview of SAMBA, followed by a detailed account of this particular prediction. We then examine data from Lacouture (1997) that confirm the prediction, and show SAMBA’s fit to the data. We conclude by discussing the implications of these results for theoretical development in the field of absolute identification.

SAMBA

SAMBA (Brown et al., 2008) is composed of three stages: the selective attention

stage, the mapping stage and the decision stage. The selective attention stage of SAMBA begins with a relatively impoverished psychophysical representation of a stimulus and produces an estimate of its magnitude. This estimate is constructed using Marley and Cook's (1984, 1986) theory, which posits that stimulus magnitudes are judged relative to a context defined by upper and lower "anchors", which are long term memories for the magnitues of very large and very small stimuli (call these L and U). On each trial, the magnitude of the stimulus is judged relative to the overall context, producing a noisy estimate. This magnitude estimate falls in the interval [0,1], and the average magnitude estimate for any given stimulus (over repeated presentations) is given by the linear function which maps the interval $[L,U]$ onto the interval [0,1]. For example, in an experiment with ten equally spaced stimuli, if stimulus #5 were presented, a magnitude estimate of close 0.45 might be expected, but on any given trial the estimate will vary somewhat from this average. The context used to judge stimuli is maintained by the observer using a noisy memory rehearsal process. The noise in this process is one of the key elements that introduces inaccuracy into SAMBA's predicted responses.

SAMBA's mapping stage transforms the magnitude estimate produced by the selective attention stage into response strengths, one for each possible response. It operates like a highly constrained set of tuning curves, where each curve produces a response strength that depends on how closely the observed magnitude estimate matches a referent for the given response. SAMBA assumes the referents are obtained by averaging magnitude estimates associated with repeated presentations of each stimulus. The mapping phase operates similarly to all tuning systems, in that the largest response strength is always assigned to the response whose referent most closely

matches the observed magnitude estimate. For example, a magnitude estimate of 0.45 might be closest to the long-term referent for stimulus #5 and so the largest response strength will be assigned to response #5. The outputs from the mapping phase are used as inputs for the decision stage of SAMBA, which uses a set of ballistic accumulators, one for each possible response (Brown & Heathcote, 2005). The ballistic accumulators instantiate a noisy max-picking algorithm. The chosen response will usually be the one with largest response strength (from the mapping stage), but not always. Larger response strengths are associated with faster responses, and bigger differences between response strengths are associated with more accurate "pick-the-max" behaviour.

In addition to these basic elements, SAMBA also includes mechanisms for sequential effects. For example, activity in the ballistic accumulators in the decision phase is assumed to decay slowly between trials. Amongst other things, this means that the response selected on the previous trial will have an advantage on the current trial, as observed in data. However, for our purposes the critical element of SAMBA is the bow mapping phase. Lacouture and Marley (1995) developed the mapping from a theoretical viewpoint. That is, they started out with a list of mathematical properties that any reasonable set of tuning curves should have. For example, for any reasonable set of tuning curves the greatest response strength should always assigned to the response whose long-term referent most closely matches the incoming magnitude estimate. Obviously, a great variety of tuning curves would satisfy this property, so other properties were included to constrain and simplify the solution, including: all response strengths should always be positive; the tuning curves should use the simplest functional form possible – a straight line; and the set of curves should be symmetric, as long as the referents are symmetric. Lacouture and Marley developed their bow

mapping as a set of linear tuning "curves" which satisfied all these constraints. Their solution was very powerful because it was also parameter free, being entirely specified by the values of the long term referents for each stimulus' average magnitude estimate. The mapping solution also predicted the ubiquitous bow effects observed in both response time and accuracy for absolute identification, even though these properties were not included as constraints for its development.

Of course, other solutions to the basic tuning curve problem could be developed. In particular, one may relax the simplifying constraints imposed by Lacouture and Marley (1995), by allowing more complex nonlinear forms for the tuning curves. Such solutions would probably also be able to accommodate the peculiar accuracy-RT dissociation we discuss below, but it is difficult to justify their extra complexity. From this point of view, one may consider the bow mapping as the simplest and most constrained set of tuning curves available, with the added benefit that they allow our model to accommodate all the required empirical data.

SAMBA's Predictions for Unequally Spaced Stimuli

SAMBA makes the prediction that if the spacing between two adjacent stimuli is increased, with other stimulus spacings unchanged, then these particular stimuli are identified with higher accuracy, but RT is relatively unaffected. This prediction is a consequence of Lacouture and Marley's (1995) mapping solution. When stimuli are unequally spaced, the long-term referents (average magnitude estimates) that define the mapping stage will reflect the unequal spacing. For example, first consider a standard absolute identification experiment with 10 equally spaced stimuli, and suppose that participants place their lower and upper anchors at a distance equivalent to one stimulus separation above and below the stimuli at the upper and lower end of the range,

respectively. In this case, the selective attention phase of SAMBA produces average magnitude estimates given by the linear mapping of the stimulus magnitudes onto the unit interval, namely: $\{\frac{1}{11}, \frac{2}{11} \ldots \frac{10}{11}\}$. However, now imagine that a set of 10 unequally spaced stimuli is constructed by first taking 14 equally spaced stimuli, then removing the central four. This stimulus set has a large central gap between stimuli #5 and #6. The selective attention phase of SAMBA then produces average magnitude estimates that respect the unequal stimulus spacing, namely $\{\frac{1}{15}, \frac{2}{15}, \ldots, \frac{5}{15}, \frac{10}{15}, \ldots, \frac{14}{15}\}$. Since the average estimates define the mapping solution, the spacing of the stimulus set is naturally encoded into the operation of the model.

On each trial of an absolute identification experiment, the selective attention phase produces a noisy magnitude estimate, say z. The mapping solution transforms this estimate into a response strength R_j for each of the j possible responses according to the formula $R_j=(2Y_j-1)z-Y_j^2+1$. The function is completely defined by Y_j, which is the average magnitude estimate for the jth stimulus – the long term referent for response j. Figure 1 illustrates the mapping solutions that arise from both the equally spaced and unequally spaced sets of ten stimuli. In both cases, the mapping supports the basic property, that if the observed magnitude estimate is close to the average magnitude estimate for stimulus j, the highest response strength will be assigned to response j. For example, suppose on a particular trial stimulus #5 is presented, and the selective attention phase produces a magnitude estimate of .45 units. In both the equal and unequal cases, the large black dot shows that the greatest response strength in this case is assigned to response #5 (i.e., the highest line above x=.45 is the one corresponding to the fifth response). It may strike the reader as surprising, on first glance, that the greatest response strength for each response is always assigned at one extreme or the other (x=0

or 1). For example, response #5 is the maximum-strength response at x=.45, but the *greatest* response strength is assigned to response #5 when x=0. This property arises from the severe simplifying constraints imposed by Lacouture and Marley (1995), in particular that the tuning curves should be linear. It is testament to the power of their solution that it still fits the data so well, even with such constraint.

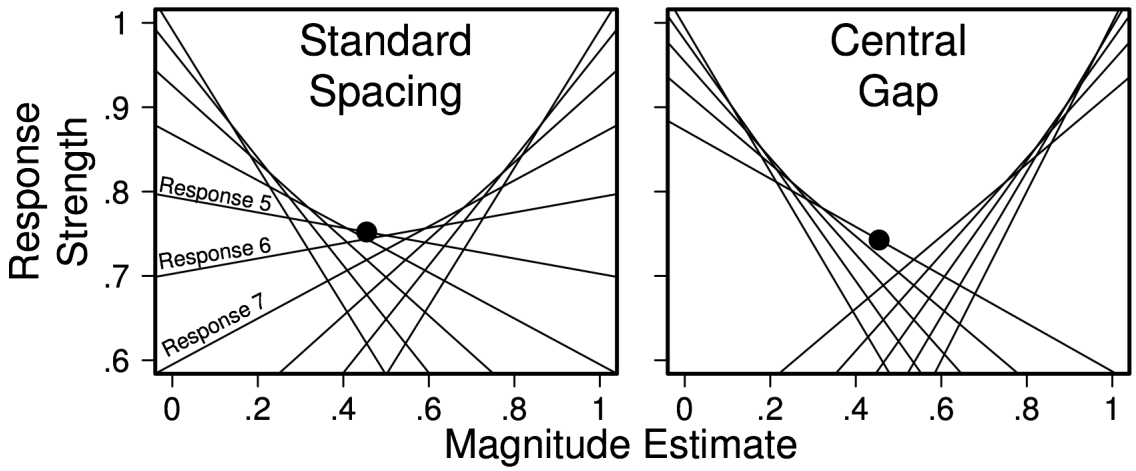

Figure 1. Mapping solution for equally spaced stimuli (left panel) and a set of ten stimuli with a central gap in stimulus spacing equivalent to four stimuli (right panel). Each line shows how the response strength varies with input magnitude estimate, for one of the 10 possible responses.

The difference in the mapping solutions for equally and unequally spaced stimuli leads to the prediction that is our focus. Consider again stimulus #5, which is adjacent to the large central gap in the unequal stimulus set (a similar argument applies to stimulus #6). In the equally spaced condition, response #5 is the maximum-strength response[1] for any magnitude estimates in the interval $z \in [\frac{4.5}{11}, \frac{5.5}{11}]$. However, in the unequally spaced condition, response #5 is the maximum-strength response for a larger range of magnitude estimates: $z \in [\frac{4.5}{15}, \frac{7.5}{15}]$. Given that the response with the largest strength is

1 It is elementary to show that responses j and j+1 have equal response strengths at the point that is midway between the long term referents for stimuli j and j+1, and this holds for both equally spaced and unequally spaced stimulus sets.

usually the response made by SAMBA's decision phase, accuracy is predicted to be higher for stimulus #5 in the unequally spaced condition than in the equally spaced condition. This is mostly due to the prediction that stimulus #5 will not often be confused with stimulus #6 (and vice versa). For example, for the unequally spaced stimuli in Figure 1, the average magnitude estimate for stimulus #5 is $\frac{5}{15}$. Due to the properties of the rehearsal stage, it is rare that if stimulus #5 were presented that the magnitude estimate would be greater than $\frac{7.5}{15}$, and hence fall in the region where response #6 would receive the largest response strength.

Turning now to predictions for response time, Figure 1 shows that the size of the response strength produced for stimulus #5 is about the same in both the equal and unequal stimulus spacing conditions. In the decision stage of SAMBA, the response strength for response #5 determines the rate of increase of activation in the corresponding ballistic accumulator. All other parameters being equal, response time is inversely related to the response strength, so SAMBA predicts about the same response times for stimulus #5 in both the equal and unequal spacing conditions. To be numerically precise, the average magnitude estimate associated with stimulus #5 in the equally spaced condition is $\frac{5}{11}$, and this results in a maximum response strength being assigned to response #5, a strength of $(2\frac{5}{11}-1)\frac{5}{11}-(\frac{5}{11})^2+1=0.752$. In the unequally spaced case, stimulus #5 generates an average magnitude estimate of $\frac{1}{3}$, but again the maximum response strength is assigned to response #5, $(2\frac{1}{3}-1)\frac{1}{3}-(\frac{1}{3})^2+1=0.778$. Critically, in both cases, the response strengths assigned to the correct response for neighbouring stimuli is larger: when stimulus #4 is presented, the average response strength assigned to response #4 is 0.769 in the equally spaced condition and 0.804 in

the unequally spaced condition. Since predicted RT is inversely related to response strength, SAMBA predicts the required dissociation, that responses to stimuli near the large gap will be more accurate than for neighbouring stimuli, but the corresponding response times will be slower.

Empirical Evidence

Several researchers have manipulated stimulus spacing, including Lockhead and Hinson (1986) and Lacouture (1997). Brown et al. (2008) demonstrated that SAMBA provides a parsimonious account of the choice probabilities of Lockhead and Hinson (RTs were not recorded). Lacouture's data set included RT measurements, allowing us to test SAMBA's prediction of the effect of stimulus spacing on both choice probabilities and RT. Participants in Lacouture's experiment spent the first hour in a standard absolute identification experiment, with 10 equally spaced stimuli. Each participant then spent a second hour in one of several conditions in which physical properties of the stimuli were manipulated. Brown et al. presented fits of SAMBA to data from the first session (equal spacing), but until now the unequal spacing conditions have never been modeled. Since Lacouture published his findings, several important, integrative theories of absolute identification have been published, some of which have even addressed the effects of unequal stimulus spacing on response choices, but none of which have addressed the effects of stimulus spacing on response times. This leaves a gap in the theoretical development, especially given that Lacouture's data present such a challenging test for models.

Participants in the second session of Lacouture's (1997) experiment experienced one of six conditions, four of which employed unequally spaced stimuli. These four conditions had larger gaps introduced either in the centre (between stimuli #5 and #6) or

at the edges (between stimuli #2 and #3 and stimuli #8 and #9). The gap location was crossed with a manipulation of gap size (large or small) to create the four conditions: a large central gap (C-L); a small central gap (C-S); large extreme gaps (E-L); and small extreme-gaps (E-S). The top row of Figure 2 provides a schematic illustration of the stimuli from these four conditions (for actual stimulus lengths, see Lacouture's Table 1). The second and third rows of Figure 2 show data from the four spacing conditions, replicating Lacouture's Figure 4. The data are represented by solid circles, with error bars showing +/-1 standard errors, calculated assuming normal distributions across subjects for mean RT and binomial distributions for accuracy. In each graph, vertical arrows show the locations of the larger gaps, and the dashed lines show predictions generated by SAMBA. The second row of Figure 2 shows response accuracy separately for each response and the third row shows mean correct response times. Notice that accuracy is greater for stimuli adjacent to gaps, and the effect is more pronounced in the large spacing conditions than the small spacing conditions. However, the improved accuracy is never accompanied by faster response times, contrary to the typical inverse RT-accuracy relationship.

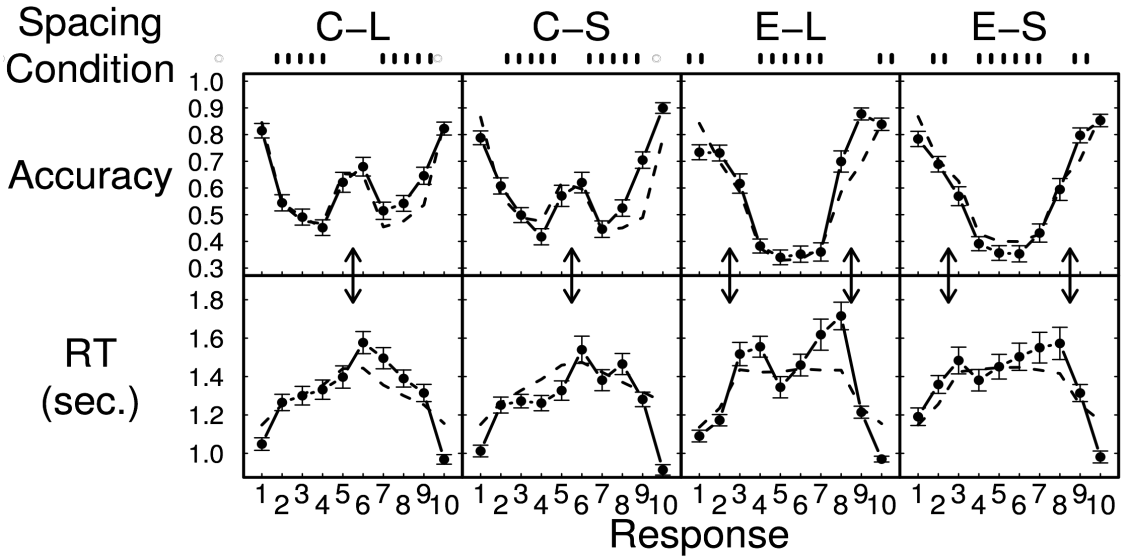

Figure 2. The top row shows a schematic representation of the stimuli used by Lacouture (1997). C-L refers to the 'large central-gap' and C-S to the 'small central-gap' condition, E-L refers to the 'large extreme-gaps' condition, and E-S 'small extreme-gaps. The second row shows response accuracy and the third row shows mean RT for correct responses, both as functions of response. Data are shown with solid lines and points, and SAMBA's fits with dotted lines.

When analyzing response choices from absolute identification tasks, it is customary to calculate sensitivity (*d'*) instead of raw percent correct (Luce, Nosofsky, Green and Smith, 1982). Sensitivity provides a bias-free measure of how often successive pairs of stimuli are confused, that is, how often stimuli #1 and #2 are confused, and stimuli #2 and #3, and so on up to stimuli #9 and #10. For any given pair, say stimuli #4 and #5, *d'* is calculated in the usual manner, using hit and false alarm rates, where "hits" are defined as responses #5 or greater, when stimulus #5 is presented, and "false alarms" are defined as responses #5 or greater, when stimulus #4 is presented. To ensure that the effects observed in Lacouture's (1997) data were not due to a response bias effect, we calculated *d'* values for each stimulus pair, shown in Figure 3. Graphing the data using *d'* shows an even more pronounced effect of stimulus spacing – stimuli that are separated by large gaps were almost never confused with one another.

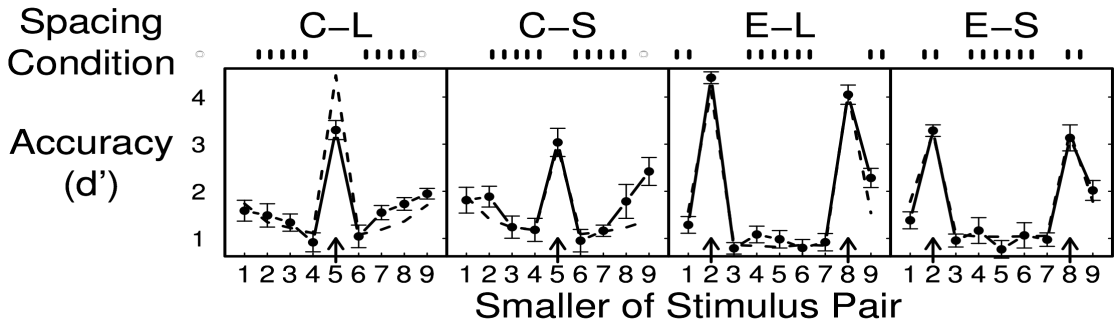

Figure 3. Sensitivity (*d'*) for each stimulus pair in the four unequally spaced stimulus conditions from Lacouture (1997). Error bars show standard errors assuming normally distributed *d'* values across participants.

Figures 2 and 3 demonstrate a clear dissociation – stimuli separated by large gaps enjoy an accuracy (and sensitivity) bonus, but no corresponding RT bonus. To confirm the statistical reliability of this dissociation, we calculated binomial tests. We used binomial tests because they provide robust analyses that directly test the ordinal hypotheses we entertain, without problematic distributional assumptions. We carried out two tests, one for the dissociation of response times and raw response accuracy and the other for the dissociation of response times and *d'*. We examined the accuracy (or *d'*) and RT values separately for each participant, and counted how frequently the dissociation in question was observed on a single-participant basis – that is, how often we observed increased accuracy for stimuli near large gaps (relative to neighbouring stimuli) without a corresponding increase in RT. For each of the sixteen participants in the E-S and E-L conditions, there were two opportunities to observe the dissocation – two increased stimulus gaps in each condition – and for each of the other sixteen participants in the C-S and C-L conditions there was one opportunity. For the raw accuracy data, under a null hypothesis of no relationship, the probability of observing the dissociation by chance is 1 in 16 at each opportunity, but we observed the dissociation 10 out of 48 times, significantly more than the three that would be expected

by chance (p=.0015). For the *d'* data, the probability of observing the dissociative ordering by chance is 1 in 12, but we observed the dissociation 20 out of 48 times, again significantly more than would be expected by chance ($p<10^{-10}$). These tests are quite convincing, especially given the limited number of participants available (only eight in each condition) and the reduced power afforded by robust non-parametric statistical tests. Note that those participants who did *not* demonstrate the critical dissociation on single-participant level did not necessarily demonstrate the opposite (i.e., the usual inverse accuracy-RT relationship). In fact, of the 96 opportunities to observe the usual inverse relationship in these critical tests, we observed it only once – for the other 95 opportunities, we either observed random ordering due to noise (65 times) or the accuracy-RT dissociations counted above (30 times).

Turning now to predictions from SAMBA, we can see from the dashed lines in Figures 2 and 3 that the model provides a good qualitative account of the data, capturing the observed dissociation between accuracy and RT. The model also provides a very close quantitative fit to the data, which is all the more surprising given the strong constraints we imposed on the parameters. To fit SAMBA to Lacouture's (1997) unequal spacing conditions, we began with the parameters reported by Brown et al. (2008) that were used to fit the standard (equal spacing) condition from the first session of Lacouture's experiment. Only three parameters were adjusted for the fits presented in Figure 2, and even these parameters were irrelevant in capturing the critical dissociation between accuracy and RT – all three parameters were instead related to the effects of practice, capturing differences in the data between the first and second experimental sessions. Firstly, we dcreased the response threshold parameter for SAMBA's decision phase to be 90% of the value it took for the first experimental session, reflecting that

participants may have become a little less careful in the second session of the experiment. Secondly, we had previously noted asymmetry in the data: in the first experimental session, responses were slower and less accurate for the large stimuli than the small stimuli. SAMBA accounted for the asymmetry by setting the lower anchor close to the smallest stimulus (*L* was set at 95% of the magnitude of the smallest stimulus) but the upper anchor quite far away from the largest stimulus (*U* was 62% larger than the largest stimulus). In the second experimental session the asymmetry disappeared: note that in Figure 2 accuracy and RT are about the same for the smaller stimuli as for the larger stimuli. To capture this return to symmetry, we set the lower anchor to 99% of the magnitude of the smallest stimulus, and the upper anchor to 101% of the magnitude of the largest stimulus. One possible interpretation for the change in symmetry between sessions could be improvement due to practice. Although absolute identification is mostly immune to practice effects, Rouder, Morey, Cowan and Pfaltz (2004) showed that learning in absolute identification is possible. Donkin, Dodds, Brown and Heathcote (submitted) have shown that this is especially true for lines of varying length, as used by Lacouture. This explanation is consistent with the change in parameters of SAMBA used to achieve the reported fits. The upper and lower anchors, *U* and *L*, were moved closer to stimuli #1 and #10 in the unequal spacing conditions, indicating that participants improved their knowledge of the task in the second session relative to the first.

It is notable that a single set of parameters was used to fit all four different spacing conditions. The differences between spacing conditions are completely determined by the properties of the stimulus spacing, which in turn determine the referents. For example, referents for the large central-gap condition are based on long-

term averages of magnitude estimates produced by SAMBA's selective attention phase, and these magnitude estimates naturally reflect the large gap between stimuli #5 and #6. The same mechanism applies to the other stimulus spacing conditions. Given the constraints we imposed on the model parameters, the quantitative fits are quite good, although SAMBA overpredicts the improvement in *d*' in the C-L condition.

Alternative Models

There are four recent models of absolute identification, besides SAMBA, that make predictions for both choice and RT. Two of these models are exemplar based accounts of general categorization behavior, applied to absolute identification, which can be seen as a special case of categorization. These two models are the exemplar-based random walk (EBRW: Nosofsky, 1997; Nosofsky & Palmeri, 1997) and the extended generalized context model for response times (EGCM-RT: Kent & Lamberts, 2005, Lamberts, 2000). Both models predict that increased accuracy should always be associated with faster RT, at least when parameters unrelated to stimulus properties are kept constant. Lacouture's (1997) dissociation of RT and accuracy was observed within blocks in which only stimulus magnitude was manipulated, so it would seem that these theories are incapable of accounting for the dissociation between accuracy and RT with unequal stimulus spacing (in particular, see Equation 5 in Nosofsky, 1997, and Equation 12 in Lamberts, 2000). The EBRW and EGCM-RT both predict the observed increase in accuracy with increased spacing between stimuli, caused by reduced similarity between stimulus representations. However, both models also predict an associated decrease in RT, which was not observed in Lacouture's data. It is possible that, with carefully chosen parameter values, these models could decrease the size of the predicted misfit. That is, there may exist parameter values that allow the models to predict increased

accuracy near large gaps, accompanied by only a small decrease in RT for those same stimuli. Even if these parameter values exist, the models still make incorrect predictions about the (statistically reliable) ordering of the data values observed above.

Another absolute identification model that predicts RT is Karpiuk, Lacouture and Marley's (1997) limited capacity, wave equality, random-walk model. This model is similar to SAMBA, in that it uses Marley and Cook's (1984) rehearsal model, but in place of SAMBA's mapping stage, Karpiuk et al. used a set of tuning curves for each response. Tuning curves, specified by free parameters, operate like SAMBA's mapping stage but with less constraint and greater flexibility. For this reason, it is quite likely that Karpuik et al.'s model is capable of capturing (but not predicting) the observed dissociation between RT and accuracy. Lacouture and Marley's (1995, 2004) mapping model employs the same mapping functions as SAMBA, and so it also predicts the dissociation between RT and accuracy.

Ashby (2000) developed a theory of categorization that includes predictions for RT as well as choices. Other versions of this theory have been applied to absolute identification data (Ashby & Lee, 1991), although the RT-inclusive version has not. Similarly to the exemplar-based categorization models, Ashby's theory predicts a monotonic relationship between mean RT and accuracy in categorization (see, e.g., Ashby, 2000, p.321 for a summary of the extensive successes, and limited failures, of this prediction) and, therefore, does not accommodate the observed dissociation, without modification. We note also that extant absolute identification models (other than SAMBA) that predict both accuracy and RT fail to predict other key phenomena. For example, none of the models described above predict the well-known sequential effects in absolute identification data, such as assimilation and contrast.

Discussion

We have presented and tested a prediction arising from the mapping stage of SAMBA. The prediction of a dissociation between accuracy and RT is surprising due to the regularity with which a monotonic relationship has been observed. Nevertheless, SAMBA's prediction is confirmed by data from Lacouture's (1997) previously unaddressed unequal spacing experiments. SAMBA accounts for the dissociation between RT and accuracy under different spacing conditions, and provides an impressive quantitative fit, given that no parameter changes were made between conditions, and all but three parameters were fixed at values estimated using data from a different condition.

In most empirical and theoretical work on absolute identification, response times have received much less attention than response choices. Despite an empirical research history pre-dating Miller's (1956) seminal review, and theoretical accounts existing for at least 50 years, models have only begun to address RT in the last 15 years. The disinterest in RT is underlined in Stewart, Brown and Chater's (2005) model summary table (*p.*886), where only 3 out of the 14 models reviewed made predictions about RT. This neglect is most likely due the belief that RT has little utility for discriminating models, which might have been true if a systematic monotonic inverse relationship between RT and accuracy always held. However, Lacouture's (1997) results show that this is not the case, and that RT and accuracy data together provide greater model constraint than accuracy data alone. In particular, Lacouture's data provide a strong test for any theoretical account of absolute identification that attempts to account for both choice and RT. SAMBA passes this test, confirming a prediction made by Lacouture and Marley's (1995) highly constrained method of obtaining tuning curves, adopted by SAMBA. Hence, Lacouture and Marley's method, motivated entirely independently on

theoretical grounds, not only predicts the ubiquitous bow effects found in absolute identification, but also a heretofore unexplored dissociation between speed and accuracy.

References

Ashby, F.G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology, 44*, 310-329.

Ashby, F.G., & Lee, W.W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General, 120*(2), 150-172.

Brown, S., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated architecture for absolute identification. *Psychological Review.*

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112*(1), 117-128.

Brown, S.D., & Heathcote, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology.*

Donkin, C., Dodds, P., Brown, S., & Heathcote, A. (2008). *Revisiting the Limits of Human Information Processing Capacity.* Manuscript submitted for publication.

Duyck, W., Lagrou, E., Gevers, W., & Fias, W. (2008). Roman Digit Naming : Evidence for a Semantic Route. *Experimental Psychology, 55*(2), 73-81.

Kane, M. J., & Engle, R. W. (2003). Working-Memory Capacity and the Control of Attention: The Contributions of Goal Neglect, Response Competition, and Task Set to Stroop Interference. *Journal of Experimental Psychology, 132,* 47-70.

Karpiuk, P., Jr., Lacouture, Y., & Marley, A. A. J. (1997). A limited capacity, wave equality, random walk model of absolute identification. In A. A. J. Marley (Ed.), *Choice, decision and measurement: Essays in honour of R. Duncan Luce* (pp.279-299). Mahwah, NJ: Erlbaum.

Kent, C., & Lamberts, L. (2005). An exemplar account of the bow and set size effects in absolute identification. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, 31, 289-305.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60,* 121-133.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39,* 383-395.

Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception & Psychophysics, 66,* 1206-1226.

Lamberts, K. (2000). Information-Accumulation Theory of Speeded Categorization. *Psychological Review, 107*, 227-260.

Lockhead, G.R., & Hinson, J. (1986). Range and sequence effects in judgment. *Perception & Psychophysics, 40*, 53-61.

Luce, R.D., Nosofsky, R.M., Green, D.M., & Smith, A.F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics, 32*(5), 397-408.

Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology, 37,* 136-151.

Marley, A. A. J., & Cook, V. T. (1986). A limited capacity rehearsal model for psychological judgments applied to magnitude estimation. *Journal of Mathematical Psychology, 30,* 339-390.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63,* 81-97.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded

categorization and absolute judgment. In A. A. J. Marley (Ed.), *Choice, decision and measurement: Essays in Honor of R. Duncan Luce* (pp. 347-365). Mahwah, NJ: Erlbaum

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded categorization. *Psychological Review, 104,* 266-300.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112,* 383-416.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59 –108.

Roeflofs, A. (2006). Functional architecture of naming dice, digits, and number words. *Language and Cognitive Processes, 21,* 78-111.

Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11,* 938-944.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112,* 881-911.

Wuehr, P., & Frings, C. (2008). A Case for Inhibition: Visual attention suppresses the processing of irrelevant objects. *Journal of Experimental Psychology: General, 137,* 116-130.

Appendix: The Latencies of Incorrect Responses

The relative speeds of correct and incorrect responses have proven very illuminating in the development of theories of choice response time (see, e.g., Brown & Heathcote, 2005, 2008). Theoretical accounts of response times in absolute identification are less well developed, so the fine model discrimination afforded by the analysis of error RT may yet be premature. Nevertheless, we note here two interesting phenomena related to incorrect RTs in Lacouture's (1997) data. Firstly, response times were slightly, but reliably, slower for incorrect responses than correct responses in the second session of Lacouture's experiment (mean difference 29msec, $t(38)=2.5$, $p<.01$). Secondly, the relative speed of correct and incorrect responses changed systematically with the stimulus magnitude. For extreme stimuli (#1 and #10), incorrect responses were much slower than correct responses (mean difference 269msec, $t(46)=7.5$, $p<.001$) but for central stimuli (#5 and #6) there was almost no difference (mean difference 10msec, $t(46)=0.3$, $p>.05$). The relative speeds of correct and incorrect responses are captured well by SAMBA – as a brief illustration, Figure A1 shows mean error response times along with SAMBA's predictions using the same format as Figure 2. The model captures the global qualitative trends in the data, but misses some of the finer quantitative properties, such as the tendency for some extreme responses to be associated with very fast errors (e.g., #1 in C-L condition and #10 in E-S and C-S).

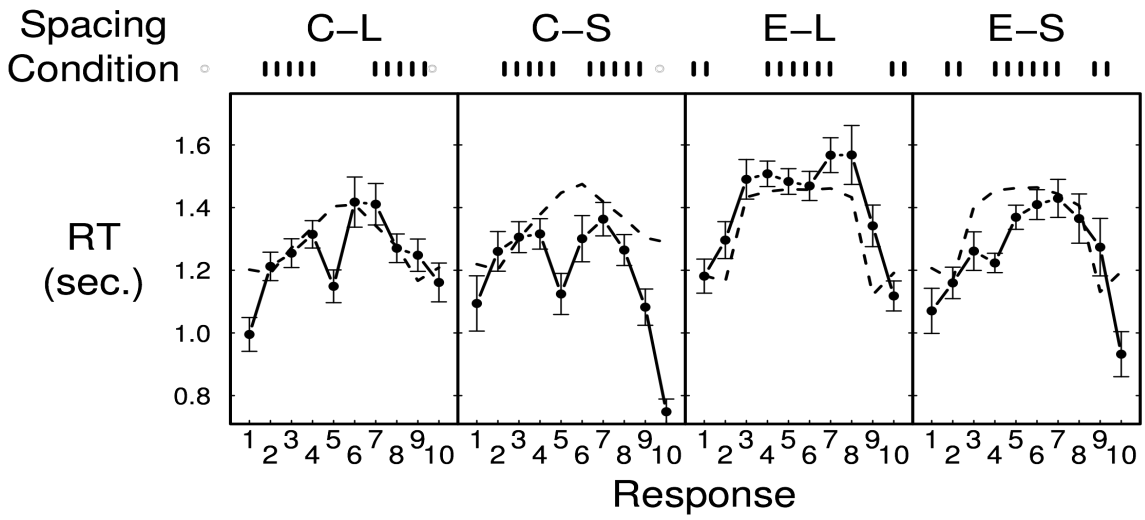

Figure A1. Mean response times for incorrect responses, along with predictions from SAMBA. Error bars show +/-1 standard error assuming that mean RT is normally distributed across participants.

We do not take this goodness of fit to be as impressive as SAMBA's ability to fit the effects of stimulus spacing that is our main focus. While the patterns of fast and slow errors may appear complex at first glance, they are less theoretically challenging than might be imagined. For example, incorrect response times were slower than correct response times, as predicted by SAMBA. Other models of absolute identification do not predict this in their current forms. For example, Kent and Lamberts' (2005) model uses a random walk, which is constrained to predict equal response times for correct and incorrect responses (see, e.g., Ratcliff, 1978). However, this limitation is not central to Kent and Lamberts' model, and can easily be remedied by the addition of certain variance components to its decision phase (as described by Ratcliff). Similarly, any model of absolute identification that predicts the ubiquitous bow effects – longer RT for central responses and shorter RT for extreme responses – will also successfully accommodate our second observation. That is, error responses will necessarily be slow for extreme stimuli, because those incorrect responses are less extreme (usually #2 and #9, rather than #1 and #10, for example). For these reasons, we think that a detailed

comparison of empirical results with theoretical predictions for incorrect response times may yet be premature for models of absolute identification.