Sequential Effects

in

Human Performance

Paul Williams

B. Psych (Hons I)

A thesis submitted in fulfilment of the requirements for the degree of

Doctorate of Philosophy (Clinical Psychology – Science)

8 November, 2016

Statement of Originality

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 the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968.

***** Unless an Embargo has been approved for a determined period.

Statement of Collaboration

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.

Statement of Authorship

I hereby certify that the work embodied in this thesis contains published papers/scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publications/scholarly work.

Thesis by Publication

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.

Paul Williams

B. Psych. (Hons I)

Acknowledgments

The period over which I have produced this dissertation has been a time of rapid change and personal growth. I have met, married, and fallen pregnant with (can I say this? it sounds nicer than 'impregnated') my wife. I've lived in three homes and held countless different work roles, culminating in practicing Clinical Psychology full-time over the last 12 months. I've raged against the bureaucratic and monolithic system that has benefitted me and infuriated me in equal parts. You don't get through periods like this without enormous support from family, friends, and many others. Thank you all. Three people though deserve special mention:

My primary supervisor and friend *Dr Ami Eidels* shepherded me through each step - and occasional meltdown - with calmness and patience. Ami has an infectious curiosity, especially for cognitive science and pedagogy. Somehow though, his vast knowledge is often overshadowed by his good-humor, graciousness, and kindness. I owe you many debts Ami, but above all a debt of intangibles.

My secondary supervisor and friend *Professor Andrew Heathcote* has also been central to the completion of this work. Andrew's talent for mathematical science is perhaps only bettered by his ability to help students and colleagues achieve goals and progress their careers. You are a genuine and kind man Andrew and it was a privilege to work within a research group where you set the culture.

Lastly, I would not have made it without my best friend and wife, *Karlye Damaso*. Your support and belief has dragged me through the periods when I was stuck, down, or second-guessing myself. In essence, you have replaced nicotine and caffeine! I am so excited to see where parenthood takes us and what (REF) has in store for us.

Publications Included in Thesis

In order of presentation:

1. Williams, P., Nesbitt, K., Eidels, A., & Elliot, D. (2011). Balancing Risk and Reward to Develop an Optimal Hot-Hand Game. . *Game Studies, 11*(1).

2. Williams, P., Nesbitt, K., Eidels, A., Washburn, M., & Cornforth, D. (2013). Evaluating Player Strategies in the Design of a Hot Hand Game. *GSTF Journal on Computing (JoC), 3*(2), 1-11.

3. Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgement and Decision Making, 11*(2), 174-184.

4. Ben-Haim, M. S., **Williams, P.,** Howard, Z., Mama, Y., Eidels, A., Algom, D. (2016) The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content. *Journal of Visualised Experiments.* (112), e53720.

5. Ross, R., **Williams, P.,** Eidels, A., (*2016*) An investigation into how emotional words affect processing in the emotional Stroop task. *Submitted to Cognition and Emotion.*

6. Williams, P., Howard, Z., Ross, R., Eidels, A. Cognitive dysfunction under emotional exposure: When participants with depression symptoms show no cognitive control. *Submitted to Psychological Science*

Statement of Contribution

This statement outlines Research Higher Degree Candidate Paul Williams' contribution to the series of manuscripts included in his thesis. The manuscripts included as part of his thesis are listed below, with a statement of his contribution to each.

Doctor Ami Eidels	Professor Andrew Heathcote
Doctor Keith Nesbitt	Doctor David Cornforth
Doctor Moshe Shay Ben-Haim	Doctor Yaniv Mama
Professor Daniel Algom	Miss Rachel Ross

Mister Zachary Howard

Mister David Elliot

~In lieu of Mr. Washburn's signature, please see attached Statement of Contribution Letter from Dr. Ami Eidels~.

Mister
Mark Washburn

……..

Endorsed by:

Associate Professor

Frances Martin

8 November, 2016

Assistant Dean Research Training

FSCIT

……..

Paul Williams led the following three manuscripts. In all cases Paul supervised data collection, completed the data analysis, and took the lead role in manuscript preparation.

Williams, P., Nesbitt, K., Eidels, A., & Elliot, D. (2011). Balancing Risk and Reward to Develop an Optimal Hot-Hand Game. . *Game Studies, 11*(1).

Williams, P., Nesbitt, K., Eidels, A., Washburn, M., & Cornforth, D. (2013). Evaluating Player Strategies in the Design of a Hot Hand Game. *GSTF Journal on Computing (JoC), 3*(2), 1-11.

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgement and Decision Making, 11*(2), 174-184.

Paul Williams led the following manuscript. Using archived data, Paul designed the research question/s, performed all analyses, and took the lead role in manuscript preparation.

Williams, P., Howard, Z., Ross, R., Eidels, A. (2016) Cognitive dysfunction under emotional exposure: When participants with depression symptoms show no cognitive control. *Submitted to Psychological Science.*

Paul Williams contributed to the following manuscript. Paul performed the analyses, and collaboratively prepared the manuscript.

Ross, R., **Williams, P.,** Eidels, A., (2016) How do emotional words affect cognitive processing? Submitted to Emotion and Cognition

Paul Williams contributed to the following methodological review manuscript. Paul collaboratively prepared the manuscript.

Ben-Haim, M. S., **Williams, P.,** Howard, Z., Mama, Y., Eidels, A., Algom, D. (2016) The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content. *Journal of Visualised Experiments.* (112), e53720.

November 7, 2016

Dr. Ami Eidels
School of Psychology
University of Newcastle
Phone: +61 – 2 – 4921 7089
E mail: Ami.Eidels@newcastle.edu.au

To: Office of Graduate Studies, UoN

Contribution statement to *thesis-by-publication*, Mr Paul Williams

Mr Paul Williams is about to submit his PhD thesis "Sequential Effects in Human Performance", under my supervision. He plans to submit a *thesis by publication*, based on a series of peer-reviewed articles he co-authored that are either published or under review.

For thesis by publication, the candidate's co-authors need to sign a statement of author contribution in which they indicate that the PhD candidate contributed substantially to each of the relevant papers. Paul was able to contact and obtain signatures from all co-authors but one – Mr Mark Washburn, who was an Honours student in my lab in 2011.

Since we were not able to contact Mark over the last two weeks, I am providing below a statement concerning Paul's contribution to the paper he co-authored with Mark and myself. While this is not Mark's signature, it represents what I know about the project and Paul's involvement, from the Principle Investigator perceptive, as well as that of a supervisor to both Paul and Mark:

"I confirm that, to the best of my knowledge as Principle Investigator of the project and supervisor for Williams and Washburn, Mr Paul Williams supervised data collection, completed the data analysis, and took the lead role in manuscript preparation of the paper *Evaluating Player Strategies in the Design of a Hot Hand*

Game, coauthored by Paul Williams (lead author), Keith Nesbitt, Ami Eidels, Mark Washburn, and David Cornforth."

Dr Ami Eidels
School of Psychology
University of Newcastle

Table of Contents

Abstract

In this thesis I explore the influence of the recent past on future human performance. That is, how does an event or outcome that occurs at point A influence performance at point *B,* if at all? The theoretical topics and statistical methods are therefore *sequential* in nature. I explore two aspects of human performance under this general framework. In the first section, comprising three chapters, I develop and employ a novel paradigm to explore basic cognition. The development of the paradigm is documented in Chapter 1 and Chapter 2, before Chapter 3 ultimately applies the paradigm to explore the hot hand belief and post-error slowing. *The hot hand belief* is the belief that the probability of a success given recent success will be greater than the probability of a hit given recent failure (Gilovitch, Vallone, & Tversky, 1985). *Post-error slowing* describes systematic increases in response time (RT) following errors in rapid choice tasks (Laming, 1968; Rabbitt, 1966). We investigate the hot hand and post-error slowing simultaneously, noting that both of these research areas had moved toward appraising the sequential influence of success and failure on dual performance dimensions: accuracy and RT (post-error slowing), or accuracy and difficulty (hot hand). We observed no post-error slowing for paid participants, and systematic post-error speeding for unpaid participants. Thus, we provide evidence for the newly emerging hypothesis that post-error slowing is not ubiquitous, but rather task and situation dependent. When post-error speeding was observed, we also observed the rarely documented hot hand effect, suggesting the hot hand may be more prevalent in low motivation contexts. In the second section, also comprising three chapters, I explore clinical applications of sequential methodologies. In Chapter 4 I document the best practice design of the emotional Stroop task (Williams, Mathews, & MacLeod,

1996) - a well-established paradigm used to explore the impact of emotional stimuli on human performance. In Chapters 5 and 6 I go on to explore sequential effects in this task for participants with and without symptoms of depression. In Chapter 5 I explore whether or not the presentation of an emotional word impacts performance on only the current trial (fast effect), or also on a subsequent trial (slow effect). Unlike previous efforts, we found no evidence of a slow effect in our data. In Chapter 6, we explore post-error slowing in the emotional Stroop task. Post-error slowing is a benchmark effect for *cognitive control*, which is the ongoing monitoring and regulation of actions and performance. We document that major depression symptoms are linked to severe deficits in cognitive control following errors and that these deficits are specific for emotional and non-emotional stimuli. These findings help constrain existing theories of error detection and correction, offer insights into the cognitive processes underpinning depression, and suggest that under emotional priming, major depression is marked by a complete failure to adapt behaviour in response to relevant environmental feedback.

General Introduction

The influence of the recent past on future human performance has been a hot topic in research psychology for some time. That is, how does an event or outcome that occurs at point A influence performance at a later point B, if at all? Examples of such sequential effects in research psychology are so common that they are considered ubiquitous. The prototypical sequential effect involves the repetition or alternation of stimuli in two alternative rapid-choice tasks. That is, as a stimuli sequence moves from one trial to the next, stimuli may repeat and therefore require the same response, or they may alternate and require the alternate response. These presentation-order sequential-effects have been found to influence choice probabilities and response time (RT) in domains such as stimulus detection (Posner & Cohen, 1984), categorization (Stewart, Brown, & Chater, 2002), and decision-making (Hogarth & Einhorn, 1992). Sequential effects have also been used as benchmarks to assess competing theoretical models in more complex decision-making paradigms such as absolute identification (Luce, Nosofsky, Green, & Smith, 1982). A prototypical absolute identification task might involve a participant identifying which of eight lines they were presented from a set which differ only in length. When stimuli are presented, participants identify the line size, from 1 through 8. The presentation order of lines leading up to the current trial has been shown to have a strong influence on choice probabilities. For example, when an incorrect response is given, it tends to be toward, rather than away from, the stimulus from the previous trial. Accuracy is also improved when successive trials display similarly sized stimuli. Models that aim to provide a complete account of absolute identification must account for these sequential effects (Brown, Marley, Donkin, & Heathcote, 2008).

Sequential effects are not, however, limited to the presentation order of stimuli in cognitive tasks. Performance on a current attempt may be systematically influenced by many preceding events, such as whether or not the previous attempt was successful or unsuccessful, or even whether a previous stimulus carried emotional content or no emotional content. The six papers that make up this thesis explore the influence of these more eclectic sequential effects. The thesis is broken into six chapters, each comprising a published (or in the case of Chapters 5 and 6, close to-be-published) paper. While the six chapters are linked and form a coherent whole; for organisational purposes the thesis is broken into two sections. Chapters 1 to 3 make up Section 1, and Chapters 4 to 6 make up Section 2.

In Section 1 (Chapters 1-3) the development of a novel cognitive game is documented. It was initially developed to explore the hot hand belief. *The hot hand belief* is the belief that the probability of a success given recent success will be greater than the probability of a hit given recent failure (Gilovitch, Vallone, & Tversky, 1985). The development of the cognitive game is documented in the first and second papers, before the third paper applies the game to explore both the hot hand belief and post-error slowing. *Post-error slowing* describes systematic increases in response time (RT) following errors in rapid choice tasks (Laming, 1968; Rabbitt, 1966). In Paper 3 we make two novel contributions. Firstly, we test the theoretical and empirical links between the hot hand and post-error slowing. Secondly, we compare three methods for calculating post-error slowing. This methodological advance allowed, to our knowledge for the first time, an assessment of the relative contribution of the local effect of errors to the global effects that may contribute to post-error adjustments, such as fatigue or boredom.

In Section 2 (Chapters 4-6), the key measurement methods developed in Section 1 are applied to a well-established task using a clinical sample. In Paper 4 we document the best practice design of the emotional Stroop task (Williams, Mathews, & MacLeod, 1996) - the well-established paradigm we use to explore the impact of emotional stimuli on performance. In the emotional Stroop task, emotional and non-emotional words are presented in sequence (one word at a time) to participants who are asked to identify the print-colour of the presented word. Amazingly, even though the emotional valence of the word is superfluous to the task, whether the stimuli are emotional or not has been shown to influence responding on the immediate and subsequent trials. In Papers 5 and 6 we apply the sequential methodologies outlined in Section 1 to data collected using the emotional Stroop task from participants with and without symptoms of depression. In Paper 5 we examine whether or not the presentation of an emotional word impacts performance on only the current trial (fast effect), or also on a subsequent trial (slow effect). In Paper 6 we examine post-error slowing for participants with and without depression symptoms in the emotional Stroop task. In this work we explore, again for the first time, the interaction effects of emotional content *and* success or failure on subsequent performance. We document that major depression symptoms are linked to severe deficits in cognitive control following errors and that these deficits are specific for emotional and non-emotional stimuli. These findings help constrain existing theories of error detection and correction, and have profound implications for understanding maladaptive behaviour from sufferers of major depression.

The thesis is organised such that each chapter contains a single paper (i.e., Chapter 1 contains Paper 1, and so on), and any additional material that may be useful to better understand the paper and its contributions. The additional material varies by

chapter, and might include a literature review, and/or additional analysis, and/or a general paper overview. The six chapters are not of equal length; rather, the length of each chapter gives an insight into the importance of that chapter to my scientific contribution. To help explain, consider that Section 1 outlines a complete body of work – from paradigm development and piloting through to the contribution derived from applying this paradigm. As such, the early chapters of Section 1 are more comprehensive than those that appear later. Because additional material is located alongside each paper, the cognitive load required of the reader is reduced and the flow of the document is improved. The most apparent difference between this approach and a more traditional thesis format is that this general introduction does not contain a large literature review. Rather, additional literature review/s, where required, are contained within the chapter dedicated to the paper that they support.

Section 1

Chapters 1, 2, and 3

Chapter 1

Chapter 1 might be considered a defining chapter as it establishes my interest in sequential effects, and ultimately sets the direction of this thesis. As such, Chapter 1 is comprehensive. It is made up of three components. Firstly, I provide a focused literature review of *the hot hand belief*. This review outlines the theoretical motivations for our exploration of sequential effects in Paper 1. Following this literature review is the *Paper 1 Overview*, which highlights the unique contributions of Paper 1. This overview might be best read in conjunction with *Paper 1*, which is presented in full to conclude the chapter.

The Hot Hand Belief

What is the Hot Hand Belief?

The effects of success and failure on performance have the capacity to engage both the research scientist and the casual observer. This is not surprising given the intrinsic interest, and self-interest, many individuals have when it comes to understanding human performance. Understanding when and how performance might fluctuate holds value for gamblers, investors, or anyone interested in trying to maximise their own, or their teams, performance. Performance streaks, or seemingly unusual runs of success and failure, have generated considerable interest in this area. Every sports or game player can remember being ‘on a hot run’, or ‘in the zone’. Similar descriptions of streakiness are often used to describe ongoing performance, such as ‘he has the hot hand’. Thus, descriptions of streakiness serve as both a descriptor of the recent past, and a predictor of the short-term future. It is not surprising then that Gilovitch, Vallone, and Tversky (1985) created far-reaching

interest when they validated the strength of people's belief in streaks, and then empirically invalidated their existence.

To begin their investigation, Gilovitch, Vallone, and Tversky (1985) surveyed 100 basketball fans, and found 91% believed that professional players had a better chance of making a shot after having hit their previous two or three shots than after having missed their previous two or three shots. Respondents further believed this benefit to be substantial. For a 50% shooter, fans estimated the chance of making a shot was 61% following a hot sequence, compared to 42% following a cold sequence. Professional basketball players themselves also endorsed the belief. All seven Philadelphia 76ers interviewed believed "it was important to pass the ball to a player who had made several shots in a row" (p. 302). *The hot hand belief* was formalised as the belief that for skilled tasks, the probability of a hit (success) given a hit will be greater than the probability of a hit given a miss (failure).

Gilovitch et al. (1985) examined basketball data and found that despite the beliefs of fans and players, there was no evidence to support the hot hand. A summary of their analyses for the general play shooting of the 1980-81 Philadelphia 76ers is provided in Table 1. A comparison of column 4 and column 6 shows that in contrast to hot hand beliefs, seven of nine players had lower success probabilities following hits. To further support these results, Gilovitch et al. carried out a controlled shooting experiment and an analysis of free throw shooting.[1] Both shooters and observers placed bets on each shot outcome, and while both shooters and observers demonstrated the hot hand belief by placing higher bets on shots following a hit, the

[1] When awarded free throws, a basketball player is required to take two 'free' shots from a fixed position. These shots occur sporadically throughout a game when a shooter is deemed to have been unfairly impeded while taking a shot.

shooting sequences showed no streaky characteristics. Free throw shooting provided similar results. On the basis of these investigations, Gilovitch et al. declared belief in the hot hand a fallacy.

Table 1

Probability of Making a Shot Conditioned on the Outcome of Previous Shots for Nine Members of the 1980-81 Philadelphia 76ers

Player	P(hit|3 misses)	P(hit|2 misses)	P(hit|1 miss)	P(hit)	P(hit|1 hit)	P(hit|2 hits)	P(hit|3 hits)
Clint Richardson	.50 (12)	.47 (32)	.56 (101)	.50 (248)	.49 (105)	.50 (46)	.48 (21)
Julius Erving	.52 (90)	.51 (191)	.51 (408)	.52 (884)	.53 (428)	.52 (211)	.48 (97)
Lionel Hollins	.50 (40)	.49 (92)	.46 (200)	.46 (419)	.46 (171)	.46 (65)	.32 (25)
Maurice Cheeks	.77 (13)	.60 (38)	.60 (126)	.56 (339)	.55 (166)	.54 (76)	.59 (32)
Caldwell Jones	.50 (20)	.48 (48)	.47 (117)	.47 (272)	.45 (108)	.43 (37)	.27 (11)
Andrew Toney	.52 (33)	.53 (90)	.51 (216)	.46 (451)	.43 (190)	.40 (77)	.34 (29)
Bobby Jones	.61 (23)	.58 (66)	.58 (179)	.54 (433)	.53 (207)	.47 (96)	.53 (36)
Steve Mix	.70 (20)	.56 (54)	.52 (147)	.52 (351)	.51 (163)	.48 (77)	.36 (33)
Daryl Dawkins	.88 (8)	.73 (33)	.71 (136)	.62 (403)	.57 (222)	.58 (111)	.51(55)

Note. Parenthetical values = Number of shots, where a shot made after 3 misses is counted in columns 2, 3, and 4, and a shot made after 2 misses is counted in columns 3 and 4, etc. Since the first shot of each game cannot be conditioned, parenthetical values in columns 4 and 6 do not sum to the parenthetical values in column 5. Adapted from "The hot hand in basketball: On the misperception of random sequences," by T. Gilovitch, R. Vallone, and A. Tversky, 1985, Cognitive Psychology, 17, p. 299. Copyright 1985 by the Academic Press, Inc.

Subsequently, the hot hand was enthusiastically researched in a variety of sporting contexts, many of which are reviewed below. While the results of this research have been hotly debated (Bar-Eli, Avugos, & Raab, 2006), recent meta-analyses indicates little evidence in favour of a hot hand effect in professional sports (Avugos, Köppen, Czienskowski, Raab, & Bar-Eli, 2013). Despite this lack of empirical support for a hot hand effect, the hot hand belief has been demonstrated

across a broad range of interests and pursuits (Bar-Eli, Avugos, & Raab, 2006), and has been commonly used to explain sub-optimal decision making in situations where the influence of recent success on future outcomes is overestimated. Examples include the decisions of financial traders (Huber, Kirchler, & Stockl, 2010; Offerman & Sonnemans, 2004), bookmakers in framing markets (Camerer, 1989), gamblers in placing bets (Croson & Sundali, 2005), and investors when selecting mutual funds (Rabin, 2002). In each case sub-optimal decision-making is suggested to arise from a misplaced belief in success following success.

Task Difficulty and the Hot Hand

One notable development in hot hand research has been the consideration of fluctuations in trial-to-trial task difficulty. Larkey, Smith, and Kadane (1989) first suggested that the difficulty of tasks might vary from one attempt to the next, resulting in accuracy measures, but not observers, being insensitive to streaks in performance. Using basketball as an example, Larkey et al. suggested that variables such as defensive pressure, shot distance, and player confidence might increase shot difficulty following success, obscuring streakiness in accuracy measures. In extending this thinking, Smith (2003) also suggested performers, even in the absence of opponents, may systematically attempt more difficult tasks following success.

The consideration of both difficulty and accuracy extends the range of outcomes that may result from hot hand investigations. Figure 1 presents the possible outcomes in previous investigations of the hot hand (panel A), and the possible outcomes when both accuracy and difficulty are considered (panel B). In both panels, p(h|h) marks the probability of a hit following a hit, while p(h|m) marks the probability of a hit following a miss. A hot hand implies p(h|h) > p(h|m) [which can

also be expressed as p(h|h) – p(h|m) > 0]. Thus, cell number 3 in panel A corresponds to a case in which a performer displays increased accuracy following a hit, or the hot hand. Notably in panel B, the performer may display increased accuracy following a hit, or the hot hand, in cells 3, 6, or 9. Interestingly, a finding of the hot hand does not necessarily indicate overall improved performance. For example, cell 9 indicates the hot hand associated with lower task difficulty following a hit, which might represent an occurrence of a difficulty-accuracy trade-off. Using the example of a simple linear model, this trade-off would be represented by a diagonal drawn through cells 1, 5, and 9. It is also illustrative to consider a performer who takes more difficult shots following success, but maintains a consistent level of accuracy. This outcome would fall in cell 2, and thus represent enhanced performance following success, but not the hot hand. This enhanced performance might support an attenuated form of the hot hand belief – increased overall performance following success, or a short spike in ability - yet would not be detected by traditional hot hand measures based on accuracy.

A.

Trial Accuracy		
p(h|h) < p(h|m)	p(h|h) = p(h|m)	p(h|h) > p(h|m)
1 Reverse Hot Hand	2 No Hot Hand	3 Hot Hand

B.

		Trial Accuracy		
		P(h|h) < P(h|m)	P(h|h) = P(h|m)	P(h|h) > P(h|m)
Trial Difficulty	td|h > td|m	1 Increased difficulty and Cold Hand	2 Increased difficulty and no change in accuracy	3 Increased difficulty and Hot Hand
	td|h = td|m	4 No change in difficulty Cold Hand	5 No change in difficulty or accuracy	6 No change in difficulty and Hot Hand
	td|h < td|m	7 Reduced difficulty and Cold Hand	8 Reduced difficulty and no change in accuracy	9 Reduced difficulty and Hot Hand

Figure 1. Panel A shows possible outcomes following success from previous investigations of the hot hand. Improved accuracy is represented in cell 3. Panel B shows possible outcomes when difficulty and accuracy are considered. Cells 3, 6, & 9 represent increased accuracy, or a traditional view of the hot hand. Under a simple linear conception, a diagonal drawn through cells 1, 5, and 9 would represent the difficulty/accuracy trade-off. Of note, Cell 2 represents overall improved performance, but not a finding of the hot hand. In contrast, Cell 9 represents a finding of the hot hand, but lies represents a difficulty/accuracy trade-off and so does not necessarily indicate improved overall performance.
* td = trial difficulty. h = hit or success. m = miss or failure.

Given the additional complexities introduced when difficulty is considered, it is an interesting exercise to independently consider fixed difficulty tasks. In a task where difficulty is fixed from trial to trial (row 2 of Figure 1, Panel B), any difference in performance that may result from success or failure can only be documented in accuracy. Table 2 adopts this novel approach and presents a summary of previous hot hand findings by conditions of 'variable' and 'fixed' difficulty. When organised in this fashion, a pattern emerges. Studies of variable difficulty, with one exception[2], find no evidence to support the hot hand. Studies of fixed difficulty however provide mixed results. In professional sports, streaky performance has been reported in tasks such as horseshoe pitching (Smith, 2003), ten-pin bowling (Dorsey-Palmenter & Smith, 2004), and billiards (Adams, 1995). In experimental studies, Gilden and Wilson (1995) found evidence for streaky golf putting. While in each case effect sizes

[2] In this exception, Klaassen and Magnus (2001) analysed points from Wimbledon singles service games, and task difficulty was again coded by variables such as the skill of the opposing player. Servers were .4% more likely to win a point if they had won the prior point. While statistically significant given a sample size of 89,000 points, the small effect size, given the difficulties of coding difficulty in sports, falls short of providing substantial support for enhanced performance.

remain well short of predictions associated with the hot hand belief, streaky performance seems far more common in fixed difficulty conditions.

Table 2

Summary of Empirical Hot Hand Research in Sports by Control of Task Difficulty

Study	Activity	Data	Point of analysis	Method/Tests	Hot Hand/Streakiness
Variable Difficulty Studies (the trial-to-trial difficulty of tasks may vary)					
Gilovitch et al. (1985) - Study 2	Basketball	Field goal data for 9 Philadelphia 76ers players during the 1980-81 season	Shot-to shot	Conditional probability and runs analysis	No
Tversky and Gilovitch (1989)	Basketball	Field goal data for 18 players across 39 games of the 1987-88 NBA season	Shot-to-shot (close temporal proximity)	As Above	No
Larkey et al. (1989)	Basketball	Field goal data for 18 players in 39 NBA games during the 1987-88 season	Shot-to shot	Varied	No*
Siwoff et al. (1988)	Baseball	All Major League games from the 1984-87 season	Game-to-game	Comparison of batting averages following 5 game cold and hot streaks	No

Albright (1993)	Baseball	Hitting data from 40 Major League players from the 1987-90 seasons	Game-to-game	Regression model including previous streak as predictor	No
Clark (2003a)	Golf	18 hole scores from 35 professional golfers on the 1997-98 PGA and Seniors Tours	Round-to-round	Cluster analysis of par or better rounds	No
Clark (2003b)	Golf	18 hole scores from 25 professional female golfers on the 1997-98 LPGA Tour	Round-to-round	As Above	No
Clark (2005)	Golf	Hole-to-hole scores for 35 professional golfers on 1997 PGA tour	Hole-to-hole	Cluster analysis of par or better holes	No
Klaassen and Magnus (2001)	Tennis	Service games from 481 men's and women's matches, Wimbledon, 1992-95	Point-to-point	Regression model including previous point as predictor	Yes
Fixed Difficulty Studies (the trial-to-trial difficulty of tasks is constant)					
Gilovitch et al. (1985) - Study 3	Basketball	Free throw data for 9 Boston Celtics players during the 1980-82 seasons	Shot-to shot	Conditional probability and runs analysis	No
Gilovitch et al. (1985) - Study 4	Basketball	26 Cornell varsity basketball participants completing one 100 shot sequence	Shot-to-shot (interrupted by predictions)	As above	No

Koehler and Conley (2003)	Basketball	23 shooters from 4 annual NBA 3-point shootout contests (56 sequences)	Shot-to-shot	As above	No
Gilden and Wilson (1995) - Exp. 1	Golf	40 participants completing one 300 putt sequence	Putt-to-putt	Modified runs test	Yes
Gilden and Wilson (1995) - Exp. 2	Golf	5 participants completing 3 (easy, medium, hard), 300 putt sequences	Putt-to-putt	As Above	Mixed
Gilden and Wilson (1995) - Exp. 3	Darts	8 participants completing 3 (easy, medium, hard), 300 throw sequences	Throw-to-throw	As above	No
Smith (2003)	Horseshoe pitching	64 pitchers in the 2000 and 2001 World Championships	Pitch-to-pitch and game-to-game	Conditional probability analysis (hot and cold hand)	Yes
Dorsey-Palmenter and Smith (2004)	Bowling	43 professional bowlers in the 2002/3 PBA season	Bowl-to-bowl	Conditional probability of bowling a strike	Yes
Adams (1995)	Billiards	Professional Players in a 9 Ball Tournament	Game-to-game	Conditional probability of clearing a table	Yes

Note. Studies not relevant to the discrepancy between hot hand beliefs and performance have been excluded from this summary. An exception to this criterion are the studies of Gilden and Wilson (1995), who used non-expert participants and thus mathematically extracted the influence of learning from resultant sequences. This procedure renders traditional conditional probability and runs analysis inappropriate. Their modified runs test provides a suitable approximation of dependence however in a well controlled study.

* While Larkey , Smith, and Kadane (1989) concluded their analysis supported the existence of the hot hand, subsequent re-analysis (Tversky & Gilovitch, 1989) found the streak in question to be miscoded. When coded correctly, this data failed to support streakiness.

It is important to note that in real life scenarios, the difficulty of tasks is rarely tightly controlled, thus the formation of the hot hand belief is unlikely to be heavily influenced by experience of fixed-difficulty conditions. Therefore, evidence supporting streaky performance in conditions of fixed difficulty is neither necessary nor sufficient to support a difficulty-based account of the hot hand belief. Nevertheless, the data presented in Table 2 provides compelling motivation for considering difficulty as an additional dimension in hot hand studies.

In addition, a growing body of evidence suggests that in variable difficulty tasks, success and failure may be associated with systematic changes in the difficulty. For example, Rao (2009) analysed 60 LA Lakers basketball games in the 2007-08 season and reported the majority of players attempted more difficult shots following a successful run. More recently, Bocskocsky Ezekowitz, and Stein (2014) employed enhanced tracking technology and found players on a "hot run" take more shots of higher difficulty, and perform at above expected performance levels if shot difficulty is taken into account. Sporting contexts though allow researchers little control over experimental factors, and so various researchers turned to a more controlled laboratory environment. For example Wilde, Gerszke, and Paulozza (1998) asked participants to tap a series of red squares after their appearance on a computer screen. Responses closest to 1500ms were rewarded the highest points, however responses faster than 1500ms were penalised. In response to a run of point scoring trials, participants adopted successively more risk by making faster taps, however, following a penalty, subsequent taps were significantly slower and less risky. These results

further establish the merits of considering task difficulty when exploring the hot hand belief.

Summary and Transition

Three key points motivated our study of the hot hand, and more generally, sequential effects. Firstly, the hot hand is of interest for a diverse range of groups due to the consistent discrepancy between people's belief and empirical data. In many fields, sub-optimal decision-making is suggested to arise from a misplaced belief in the hot hand effect. Secondly, by distinguishing between fixed and variable difficulty tasks, I established that task difficulty is identifiable as a potentially important dimension of the hot hand that had previously escaped systematic investigation. Lastly, I provided a diverse range of evidence to support the hypothesis that in variable difficulty tasks, performers may systematically alter task difficulty in response to success and failure. These motivating factors led to the development of a controlled experimental game that could measure changes in both task difficulty and accuracy. The development of this game is documented in Paper 1. An overview highlighting important contributions from Paper 1 is provided below.

Paper 1 provides an example of experimental development – in which both computer game design and experimental design principles were utilised. This novelty was driven by our unique requirements. On the one hand, we required our design to be heavily informed by gaming principles. This requirement was necessary to overcome potential criticisms that the game was not in any way similar to a sporting environment – and was therefore was not suitable for exploring the hot hand effect. Such criticisms had been levelled at controlled hot hand experiments previously (Smith, 2003). On the other hand, we required our game to meet strict experimental

design specifications, or else the game would not be suitable to explore the hot hand in a rigorous and scientific manner.

Our review of literature regarding computer-game design highlighted that balancing risk and reward was a key element in the design of engaging games. Adams (2010) acknowledged that this principle was a fundamental rule for designing computer games, and best summed up this position in acknowledging "A risk must always be accompanied by a reward" (p. 23). A well-calibrated risk and reward structure was therefore imperative to provide entertainment value and engage participants. This understanding informed our choice risk-reward design, specifically, that players would aim to maximise the number of successful attempts in a fixed time period. Under this design, players could trade off speed for accuracy, bringing an element of risk and strategy into the game play.

Of course, there was a danger in over-utilising computer gaming principles. Computer games, in a similar manner to sports games and stock markets, are 'noisy' statistical environments. In these domains, many variables contribute to the outcome of an unknown result – which might add to excitement and unpredictability - but also makes it difficult for players, spectators, and researchers, to isolate the risks and rewards of any given choice or event. For us to explore the hot hand in a scientific manner, we required a platform from which risk adopted by players' following success and failure could be precisely measured. Thus, our requirements demanded an exploration of risk and reward from both a game design and cognitive science perspective. Chapter 1 illustrates the iterative, player-centric development of a top-down shooter game that we hoped would be suitable to explore the hot hand effect. Importantly, the top-down shooter game was designed to explore the effect of success and failure on both accuracy and task difficulty.

The success of our cognitive game was threatened by potential flaws derived from both the gaming and experimental perspectives. From a game-design perspective, if one risk level provided substantially more reward than any other, players would learn this reward structure and be unlikely to change strategy throughout play. In the game design literature, this flaw is considered an *exploit*. From an experimental (in this case statistical) perspective, the core challenge was to obtain a large enough number of trials, coupled with a probability of success somewhere in the range of 40-60%, on average. If people failed most of the time, we would not record enough runs of success. If people succeeded most of the time, we would not observe enough runs of failure. This was an experimental, or statistical, constraint placed upon our game design. Interestingly, while this level of success seems low for a psychological experiment, it is precisely what might be expected of a basketball shooter (Erčulj & Štrumbelj, 2015), and was therefore suitable for study of the hot hand.

The ultimate goal for a player of the top-down shooter game was to shoot down as many alien spaceships as possible within a fixed amount of time. On each trial, a single alien ship decelerated at a constant rate, moving more slowly after each pass across the screen, so it became easier to hit as time progressed within a trial. Under this design, players could trade off speed for accuracy, bringing an element of risk and strategy into the game play. If a player adopted more risk on each attempt (and so fired quickly, while the alien ship was still moving rapidly), they would be able to attempt more shots in their fixed time period. However, if a player adopted less risk on each attempt (and so fired slowly), they would be more likely to successfully shoot down each alien, but at the expense of less overall attempts. This meant the number of overall shots made, as well as the number of hits, depend on

player performance and strategy. We suspected this would promote ideal conditions to capture players systematically altering task difficulty following hits and misses. The game as described represented a variable difficulty hot hand task; a player could change the desired level of risk in response to a previous success or failure. It is important to note though that the game was coded such that the rate at which the alien ships decelerated was controlled by a single parameter. If this deceleration parameter was set to zero, the alien shooter game quickly became a fixed-difficulty task (because the alien would not slow down over time, and thus difficulty was constant).

Most importantly, we adopted the iterative design process supported by best-practice game-design methods, and we drove this iterative process with empirical data, consistent with best-practice principles of experimental cognitive psychology.

Given we were designing the game to explore cognitive processes, our data-driven changes were crucial because it was unlikely that qualitative feedback would have allowed us to make the required changes to game mechanics. As an example, this process helped us discover that players tended to explore shooting on different passes in the practice block, but settled for a single pass (i.e., in a fixed point in time; for example, always on the fifth time the alien ship crosses the screen horizontally) in game blocks. We referred to this strategy in Paper 1 as *investment.* That is, players invested in learning to shoot on a particular pass, and they then chose to exploit this learning rather than risk making an attempt with the alien travelling at an unfamiliar speed. In gaming terminology, once a player had well learned her shot timing for a certain pass, shooting on this pass became an exploit. The issue of exploits in games is often debated in gaming circles and is also well studied in psychology (e.g., Hills, Todd, & Goldstone, 2008; Walsh, 1996). Of course, an exploit that strongly

encouraged players to always fire on the same pass meant that we could not achieve our objective: to assess whether players might systematically shift between passes.

Data driven game development also allowed us to model and remodel our reward structure based on precise measures of player performance. In stages one and two players were penalised by waiting 1.5 seconds each time they missed an alien. In stage three we reduced this penalty to 0.25 seconds based on our analysis and modelling of player behaviour. This relatively minor change was enough to modify players' behaviour and encourage earlier shots at the alien spacecraft. Our game was quite simple in nature, yet as these examples illustrate our data driven approach highlighted key problems with our initial design. Ultimately, we felt the game design at which we arrived was suitable to investigate hot hand phenomena. Unfortunately, further testing highlighted that alterations were required to our game. The Alien shooter was therefore a valuable development stage, but did not reflect our final experimental paradigm. This additional development is documented in Chapter 2.

Game Studies. Volume 11, Issue 1. February 2011.

Balancing Risk and Reward to Develop an Optimal Hot-Hand Game

by Paul Williams, Keith V. Nesbitt, Ami Eidels, David Elliott

Abstract
This paper explores the issue of player risk-taking and reward structures in a game designed to investigate the psychological phenomenon known as the 'hot hand'. The expression 'hot hand' originates from the sport of basketball, and the common belief that players who are on a scoring streak are in some way more likely to score on their next shot than their long-term record would suggest. That is, they are on a 'hot streak', or have the 'hot hand'. There is a widely held belief that players in many sports demonstrate such streaks in performance; however, a large body of evidence discredits this belief. One explanation for this disparity between beliefs and available data is that players on a successful run are willing to take greater risks due to their growing confidence. We are interested in investigating this possibility by developing a top-down shooter. Such a game has unique requirements, including a well-balanced risk and reward structure that provides equal rewards to players regardless of the tactics they adopt. We describe the iterative development of this top-down shooter, including quantitative analysis of how players adapt their risk taking under varying reward structures. We further discuss the implications of our findings in terms of general principles for game design.

Key Words: risk, reward, hot hand, game design, cognitive, psychology

Introduction

Balancing risk and reward is an important consideration in the design of computer games. A good risk and reward structure can provide a lot of additional entertainment value. It has even been likened to the thrill of gambling (Adams, 2010, p. 23). Of course, if players gamble on a strategy, they assume some odds, some amount of risk, as they do when betting. On winning a bet, a person reasonably expects to receive a reward. As in betting, it is reasonable to expect that greater risks will be compensated by greater rewards. Adams not only states that "A risk must always be accompanied by a reward" (2010, p. 23) but also believes that this is a fundamental rule for designing computer games.

Indeed, many game design books discuss the importance of balancing risk and reward in a game:

- "The reward should match the risk" (Thompson, 2007, p.109).
- "... create dilemmas that are more complex, where the players must weigh the potential outcomes of each move in terms of risks and rewards" (Fullerton, Swain, & Hoffman, 2004, p.275).
- "Giving a player the choice to play it safe for a low reward, or to take a risk for a big reward is a great way to make your game interesting and exciting" (Schell, 2008, p.181).

Risk and reward matter in many other domains, such as stock-market trading and sport. In the stock market, risks and rewards affect choices among investment options. Some investors may favour a risky investment in, say, nano-technology stocks, since the high risk is potentially accompanied by high rewards. Others may be more conservative and invest in solid federal bonds which fluctuate less, and therefore offer less reward, but also offer less risk. In sports, basketball players sometimes take more difficult and hence riskier shots from long distance, because these shots are worth three points rather than two.

Psychologists, cognitive scientists, economists and others are interested in the factors that affect human choices among options varying in their risk-reward structure. However, stock markets and sport arenas are 'noisy' environments, making it difficult (for both players and researchers) to isolate the risks and rewards of any given event. Computer games provide an excellent platform for studying, in a well-controlled environment, the effects of risk and reward on players' behaviour.

We examine risk and reward from both a cognitive science and game design perspectives. We believe these two perspectives are complementary. Psychological principles can help inform game design, while appropriately designed games can provide a useful tool for studying psychological phenomena.

Specifically, in the current paper we discuss the iterative, player-centric development (Sotamma, 2007) of a top-down shooter that can be used to investigate the psychological phenomenon known as the 'hot hand'. Although the focus of this paper is on the process of designing risk-reward structures into a game to clearly understand the design requirements of a hot hand game, we begin with an overview of this phenomenon and the current state of research. In subsequent sections we describe three stages of game design and development. In our final section we relate our findings back to more general principles of game design.

The Hot Hand

The expression 'hot hand' originates from basketball and describes the common belief that players who are on a streak of scoring are more likely to score on their next shot. That is, they are on a hot streak or have the 'hot hand'. In a survey of 100 basketball fans, 91% believed that players had a better chance of making a shot after hitting their previous two or three shots than after missing their previous few shots (Gilovitch, Vallone, & Tversky, 1985).

While intuitively these beliefs and predictions seem reasonable, seminal research found no evidence for the hot hand in the field-goal shooting data of the 1980-81 Philadelphia 76ers, or the free-throw shooting data of the 1980-81 and 1981-82 Boston Celtics (Gilovitch et al., 1985). With few exceptions, subsequent studies across a range of sports confirm this surprising finding (Bar-Eli, Avugos, & Raab, 2006) - suggesting that hot and cold streaks of performance could be a myth.

However, results of previous hot hand investigations reveal a more complicated picture. Specifically, previous studies suggest that a distinction can be made between tasks of 'fixed' difficulty and tasks of 'variable' difficulty. A good example of a 'fixed' difficulty task is free-throw shooting in basketball. In this type of shooting the distance is kept constant, so each shot has the same difficulty level. In a 'variable' difficulty task, such as field shooting during the course of a basketball game, players may adjust their level of risk from shot-to-shot, so the

difficulty of the shot varies depending on shooting distance, the amount of defensive pressure, and the overall game situation.

Evidence suggests it is possible for players to get on hot streaks in fixed difficulty tasks such as horseshoe pitching (Smith, 2003), billiards (Adams, 1996), and ten-pin bowling (Dorsey-Palmenter & Smith, 2004). In variable difficulty tasks, however, such as baseball (Albright, 1993), basketball (Gilovitch et al., 1985), and golf (Clark, 2003a, 2003b, 2005), there is no evidence for hot or cold streaks - despite the common belief to the contrary.

The most common explanation for the disparity between popular belief (hot hand exists) and actual data (lack of support for hot hand) is that humans tend to misinterpret patterns in small runs of numbers (Gilovitch et al., 1985). That is, we tend to form patterns based on a cluster of a few events, such as a player scoring three shoots in a row. We then use these patterns to help predict the outcome of the next event, even though there is insufficient information to make this prediction (Tversky & Kahneman, 1974). In relation to basketball shooting, after a run of three successful shots, people would incorrectly believe that the next shot is more likely to be successful than the player's long term average. This is known as the hot-hand fallacy.

A different explanation for this disparity suggests shooters tend to take greater risks during a run of success, for no loss of accuracy (Smith, 2003). Under this scenario, a player does show an increase in performance during a hot streak - as they are performing a more difficult task at the same level of accuracy. This increase in performance may in turn be reflected in hot hand predictions, however would not be detected by traditional measures of performance. While this hypothetical account receives tentative support by drawing a distinction between fixed and variable difficulty tasks (as the hot hand is more likely to appear in fixed-difficulty tasks, where players cannot engage in a more difficult shot), this hypothesis requires further study.

Unfortunately, trying to gather more data to investigate the hot hand phenomenon from sporting games and contests is fraught with problems of subjectivity. How can one assess the difficulty of a given shot over another in basketball? How can one tell if a player is adopting an approach with more risk?

An excellent way to overcome this problem is to design a computer game of 'variable' difficulty tasks that can accurately record changes in player strategies. Such a game can potentially answer a key question relevant to both psychology and game design - how do people (players) respond to a run of success or failure (in a game challenge)?

The development of this game, which we call a 'hot hand game', is the focus of this paper. Such a game requires a finely tuned risk and reward structure, and the process of tuning this structure provides a unique empirical insight into players risk taking behaviour. At each stage of development we test the game to measure how players respond to the risk and reward structure. We then analyse these results in terms of player strategy and performance and use this analysis to inform our next stage of design.

This type of design could be characterised as iterative and player-centric (Sotamaa, 2007). While the game design in this instance is simple, due to the precise requirements of the psychological investigation, player testing is more formal than might traditionally be used in game development. Consequently, changes in player strategy can be precisely evaluated. We

find that even subtle changes to risk and reward structures impact on player's risk-taking strategy.

Game Requirements and Basic Design

A hot hand game that addresses how players respond to a run of success or failure has special requirements. First and foremost, the game requires a finely-tuned risk and reward structure. The game must have several (5-7), well-balanced risk levels, so that players are both able and willing to adjust their level of risk in response to success and failure. If, for example, one risk level provides substantially more reward than any other, players will learn this reward structure over time, and be unlikely to change strategy throughout play. We would thus like each risk level to be, for the average player, equally rewarding. In other words, regardless of the level of risk adopted, the player should have about the same chance of obtaining the best score.

The second requirement for an optimal hot hand game is that it allows measurement of players' strategy after runs of both successes and failures. If people fail most of the time, we won't record enough runs of success. If people succeed most of the time, we won't observe enough runs of failure. Thus, the core challenge needs to provide a probability of success, on average, somewhere in the range of 40-60%.

The game developed to fulfil these requirements was a top-down shooter developed in Flash using Actionscript. While any simple action game based on a physical challenge with hit-miss scoring could be suitably modified for our purposes, a top-down shooter holds several advantages. Firstly, high familiarity with the style means the learning period for players is minimal, supporting our aims of using the game for experimental data collection. Secondly, the simple coding of key difficulty parameters (i.e. target speeds and accelerations) allows the reward structure to be easily and precisely manipulated. Lastly, a 'shot' of a top-down shooter is analogous to a 'shot' in basketball, with similar outcomes of 'hit' and 'miss'. This forms a clear and identifiable connection between the current experiment and the origins of the hot hand.

In the top-down shooter, the goal of the player is to shoot down as many alien spaceships as possible within some fixed amount of time. This means the number of overall shots made, as well as the number of hits, depend on player performance and strategy. However, the games duration is fixed. The game screen shows two spaceships, representing an alien and the player-shooter (Figure 1). The simple interface provides feedback about the current number of kills and the time remaining. During the game the player's spaceship remains stationary at the bottom centre of the screen. Only a single alien spaceship appears at any one time. It moves horizontally back-and-forth across the top of the screen, and bounces back each time it hits the right or left edges. The player shoots at the alien ship by pressing the spacebar. For each new alien ship the player has only a single shot with which to destroy it. If an alien is destroyed the player is rewarded with a kill.

Figure 1: The playing screen.

Each alien craft enters from the top of the screen and randomly moves towards either the left or right edge. It bounces off each side of the screen, moving horizontally and making a total of eight passes before flying off. Initially the alien ship moves swiftly, but it decelerates at a constant rate, moving more slowly after each pass. This game therefore represents a variable difficulty task; a player can elect a desired level of risk as the shooting task becomes less difficult with each pass of the alien.

The risk and reward equation is quite simple for the player. The score for destroying an alien is the same regardless of when the player fires. Since the goal is to destroy as many aliens as possible in the game period, the player would benefit from shooting as quickly as possible; shooting in the early passes rewards the player with both a kill and more time to shoot at subsequent aliens. However, because the alien ship decelerates during each of the eight passes, the earlier a player shoots the less likely this player will hit the target. If a shot is missed, the player incurs a 1.5 second time penalty. That is, the next alien will appear only after a 1.5 second delay which is additional to the interval experienced for an accurate shot.

Stage One--Player Fixation

After self-testing the game, we deployed it so that it could be played online. Five players were recruited via an email circulated to students, family and friends. Players were instructed to shoot down as many aliens as possible within a given time block. They first played a practice level for six minutes before playing the competitive level for 12 minutes. The number of alien ships a player encountered varied depending on the player's strategy and accuracy. A player could expect to encounter roughly 10 alien ships for every 60 seconds of play. At the completion of the game the player's response time and accuracy were recorded for each alien ship.

Recall that one of the game requirements was that players take shots across a range of difficulty levels, represented by passes (later passes mean less difficult shots)--this simple test provides evidence that a player is willing to explore the search space and alter her or his risk-taking behaviour throughout the game. Typical results for Players one and two are shown in

Figure 2. In general players tended to be very exploratory during the practice level of the game, as indicated by a good spread of shots between alien passes one and eight. During the competitive game time however players tended to invest in a single strategy, as indicated by the large spikes seen in the competition levels of Figure 2. This suggests that players, after an exploratory period, attempted to maximise their score by firing on a single, fixed pass.

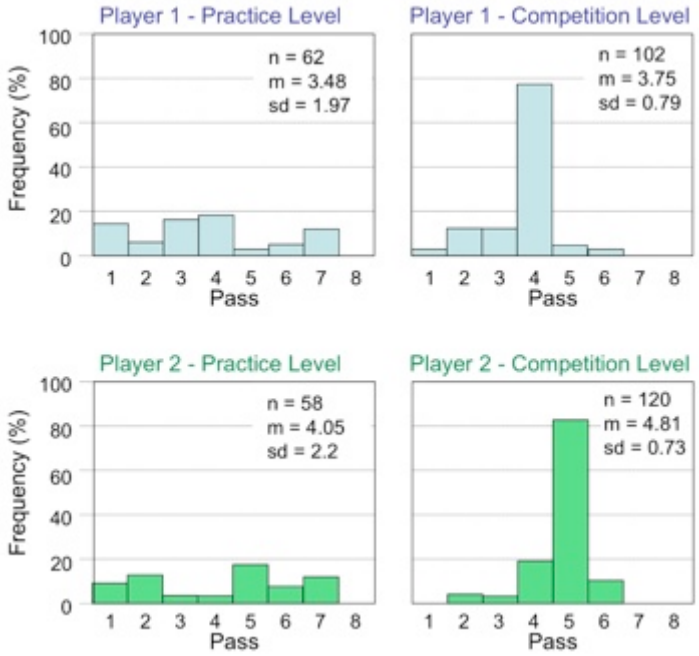


Figure 2: Results for two typical players in Stage one of game development. The upper row shows data for Player 1, and the bottom row shows data for Player 2. The left column presents the frequency (%) of shots taken on each pass in the practice level, while the right column indicates the frequency (%) of shots taken on each pass in the competition level. Note that players experimented during the practice level, as evidenced by evenly spread frequencies across passes in the left panels, but then adopted a fixed strategy during the competitive block, as evidenced by spikes at pass 4 (Player 1) and pass 5 (Player 2). For each panel, n is the overall number of shots attempted by the player in that block, m is the mean firing pass, and sd is the standard deviation of the number of attempted shots.

In experimental terms, this fixation on a single strategy is known as ‘investment’. At the end of the game the players reported that, because of the constant level of deceleration, they could always shoot when the alien was at a specific distance from the wall if they stuck to the same pass. Players thus practiced a timing strategy specific to a particular alien pass (i.e., a specific difficulty level). The number of kills per unit time (i.e., the reward) was therefore always highest for that player when shooting at the same pass. In the example graphs (Figure 2), one player ‘invested’ in learning to shoot on pass four, the other, on pass five. This type of investment runs counter to one requirement from a hot hand game, creating a major design flaw that needed to be fixed in the next iteration.

Stage Two--Encouraging Exploratory Play

The aim of the second stage of design was to overcome the problem of player investment in a single strategy. The proposed solution was to vary the position of the player's ship so that it no longer appeared in the same location at the centre of the screen but rather was randomly shifted left and right of centre each time a new alien appeared (Figure 3). Thus, on each trial, the shooter's location was sampled from a uniform distribution of 100 pixels to the left or to the right of the centre. This manipulation was intended to prevent the player from learning a single timing-sequence that was always successful on a single pass (such as always shooting on pass four when the alien was a certain distance from the side of the screen).

Figure 3: The screen in Stage two of game development. The blue rectangle appears here for illustration purposes and indicates the potential range of locations used to randomly position the player's ship. It did not appear on the actual game screen.

Once again we deployed an online version of the game and recorded data from six players. Players once again played a practice level for six minutes before they played the competitive level for 12 minutes.

The results for all individual players in the competitive game level are shown in Figure 4. Introducing random variation into the players firing position significantly decreased players' tendency to invest in and fixate on a single pass. This decrease in investment is highlighted by the increase in the variance seen in Figure 4 when compared to Figure 2. Thus, the slight change in gameplay had a significant effect on players' behaviour, encouraging them to alter their risk-taking strategy throughout the game. Furthermore, this change helps to meet the requirements necessary for hot hand investigation.

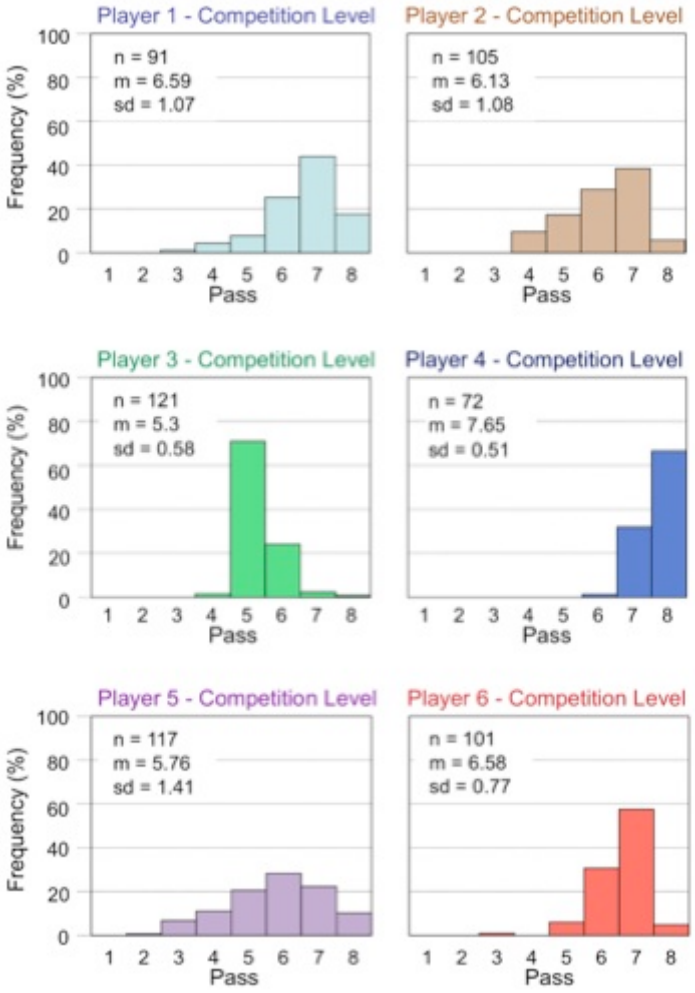


Figure 4: Individual player results for the competition level in Stage two testing. Player's tendency to fire on a single pass in the competition level has been significantly reduced compared to Stage One, as evidenced by the reduction in spikes and, in most cases, increase in variance. For each panel, n is the overall number of shots attempted by the player in that block, m is the mean firing pass, and sd is the standard deviation of the number of attempted shots.

In Figure 5 we present data averaged across all players for both the practice and competition levels. This summary highlights how the game's reward structure influenced player strategy throughout play. The left column corresponds to the practice level (not shown in Figure 4), while the right column corresponds to the competition level.

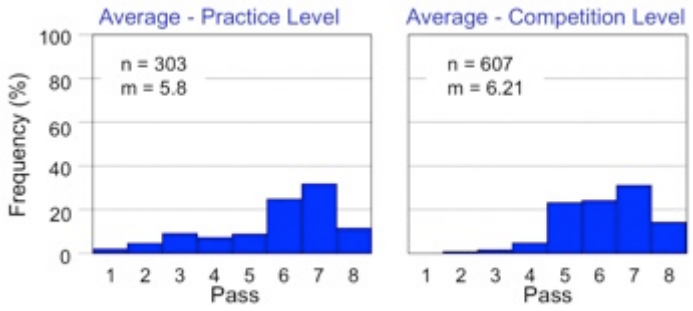


Figure 5: Average player results for Stage two. The left column presents the frequency (%) of shots taken on each pass in the practice level, while the right column indicates the frequency (%) of shots taken on each pass in the competition level. For each panel, m is the mean firing pass and n is the overall number of shots attempted by all players in that block. A comparison of mean firing pass for practice and competition levels highlights that as the game progressed, players fired later.

An inspection of Figure 5 highlights the fact that players' shooting strategy altered in a predictable manner as the game progressed. For example, the mean firing pass for the practice level (m = 5.8) was smaller than that seen in the competition level (m = 6.21). Thus players tended to shoot later in the competition level. This suggests that the reward structure of the game was biased towards firing at later passes, and that as players became familiar with this reward structure they altered their gameplay accordingly.

Given the need to minimise such bias for hot hand investigation, we examined the risk and reward structure on the basis of average player performance. We were particularly interested in the probability of success for each pass, and how this probability translated into our reward system. Recall that firing on later passes takes more time but is also accompanied by a higher likelihood for success. As the aim of the hot hand game is to kill as many aliens as possible within a 12 minute period, both the probability of hits as well as the time taken to achieve these hits are important when considering the reward structure.

We therefore analysed how many kills per 12-minute block the average, hypothetical player would make if he or she were to consistently fire on a specific pass for each and every alien that appeared. For example, given the observed likelihood of success on pass one, how many kills would a player make by shooting only on pass one? How many kills on pass two, and so on. Results of this examination are reported in Figure 6. Figure 6A shows the average number of shots taken by players on each pass of the alien (overall height of bar) along with the average number of hits at each pass (height of yellow part of the bar). Figure 6B uses this data to plot the observed probability of success and shows that the probability for success is higher for later passes. This empirically validates that later passes are in fact 'easier' in a psychological sense.

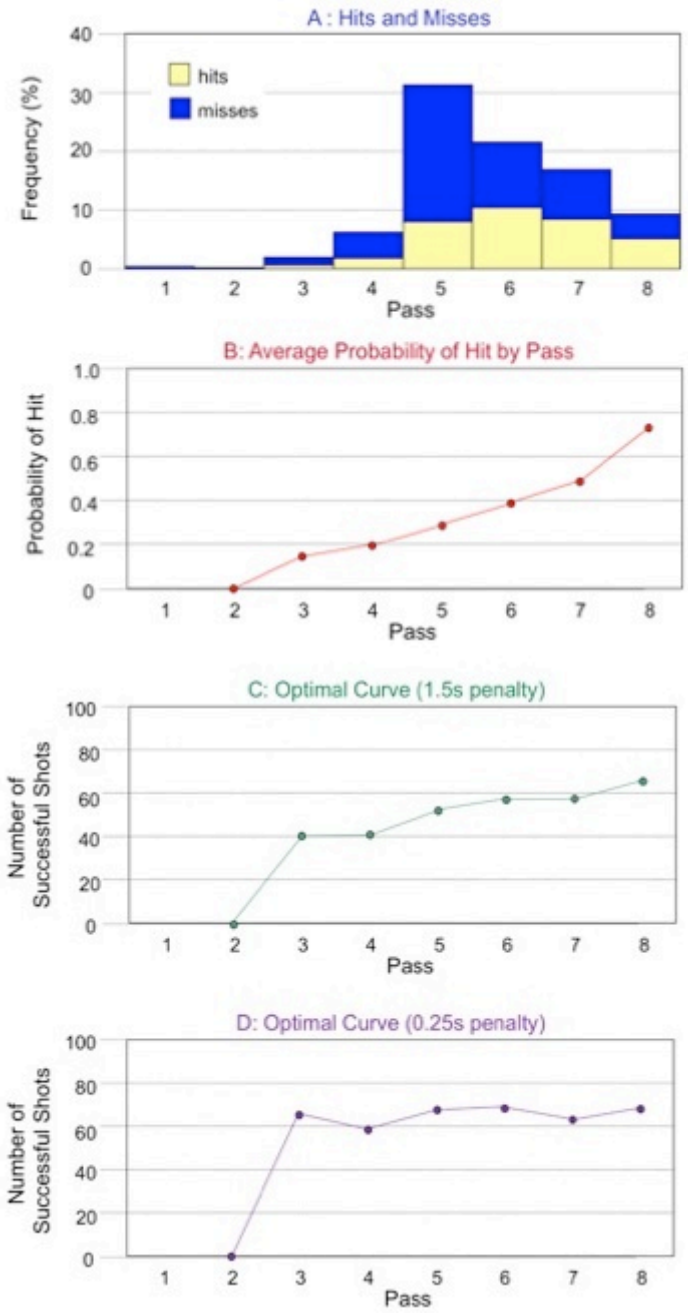


Figure 6: Averaged results and some modelling predictions from Stage two of game development. In Panel A, the frequency (%) of shots attempted on each pass is indicated by the overall height of each bar. The proportion of hits and misses are indicated in yellow and blue. Panel B depicts the average probability of a hit for each pass, given by the number of hits out of overall shot attempts. Based on the empirical results, Panels C and D show the predicted number of successful shots if players were to consistently shoot on only one pass for the entire game (see text for details).

These probabilities allow empirical estimation of the number of total kills likely to be attained by the hypothetical average player if they were to shoot on only one pass for an entire 12 minute block. By plotting the number of total kills expected for each pass number, we produce an optimal strategy curve for the current game, as shown in Figure 6C. The curve is

monotonically increasing, indicating that the total number of kills expected of an average player increases as the pass number increases. In other words, players taking less difficult shots are expected to make more hits within each game. The reward structure is clearly biased toward later passes, which validates the change in player strategy (i.e. firing on later passes) as the game progressed. As the players became accustomed to the reward structure, their strategy shifted accordingly to favour later, easier shots.

In game terms it might be considered an exploit to shoot on pass eight. Figure 6C indicates that consistently firing on pass 8 would clearly result in the greatest number of kills, making it the 'optimal' strategy for the average player. Given that an exploit of this kind reduces the likelihood of players to fire earlier in response to a run of successful shots, the current design still failed to meet the requirements for our hot hand game.

One simple adjustment to overcome this issue was to reduce the penalty period after an unsuccessful shot. While the current time penalty for a missed shot was set to 1.5 seconds, the ability to vary this penalty allows a deal of flexibility within the reward structure. Given that players make many more shots, and thus many more misses, if they choose to fire on early passes - decreasing the time penalty for a miss substantially increases the relative reward for firing on early passes.

In line with this thinking, Figure 6D shows the predicted number of kills in 12 minutes for the average player if the penalty for missing is reduced from 1.5 seconds to 0.25 seconds. This seemingly small change balances the reward structure so that players are more evenly rewarded, at least for passes three to eight. Estimation of accuracy rate on passes one and two were based on a small number of trials, which makes them problematic for modelling; participants avoided taking early shots, perhaps because the alien was moving too fast for them to intercept. Allowing for players to fire on passes three to eight still provided us with sufficient number of possible strategies for a hot hand investigation.

Stage Three--Balancing Risk and Reward

In stage two of our design we uncovered an exploitation strategy in the risk and reward structure of the game where players could perform optimally by shooting on pass eight of the alien. We suspect this influenced players to fire at later passes of the alien, particularly as the game progressed. Using empirical data to model player performance suggested that reducing the time penalty for a miss to 0.25 seconds would overcome this problem.

A modified version of the game, with a 0.25 seconds penalty after a miss, was made available online and data were recorded from five players. Averaged results show that players shot at roughly the same mean pass of the alien in the practice level and the competitive level (Figure 7). This pattern is in contrast with Figure 4, which highlighted a tendency for players to fire at later passes in the 12 minute competitive level. This data confirms the empirical choice of a 0.25 second penalty, and provides yet another striking example of how subtle changes in reward structure may influence players' behaviour.

Recall that we began the development of a hot-hand game with the requirement that for each level of assumed risk the game should be equally rewarding (total number of kills) for the average player. By balancing the reward structure, the design from stage three is now consistent with this requirement for investigating the hot hand.

Finally, we required the game to have an overall level of difficulty such that players would succeed on about 40-60 percent of attempts. Performance within this range would allow us to compare player strategy in response to runs of both success and failure. That is, testing for both hot and cold streaks. As highlighted by Figure 8, the overall probability of success does indeed meet this criteria--the overall probability of success (hits) was 43%. Thus, the game now meets the essential criteria required to investigate the hot hand phenomenon.

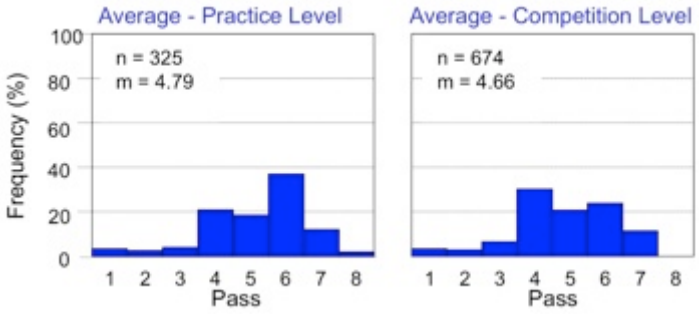


Figure 7: Average player results for Stage three of game development. The left plot presents the frequency (%) of shots attempted on each pass in the practice level, while the right plot indicates the frequency (%) of shots attempted on each pass in the competition level. For each panel, m is the mean firing pass and n is the overall number of shots taken by all players in that block. As indicated by the mean firing pass, under a balanced reward structure players no longer attempted to shoot on later passes as the game progressed.

Recall that we began the development of a hot-hand game with the requirement that for each level of assumed risk the game should be equally rewarding (total number of kills) for the average player. By balancing the reward structure, the design from stage three is now consistent with this requirement for investigating the hot hand.

Finally, we required the game to have an overall level of difficulty such that players would succeed on about 40-60 percent of attempts. Performance within this range would allow us to compare player strategy in response to runs of both success and failure. That is, testing for both hot and cold streaks. As highlighted by Figure 8, the overall probability of success does indeed meet this criteria; the overall probability of success (hits) was 43%. Thus, the game now meets the essential criteria required to investigate the hot hand phenomenon.

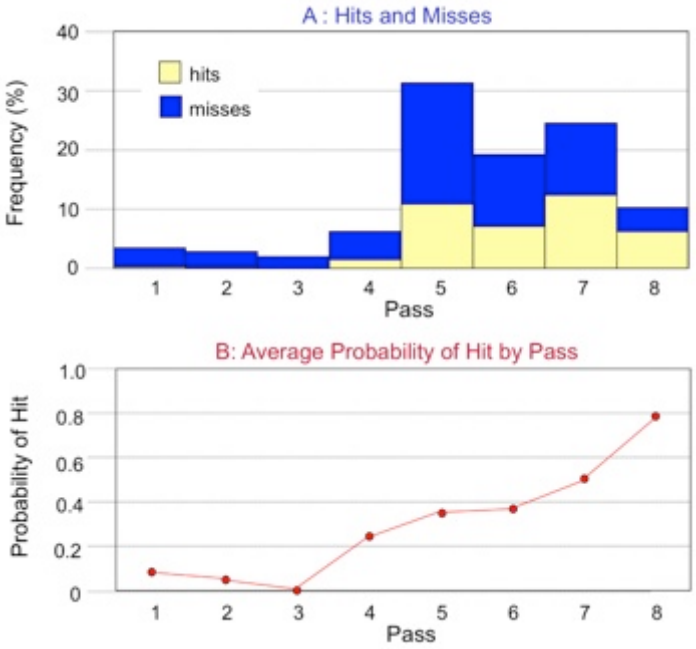


Figure 8: Averaged results from the competition level of Stage three of game development. In Panel A, the frequency (%) of shots attempted on each pass is indicated by the overall height of each bar. The proportion of hits and misses are indicated in yellow and blue. Panel B depicts the average probability of success for each pass, given by the number of hits out of overall shot attempts. In Panel B, ps is the overall probability of success (hits).

Discussion

We set out to design a computer game as a tool for studying a fascinating and widely studied psychological phenomenon called the 'hot hand' (e.g., Gilovitch, Valone, & Tversky, 1985). For this we needed a game that allowed us to investigate player risk-taking in response to a string of successful or unsuccessful challenges.

We designed a simple top-down shooter game where players had a single shot at an alien spacecraft as it made eight passes across the screen. During the game the player faced this same challenge a number of times. The goal of the game was to kill as many aliens as possible in a set amount of time. The risk in the gameplay reduced on each pass as the alien ship slowed down. Shooting successfully on earlier passes rewarded the player with a kill and made a new alien appear immediately. Missing a shot penalised the player with an additional wait time before the next alien appeared.

As a hot hand game it was required to meet specific risk and reward criteria. Players should explore a range of risk-taking strategies in the game and they should be rewarded in a balanced way commensurate with this risk. We also wanted the game challenge to have an average success rate roughly equal to the failure rate, between 40 and 60 percent so that we could use the game to gather data about player's behaviour in response to both success and failure.

To achieve our objective we developed the game in an iterative fashion over three stages. At each stage we tested an online version of the game, gathering empirical data and analysing the players' strategy and performance. In each successive stage of design we then altered the game mechanics so they were balanced in a way that met our specific hot hand requirements. The design changes and their effects are summarised in Table 1.

Stage	Requirements
Design Problems	• Design must meet specific requirements for experimenting with hot hands • Players must be willing to adjust their level of risk throughout the game, and thus explore a range of possible strategies • Must allow the measurement of a players' strategy after runs of both successes and failures by ensuring the game challenge has an average success rate around 50%
Stage	**One**
Design Problems	• Players invest in a single strategy rather than exploring a range of options
Design Solutions	• Changed mechanics (add variability to player's ship position)
Stage	**Two**
Design Effects	• Random location of ship removed investment strategy
Design Problems	• The reward structure favours an exploitation strategy of firing on pass 8 (or as late as possible).
Design Solutions	• Reduce missed time penalty from 1.5 sec to 0.25 sec based on average probability of success
Stage	**Three**
Design Effects	• Players rarely attempt shots on pass 1-2 but the game elicits a good range of firing on passes 3-8 • Success rate for the challenge is within 40-60% range • Game is suitable for our hot hands experiment

Table 1. A summary of changes to design in each of the stages and the effect of these changes on meeting the hot hand requirements.

Books on game design tend to prescribe an iterative design process. Iterative processes allow unforseen problems to be addressed in successive stages of design. This is especially important in games where the requirements for the game mechanics are typically only partially known and tend to emerge as the game is built and played. Salen and Zimmerman describe this iterative process as "play-based" design and also emphasise the importance of "playtesting and prototyping" (2004, p. 4). For this purpose successive prototypes of the game

are required. Indeed we began with only high-level requirements and used this same iterative, prototyping approach to refine our gameplay.

The main difference in our approach is that we more formally measured player's strategies and exploration behaviours in each stage of design. Given that our game requirements are rather unique, it is unlikely that subjective feedback alone would have allowed us to make the required subtle changes to game mechanics. For example, during the initial testing of the game we found that players tended to invest in a single playing strategy. Further analysis also revealed a potential exploit in the game as players could easily optimise their total number of kills by shooting on the last pass of each alien ship.

The issue of exploits in games is often debated in gaming circles and is also well studied in psychology. Indeed trade-offs between exploitation and exploration exist in many domains (e.g., Hills, Todd, & Goldstone, 2008; Walsh, 1996). External and internal conditions determine which strategy the organism, or the player, will take in order to maximise gains and minimise loses. For example, when foraging for food, the distribution of resources matters. Clumped resources lead to a focused search in the nearby vicinity where they are abundant (exploitation), whereas diffused resources lead to broader exploration of the search space.

Hills et al. showed that exploration and exploitation strategies compete in mental spaces as well, depending on the reward for desired information and the toll incurred by search time for exploration. In the context of our game, a shooting strategy of consistently attempting the easiest shooting level (pass 8) produced the highest reward. This encouraged players to drift toward later firing as the game progressed, and in turn inhibited players from exploring alternate (earlier firing) strategies. It is unlikely we could have predicted this without collecting empirical data from players.

A further advantage of gathering empirical data was that it allowed us to remodel our reward structure based on precise measures of player performance. In stages one and two players lost 1.5 seconds each time they missed an alien. In stage three we reduced this penalty to 0.25 seconds based on our analysis and modelling of player behaviour. This relatively minor change was enough to change players' behaviour and encourage them to risk earlier shots at the alien. The fact that our game is quite simple in nature reinforces both the difficulty and importance of designing a well-balanced risk and reward structure.

Another common principle referred to in game literature is player-centred design which is defined by Adams as "a philosophy of design in which the designer envisions a representative player of a game the designer wants to create." (2010, p. 30). Although player-centred design is often a common principle referred to in game-design texts there is some suggestion that design is often based purely on designer experience (Sotamaa, 2007). Involving players in the design process typically involve more subjective feedback from approaches such as focus groups and interviews which have been generally used in usability design. In our study, when designing even a simple game challenge it is clear that the use of empirical data to measure how players approach the game and how they perform can be another vital element in balancing the gameplay.

We also recognise some dangers with this approach, as averaging player performance can hide important differences between players. It would be nice to have a model of an ideal player but it is unlikely such a player exists. In fact there are many different opinions about who the 'player' is (Sotamaa, 2007). The empirical data therefore need to be gathered from

the available players' population. If there are broad differences among these players then it may require the designer to sample different groups, for example, a group of casual players and a group of hard-core gamers.

Importantly for future research, the game design at which we arrived is now suitable to investigate the hot hand phenomena. Such a game can potentially answer a number of questions:

1. How do players respond to a run of success or failure in a game challenge?

2. Will a player take on more difficult challenges if they are on a hot streak?

3. Will they lower their risk if they are on a cold streak?

4. How will this variable risk level impact on their overall measure of performance?

5. How can the hot hand principle be used in the design of game mechanics?

Answers to such questions will not only be of interest to psychologists, but could also further inform game design. For example, it might allow the designer to engineer a hot streak so that players would take more risks or be more explorative in their strategies. Of course in a game it might even be appropriate to use a cold streak to discourage a player's current strategy. The game mechanics could help engineer these streaks in a very transparent way without breaking player immersion. Further investigations of the hot hand hold significant promise for both psychology and game design.

References

Adams E. (2010). *Fundamentals of Game Design* (2nd Edition). New Riders, Berkley, CA, USA.

Adams, R. M. (1996). Momentum in the performance of professional tournament pocket billiards players. *International Journal of Sport Psychology*, 26, 580-587.

Albright, S. C. (1993). A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, 88(424), 1175-1183.

Bar-Eli, M., Avugos, S., & Raab, R. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, 7, 525-553.

Clarke, R. D. (2003a). Streakiness among professional golfers: Fact or fiction? *International Journal of Sports Psychology*, 34, 63-79.

Clarke, R. D. (2003b). An analysis of streaky performance on the LPGA tour. *Perceptual and Motor Skills*, 97, 365-370.

Clarke, R. D. (2005). Examination of hole-to-hole streakiness on the PGA tour. *Perceptual and Motor Skills*, 100, 806-814.

Dorsey-Palmateer, R., & Smith, G. (2004). Bowlers' hot hands. *The American Statistician*, 58(1), 38-45.

Gilden, D. L., & Wilson, S. G. (1995). Streaks in skilled performance. *Psychonomic Bulletin and Review*, 2(2), 260-265.

Gilovitch, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 295-314.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: evidence for generalized cognitive search processes. *Psychological Science*, 19, 802-808.

Koehler, J. J., & Conley, C. A. (2003). The "Hot Hand" myth in professional basketball. *Journal of Sport and Exercise Psychology*, 25, 253-259.

Fullerton, T., Swain, C., and Hoffman, S. (2004) *Game Design Workshop: Designing, Prototyping, and Playtesting Games*. CMP Books. USA.

Salen, K. and Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. The MIT Press, USA.

Schell, J. (2008) *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann Publishers. USA.

Smith, G. (2003). Horseshoe pitchers' hot hands. *Psychonomic Bulletin and Review*, 2003(10), 3.

Sotamaa O. (2007) Perceptions of Player in Game Design Literature. *Proceedings of DiGRA 2007 Conference*. pp. 456-465.

Thompson, J. (2007) *The Computer Game Design Course Thames and Hudson*, London, UK.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Walsh, P. D. (1996). Area restricted search and the scale dependence of patch quality discrimination. *Journal of Theoretical Biology*, 183, 351-361.

Chapter 2

In Chapter 1 the motivation for exploring sequential effects was outlined, specifically the hot hand effect. I then documented the development of a top-down alien shooter game that was engineered to explore - both in terms of accuracy and task difficulty - the hot hand effect in a controlled environment. Chapter 2 extends on this work and documents the continued development of our cognitive game paradigm. Chapter 2 contains two components. The first component is the *Paper 2 Overview and Additional Material*. I recommend Paper 2 be read before, or in conjunction with this section. The Paper 2 Overview and Additional material is broken into three sub-headings, which are introduced in more detail below. To end the chapter, *Paper 2* is presented in full.

Paper 2 Overview and Additional Material

This *Overview and Additional Material* has 3 sub-sections. An understanding of Paper 2 is critical in understanding the organisation of these sub-sections. I therefore recommend Paper 2 be read before, or in conjunction with, this thesis component. Paper 2 is presented in two distinct halves, and each half has a corresponding sub-section. The first sub-section discusses the first half of Paper 2; the implementation and evaluation of the top-down alien shooter. As hinted in Chapter 1, the top-down alien shooter ultimately failed a critical benchmark. In the first sub-section below, I extend upon the analysis conducted in Paper 2 and seek to explain the cause of the alien shooter failure. The second sub-section discusses the development of a second cognitive game, *the Buckets game.* This game is closely related to, but distinct from, the top-down alien shooter. This redesign allowed us to improve many

aspects of the paradigm. Given some of these aspects were not documented in Paper 2, I outline them and their importance below. The third sub-section is a *Summary and Transition* section, which as the name suggests, summarises the contributions of Paper 2 and sets the stage for Chapter 3.

The Top-Down Alien Shooter Exploit

In the first half of Paper 2 we uncovered a crucial flaw in the top-down alien shooter game. The crucial flaw was an *exploit* that was discovered by a large subset of players. These players learned to shoot quite accurately on very early passes, and due to this learning scored well beyond what should have been possible based on our risk-reward model. While we identify this exploit in the first half of Paper 2 – we do not explore the cause of this exploit. I do so here with additional analyses that were not integral to the goals of Paper 2.

To lay the groundwork for these analyses, it will help to recall that in an early iteration of the top-down shooter game outlined in Paper 1, we documented a similar flaw; players explored shooting on many passes for the first few trials, but then quickly settled for firing on a single pass. For example, one player may have settled on pass 4 and then shot only on pass 4, while others may have settled on pass 5 or 3. We discovered that when the player-shooter was fixed in the screen centre for all trials, players would learn to shoot accurately for one pass but not others. To clarify, consider that the initial speed and deceleration of the alien was identical for each trial. Therefore, an accurate shot on pass 4 could always be judged by the position of the alien at the time of firing. Once these timing cues were learned, for example on pass 4, this cue would provide a benefit for shots taken on pass 4, and it would provide misleading information for shots taken at pass 5 or 6. A player who tended to fire on

pass 4 would therefore find a shot at pass 5 or 6 more difficult, regardless of the slowing of the alien. Pass 4 would therefore become an exploit for that player. Randomising the position of the player-shooter (±100 pixels) was introduced to overcome this anomaly - by forcing participants to consider the spacecraft relative to the player-shooter.

To link this observation to the critical flaw described in Paper 2, it is helpful to consider the performance of early shooters and later shooters on the top-down shooter game. The observed difficulty-by-pass for all shots taken by early (panel A; median firing pass <=3) and late shooters (panel B; median firing pass >=7) is presented in Figure 1. Clearly, early shooters held a significant advantage for early passes, and found shooting on later passes more difficult. This suggests a similar flaw to what we observed in earlier pilot testing, but only for those players who invested in learning to shoot on early passes.

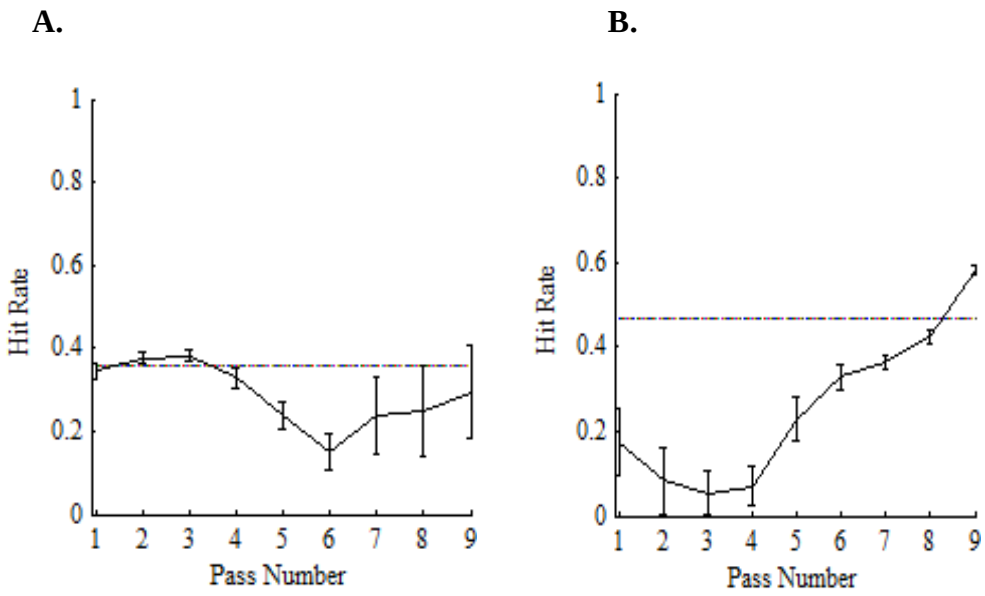


Figure 1. Hit rate by pass for all shots taken by early (panel A) and late (panel B) shooters. Early shooters are those participants with a median firing pass of 3 or below, while late shooters are those with a median firing pass of 7 or above. The dotted line indicates the overall probability of a hit. Error bars indicate the standard error of a proportion. Notably, early shooters found shooting at later passes more difficult. The rise in accuracy for late shooters on pass 1 is likely the result of chance success, which is approximately .1 for pass 1, and reduces as the alien slows.

To understand why this exploit only occurred for early shooters, it helps to consider that each bullet (22 pixels in height) took a constant 50ms to clear the alien (33 pixels in height) vertically. This 50ms was constant, so no matter whether it was pass 1 or pass 6, the bullet stayed within the vertical 'hit zone' of the alien for 50ms. Of course, the horizontal distance covered by the alien within this 50ms was much greater on early passes. This means it is possible that for early passes, the random placement of the player shooter may have had less impact on the exploit problem than it did for later passes. That is, if players invested in learning to shoot on very early passes, the randomisation of the player shooter may have had less impact, and players may have been able to shoot accurately by judging the position of the alien relative to the screen background, rather than its position relative to the shooter. These players may have been more accurate on early passes than later passes.

In combination, two figures presented below provide strong support for this explanation. Figure 2 presents the distance travelled by the Alien ship in 50ms for each pass, measured in horizontal pixels. Figure 3 presents the survivor function (one minus the cumulative distribution function) of the difference in position of the player shooter between one trial and the next. In Figure 3, the probability of seeing a difference smaller than the distance in pixels represented on the x-axis is represented by the area under the curve to the left of that distance. A comparison of Figures 2 and 3 highlights that for early passes, in the time it takes a bullet to clear the height of the alien (50ms), the speed of the alien means that the player does not have to consider the position of the shooter from one trial to the next on a large proportion of trials. Figure 2 shows on pass 1 for example, the distance covered by the alien while the bullet is in the vertical 'hit zone' is almost 70 pixels. Figure 3 highlights that this 70 pixels completely negated the movement of the shooter from one trial to the next on a

large proportion (>65%) of trials. This negation provides a significant benefit to players who persevere and learn to time their shots on early passes.

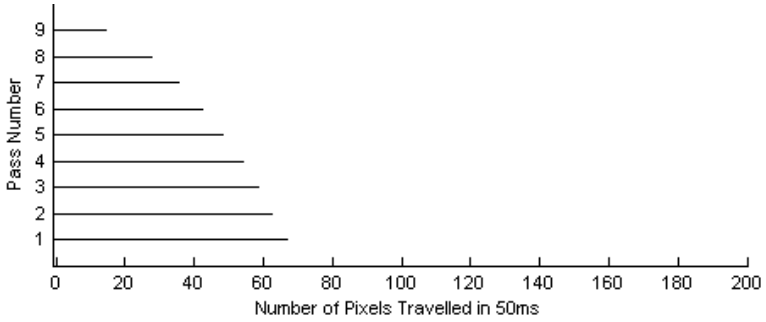


Figure 2. The distance travelled by the alien spacecraft in 50ms. This time represents the time from when a bullet tip reaches the height of alien spacecraft rectangles, until the tail of the bullet clears the top of alien spacecraft rectangles. On earlier passes, the spacecraft travels further due to its higher initial speed.

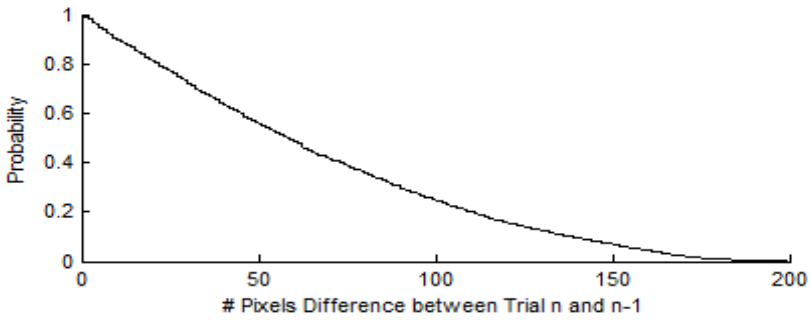


Figure 3. The survivor function (one minus the cumulative probability distribution) for the difference between player-shooter positions on each trial. The probability of seeing a difference less than a figure of interest is represented by the area under the curve to the left of that figure. Notably, differences close to 0 are the most common, while differences close to 200 are the least common.

The Buckets Game

The novel contribution of the Buckets game, and by extension Paper 2, was that it provided a platform with which we were able to accurately measure both performance outcomes (percentage of correct trials, or 'hits') as well as the difficulty

of each shot, to the tenth of a second. In the Buckets game, we moved toward a platform that allowed us a higher degree of experimental control relative to the top-down alien shooter game. We maintained some benchmarks from the shooter game, such as (a) a risk and reward profile in which the goal was to achieve a maximum number of hits in a fixed period of time, (b) no exploits and/or investment strategies, and (c) maintain an accuracy level of around 50% (this ruled out a two alternate choice design, which would have a 50% guess rate). By achieving these benchmarks, Paper 2 completed our piloting procedure and culminated in the delivery of a paradigm that was able to explore changes in performance and difficulty (risk taking) on successive trials.

Interesting though, while it was not a focus for Paper 2, we had added two new benchmarks for the Buckets game: (a) allow a fine grain measurement of the difficulty variable, and (b) allow the assessment post-error slowing. While Paper 2 does not discuss these goals and their relevance to the evolution of our research project directly - I briefly discuss these factors below.

The top-down alien shooter game had an ordinal measurement of risk; passes 1 through 7, with 1 being the hardest and 7 being the easiest. Thus, difficulty could only vary across 7 discrete levels. This ordinal, discrete scale of measurement decreased our statistical power relative to continuous variables such as response time, for which millisecond precision can often be attained. When thinking about the re-design of our cognitive game, we had noted this lack of granularity in the dependent variable and had determined to move toward a measurement of difficulty that was as close to continuous in nature as possible. It was this determination - in conjunction with the desire to rid the game of potential exploits - that led us to the Buckets game design. In the Buckets game, evidence toward the correct decision was delivered

every 100ms. We were therefore able to generate a game that had a risk and reward profile similar to the top-down alien shooter, overcame any potential exploit, and provided a fine grained measurement of the difficulty variable (tenth of a second). We also had very precise control of difficulty and could test many different rates of evidence introduction in piloting.

We were also acutely aware that by introducing evidence for the correct decision gradually and into a noisy perceptual environment, we had engineered a link to cognitive decision-making models. In sequential-sampling decision-making model, such as the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008), or the Diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998), evidence is accrued over time, in a noisy perceptual system, toward a decision threshold. A decision is made when the evidence for a particular choice reaches the threshold. These models have an extensive history of successfully describing the decision process in two-alternate decision-making paradigms, and more recently the LBA model had been applied to multiple-choice decisions (Eidels et al., 2010). This class of models had also been recently applied to better understand post-error slowing (Dutilh et al., 2012a; Dutilh et al., 2012b; Dutilh, Forstmann, Vandekerckhove &Wagenmakers, 2013). Thus, by moving toward a design that allowed comparison with perceptual decision-making, our Buckets game was uniquely positioned to investigate the hot hand and post-error slowing simultaneously. While Paper 2 does not mention post-error slowing directly, the evidence of this shift can be seen in the final paragraph of Paper 2. We noted that "this work extends beyond the understanding of how players make decision in games and sporting contests", and proceeded to highlight potential business applications, as well as users making decisions, risk taking, and strategy adoption. This shift is unpacked in more detail in the following transition section.

Summary and Transition

In sum, Paper 2 extended upon the iterative, data-driven game development cycle outlined in Paper 1. Unfortunately, a critical flaw in the top-down shooter game of Paper 1 invalidated it as a research platform suitable to investigate the hot hand effect. Paper 2 describes a fully coded and working game platform, the Buckets game, which provides a controlled testing environment to measure both performance outcomes (% correct) as well as the difficulty of each attempt. Paper 2 therefore completes the piloting stages of our paradigm development.

Importantly, by refining our measure of difficulty so that it was precise and near continuous, we had created a paradigm that brought the appraisal of the hot hand into line with other modern examinations of human performance. While we describe what might be termed a difficulty-accuracy trade-off, other well-known examinations of human performance describe the speed-accuracy trade-off (Fitts, 1954). This trade-off indicates that performing an action more quickly increases task difficulty, and leads to accuracy decreasing as a function of performance speed. As noted above, considerations of response times have led to well-developed models describing the speed/accuracy trade-off in decision-making (e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Brown & Heathcote, 2008). Considering both speed (i.e., difficulty) *and* accuracy provides a more complete account of performance than either measure alone (Brown & Heathcote, 2008).

Critically for the next stage of this thesis, bringing the appraisal of the hot hand into line with that of human decision-making allowed the parallels between the hot hand and post-error slowing to be illuminated. For example, the increased difficulty of basketball shots following success (Bocskocsky, Ezekowitz, & Stein, 2014; Rao, 2009) resembles performance in rapid-decision tasks commonly utilised in

cognitive psychology. In these tasks gradual speeding (analogous to more difficult shots) is observed over runs of correct responses that precede an error (Dudschig & Jentzsch, 2009; Laming, 1979a; see Luce, 1986, for a review), and a slower response time is typically found following an error (post-error slowing; Laming, 1968; Rabbitt, 1966). This empirical similarity is explored further in Chapter 3.

DOI 10.7603/s40601-013-0006-0

GSTF International Journal on Computing (JoC), Vol. 3 No. 2, July 2013

Evaluating Player Strategies in the Design of a Hot Hand Game

Paul Williams, Keith Nesbitt, Ami Eidels, Mark Washburn, David Cornforth

Abstract— **The user's strategy and their approach to decision-making are two important concerns when designing user-centric software. While decision-making and strategy are key factors in a wide range of business systems from stock market trading to medical diagnosis, in this paper we focus on the role these factors play in a serious computer game. Players may adopt individual strategies when playing a computer game. Furthermore, different approaches to playing the game may impact on the effectiveness of the core mechanics designed into the game play. In this paper we investigate player strategy in relation to two serious games designed for studying the 'hot hand'. The 'hot hand' is an interesting psychological phenomenon originally studied in sports such as basketball. The study of 'hot hand' promises to shed further light on cognitive decision-making tasks applicable to domains beyond sport. The 'hot hand' suggests that players sometimes display above average performance, get on a hot streak, or develop 'hot hands'. Although this is a widely held belief, analysis of data in a number of sports has produced mixed findings. While this lack of evidence may indicate belief in the hot hand is a cognitive fallacy, alternate views have suggested that the player's strategy, confidence, and risk-taking may account for the difficulty of measuring the hot hand. Unfortunately, it is difficult to objectively measure and quantify the amount of risk taking in a sporting contest. Therefore to investigate this phenomenon more closely we developed novel, tailor-made computer games that allow rigorous empirical study of 'hot hands'. The design of such games has some specific design requirements. The gameplay needs to allow players to perform a sequence of repeated challenges, where they either fail or succeed with about equal likelihood. Importantly the design also needs to allow players to choose a strategy entailing more or less risk in response to their current performance. In this paper we compare two hot hand game designs by collecting empirical data that captures player performance in terms of success and level of difficulty (as gauged by response time). We then use a variety of analytical and visualization techniques to study player strategies in these games. This allows us to detect a key design flaw the first game and validate the design of the second game for use in further studies of the hot hand phenomenon.**

Index Terms—**Evaluation, Games, Psychology, User-centered design.**

Manuscript received May 31, 2013.

P. Williams is with the University of Newcastle, Newcastle, 2299, Australia.

K. Nesbitt is with the University of Newcastle, Newcastle, 2299, Australia. (e-mail: keith.nesbitt@newcastle.edu.au)

A. Eidels is with the University of Newcastle, Newcastle, 2299, Australia.

M. Washburn is with the University of Newcastle, Newcastle, 2299, Australia.

D. Cornforth is with the University of Newcastle, Newcastle, 2299, Australia.

I. Introduction

DECISION-MAKING, risk-taking and strategy are important dimensions to many key business tasks, including trading shares, buying and selling real estate, project management and medical diagnosis. This paper examines a particular facet of decision making related to sports called the 'hot hand'. While the domain under study is sports-related the outcomes promise to be more generally applicable to software in more traditional business domains. This work also provides an interesting case study in the use of serious computer games to study decision-making. During the development of these games the interesting question of how unexpected user strategies might impact on outcomes is raised. Furthermore the outcomes highlight the importance of using empirical data to test user strategy when developing software.

Computer games often require players to exert significant perceptual and cognitive effort to be successful. This effort has been harnessed for tasks such as predicting the structure or proteins [1], labeling objects in images [2] and recognizing parts of images [3]. Computer games have also been widely spoken of as new multimedia platforms for general learning [4] and communicating about science [5].

There is also a significant potential for using computer games to assist with psychological research. Indeed a number of studies have used existing games such as Tetris and Madden to explore aspects of cognition [6, 7, 8]. Game engines have also been used to support studies in spatial cognition and social behavior [9, 10, 11, 12, 13].

In this paper we describe the development of two serious games to assist in the study of the psychological phenomenon known as the 'hot hand' [14]. To be useful in such a study these games need to meet particular design criteria in terms of player performance. In interface terms this performance is related to the efficiency and effectiveness of the user. As in typical usability studies we gathered empirical data under experimental conditions to test that our games meet our design criteria. Using this approach we found that the first game had an unintentional design flaw. This flaw made it less suitable for studying the hot-hand phenomenon. Therefore we developed a second game to address this problem. After following a similar empirical testing procedure the second game was found to meet our hot-hand requirements.

In the next section we discuss the hot-hand phenomenon and the particular design requirements for a game that allows the study of the hot hand. In the subsequent sections we

describe our first game design, called 'Aliens', the methods we used to test it, and the results of our usability analysis. We then describe a second game design, called 'Buckets', and provide an analysis of results from this study in a similar manner. In the final section of the paper we compare and contrast the results from the two game designs and discuss directions for future work.

A. Hot Hand

The term 'hot hand' describes the belief that the probability of a hit (success) following a hit should be greater than the probability of a hit following a miss (failure). In seminal research, it was found that 91% of basketball fans believed professional players had a better chance of making a shot after having hit their previous two or three shots than after having missed their previous two or three shots [15]. Professional basketball players also endorsed the belief; with each interviewed agreeing "it was important to pass the ball to a player who had made several shots in a row" (p. 302).

While intuitively these beliefs and predictions seem reasonable, no evidence for the hot hand was found in the field-goal shooting data of the 1980-81 Philadelphia 76ers. Likewise, further analysis of data from professional basketball [16], baseball [17] and golf [18, 19, 20] all failed to support the intuitive belief in the hot hand. This lack of empirical evidence led some theorists to suggest that the belief in the hot hand is a psychological fallacy [15, 21, 22]. That is, hot and cold streaks in performance are a myth that players and spectators endorse.

The most common explanation for the disparity between the popular belief that hot hand exists and actual data that shows no support for hot hand is that humans tend to misinterpret patterns in small runs of numbers [15]. That is, we tend to form patterns based on a cluster of a few events, such as a player scoring three shots in a row. We then use these patterns to help predict the outcome of the next event, even though there is insufficient information to make this prediction [23]. This is somewhat akin to the 'gamblers fallacy' that also arises from a belief in the law of small numbers [24], although for reasons we shall not discuss here the latter actually makes opposite predictions (people expect gamblers to fail after successful streaks).

However, the somewhat elusive hot- hand effect has been reported in the literature. Players have been reported to get on hot streaks in tasks such as horseshoe pitching [25], billiards [26] and ten-pin bowling [27]. Most recently, [28] found strong evidence for hot hand performance in volleyball. Although early hot hand findings (i.e., the lack of hot hand) seemed at odds with intuitive predictions, there now seems to be more to the hot hand picture than can simply be explained by a cognitive fallacy.

Under close examination, empirical studies of the hot hand seem to follow a qualitative pattern. On the one hand (no pun intended), in tasks where the difficulty of each shot is largely 'fixed' the hot hand seems common. This is true in tasks like horseshoe pitching and ten-pin bowling. Even in games like volleyball the defensive side must remain on the opposite side of the net and cannot influence the striker greatly. On the other hand, in sports where the difficulty of each shot attempt is 'variable' there is no evidence in the data for hot or cold streaks. This is true in sports like basketball where the defense can interfere.

Hot hand may be a myth resulting from a cognitive fallacy, however, the pattern highlighted by grouping 'fixed' and 'variable' studies seems to support alternative interpretations. One such explanation was provided by Smith [25] who suggested shooters might systematically take more difficult shots in response to a run of hits. Under this scenario, a player does show an increase in performance during a hot streak - as they are performing a more difficult task at the same level of accuracy. This increase in performance would not be detected by traditional accuracy measures, but may be detected by teammates and spectators.

While this hypothetical *difficulty-account* receives tentative support by drawing a distinction between fixed and variable difficulty tasks (as the hot hand is more likely to appear in fixed-difficulty tasks, where players cannot engage in a more difficult shot), further support must be provided for two underlying assumptions. These assumptions are that (1) when task difficulty is considered, players' performance can be different to what is predicted or expected, and (2) that people sometimes take on more difficult tasks in response to success and easier tasks in response to failure.

The first assumption under investigation can be framed in terms of a difficulty-accuracy trade-off. To explain, consider that as a task becomes more difficult, people tend to perform the task with less accuracy. Is it possible however that people performance might differ from this intuitive difficulty-accuracy trade-off? More specifically even, can people maintain performance levels as a task becomes more difficult?

Psychological research suggests this is possible. In fact two prominent groups of cognitive theories account for such findings. Energetical theories [29, 30] suggest increases in task difficulty lead to an increase in arousal, which in turn increases the maximum level of mental effort available to a task. On the other hand, perceptual load theory [31] suggests high perceptual load (i.e., higher difficulty) leads to a decrease in distraction from other information, thus allowing greater focus on more difficult tasks. A large body of evidence supports this account in perception [32].

Importantly these findings are not restricted to the laboratory. For instance in a famous study on Munich taxi-cabs, half of a fleet of otherwise identical taxis were fitted with an anti-lock braking system (ABS). ABS brakes improve driver control under braking [33], and as a result make driving easier and safer. However, over a 12 month period in which distance travelled and driver ability were controlled, no difference was found in the number or severity of accidents for taxis with and without ABS brakes. Wilde [34] suggested this and other similar findings demonstrate that people are willing to accept a consistent level of risk. They will maintain this fixed risk level even when conditions vary; in the taxi-cabs study, for example, the level of driving errors (risk) had

remained constant despite safer driving conditions. Likewise, Wilde [34] argued that if tasks become more difficult people might become more careful. By managing risk in this way, it is plausible that people can maintain consistent accuracy across different levels of difficulty.

The second assumption that requires support involves people's reactions to success and failure. Is essence - is there evidence to suggest that people may attempt more difficult tasks after successes, and less difficult tasks after failures?

Experimental support provides some evidence for this claim. Wilde, Gerszke, and Paulozza [35] asked participants to tap a series of red squares after their appearance on a computer screen. Responses closest to 1500ms were rewarded the highest points, however responses faster than 1500ms were penalized. In response to a run of point scoring trials, participants adopted successively more risk by making faster taps, however, following a penalty, subsequent taps were significantly slower and less risky. These results suggest people may take riskier options after successes, and less risky options after failures. It follows that performers may systematically adjust task difficulty in response to success and failure.

Attempts have also been made to assess this assumption outside of the laboratory. Rao [36] analyzed 60 LA Lakers basketball games in the 2007-08 season, and reported that while the majority of players attempted more difficult shots following a successful run, no tendency was found for players to attempt less difficult shots following an unsuccessful run. While Rao's results are of interest, the complexities of sports analysis must be considered. It is debatable whether any coding system can accurately assess the variety of contexts in which basketball shots are taken; particularly given individual players differ in shooting strengths and weaknesses.

We are thus faced with a dilemma. It seems we can support our two key assumptions, however more data needs to be gathered to investigate potential explanations of the hot hand. Unfortunately trying to gather more data from sporting games and contests is fraught with problems of subjectivity. How can one objectively assess the difficulty of a given shot over another in basketball? How can one accurately tell if a player is adopting an approach with more risk?

Our proposed solution to this problem is to design computer challenges of matched 'variable' and 'fixed' difficulty tasks that can be employed to test various hypotheses surrounding the hot hand. This presents challenges in designing tasks or game challenges that have particular usability characteristics. This paper focuses on the characteristics required in variable-difficulty hot-hand games.

A variable difficulty hot-hand game requires some careful design and testing if it is to be used to gain insight into how players respond to a run of success or failure. Namely, a hot hand game must provide a challenge with binary outcomes, that is, a challenge in which a player either succeeds or fails. The player must also be given clear feedback on each outcome, the same way a basketball player knows for sure whether he had hit or missed.

We intend to use the game to study a precise psychological phenomenon related to hot and cold streaks in performance. Therefore, a further requirement for a hot hand game is that it allows measurement of players' strategy after runs of both successes and failures. If people fail most of the time, we won't record enough runs of success. If people succeed most of the time, we won't observe enough runs of failure. Thus, the core challenge needs to provide a probability of success, on average, somewhere in the range of 40-60%.

However the most significant requirement for a hot hand game is that it requires a finely tuned risk and reward structure [37]. The game must allow players to take risks and to be adequately rewarded for the risk. If, for example, one risk level provides substantially more reward than any other, players will learn this reward structure over time, and be unlikely to change strategy throughout play. We would thus like each risk level to be, for the average player, equally rewarding. In other words, regardless of the level of risk adopted, the player should have about the same chance of obtaining the best score. In the games described below this is managed by balancing speed in the task with accuracy. The faster a player responds the less likely they are to succeed. This is balanced by allowing more opportunities for the player to attempt the task when they respond faster. So even though at faster speeds players may make more errors they will receive more chances to succeed.

In this paper we outline the development and analysis of two 'variable' difficulty tasks. One for a game called Aliens, and the other for a game called Buckets. The tasks in these games are designed so that changes in player strategies can be accurately recorded as the game progresses. We compare the two game designs by collecting empirical data that captures player performance in terms of success and shot difficulty (response time). In terms of usability these measures equate to effectiveness and efficiency.

Having collected the data we then used a variety of analytical and visualization techniques to study player strategies in these games. This allowed us to detect a key design flaw in the Aliens game, which made the game less suitable for hot-hand investigation, leading to the design of the Buckets game. Testing of this game showed that we had successfully removed the flaw and the resultant game was suitable for further study of the hot hand.

II. EXPERIMENT 1: 'ALIENS'

A. The Aliens Game

The Aliens game is a simple first person shooter game developed in Flash and Actionscript (see Fig. 1). The players' goal is to shoot down as many alien spacecraft as possible within the overall time allowed. The game consists of a repeated challenge where a single alien spacecraft appears and moves across the game screen and the player's spacecraft is allowed a single shot to hit the alien. Entry and exit by each new alien (a trial) can therefore result in a hit or miss.

All trials are separated by a brief period where no alien is on the screen. New trials always begin unless time has run out. In the case where a trial is underway as time runs out, the trial

continues until completion but no result is recorded. For each new trial the player's spacecraft is fixed in a random position within an area ±100 pixels from the screen center (see Fig. 2).

Fig. 1. Screenshot of the Aliens game in operation.

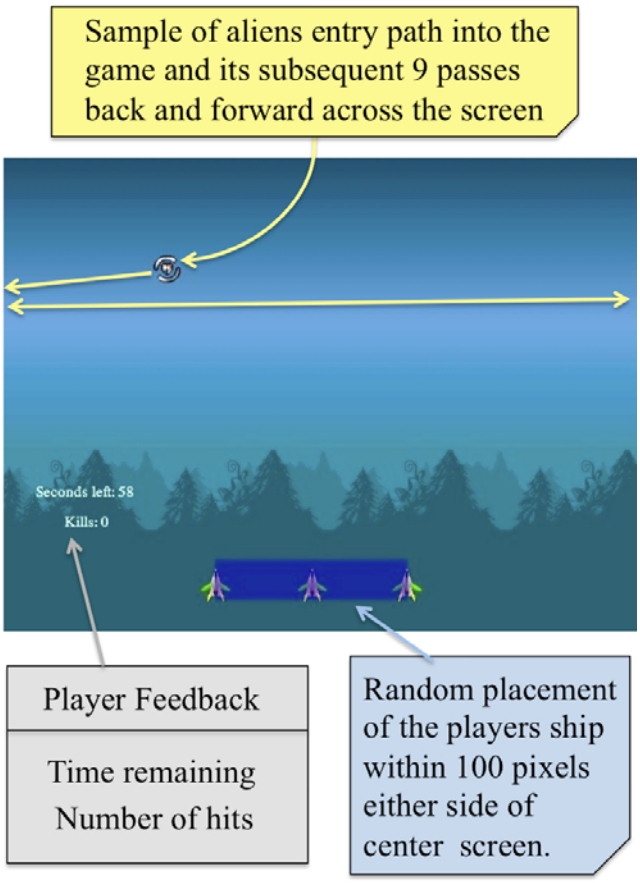


Fig. 2. The main mechanics of the Aliens game.

On each trial, an alien spacecraft (hereupon alien) enters the top of the game screen and moves either left or right in a downward arc until reaching a set height from the top of the screen (see Fig 2). The alien then travels from side to side at this height passing over the player-shooter a maximum of nine times. The aim for a player is to time their shot so that a bullet from the player's spacecraft intercepts the alien on one of these nine passes. A shot is declared a miss once the bullet clears the maximum height of the alien without making contact. A shot is declared a hit if the bullet intercepts the alien (pixel contact). If a player fails to take a shot during the nine passes of the alien then it is considered a non-attempt.

The player is allowed only one shot per alien. If an alien is hit, it explodes onscreen. Each shot and hit is accompanied by appropriate auditory effect. The trial completes immediately if the player is successful with their shot. If a shot is missed, a penalty period ensues while the alien completes the nine passes and exits the screen.

Importantly, in each successive pass the alien spacecraft decelerates. An assumption is that the slower the alien moves the easier it becomes to target. Therefore, the longer a player waits to take her shot, the easier it becomes to hit the alien. Since a player is allowed only one shot per alien, the game incorporates an element of strategy - shooting on earlier passes allows more time for additional attempts at shooting aliens. However, earlier passes present more difficult shots, increasing the players' risk of failure.

For all complete trials the initial direction of the alien (left/right), the position of the player-shooter (±100 pixels from the center), the difficulty of the attempted shot (pass number 1-9, where 1 indicates the highest shot difficulty), and the outcome of the shot attempt (h = 1; m = -1; non-attempt = 0) are recorded. The block and trial number are also recorded for each shot.

B. Methods

The experiment was run in a dimly lit room on IBM compatible computers using Windows XP and standard keyboards. Seventeen-inch CRT monitors were used for the experiment with screen resolution set to 1024 by 768 pixels. The experiment was coded in Actionscript 3.0 and run in Mozilla Firefox version 3.1 browsers for Windows. Participants wore Sennheiser headphones.

Twenty-nine participants from the University of Newcastle, Australia volunteered in response to recruitment posters. All participants had normal hearing and normal or corrected-to-normal vision. Each of the participants in the study was reimbursed $AUS10 for taking part in the experiment.

Participants first played two three minute time periods (blocks) for training purposes. After these training blocks they played a further three blocks of trials each lasting 12 minutes. To progress between each block, participants had to press the spacebar. The spacebar was also used in the game to fire each bullet. Participants were asked to maintain a comfortable, self-selected distance from the screen throughout.

Both verbal and onscreen instructions outlined the goal and rules of the game for participants. Players were advised that the first two blocks should be treated as "practice", and that shooting down as many aliens as possible would require "both speed and accuracy".

A simple visual interface provided feedback on the player's

current performance and game status by registering the number of kills (hits) and the time remaining during each block (game) (see Fig. 2). At the completion of a block, participants also received summary feedback on the number of kills made for that block, and their grand total number of kills during the experiment. Participants were encouraged to use this feedback to monitor performance and set future goals.

C. *Results and Discussion*

A summary of results for the participants is shown in TABLE I. On average each participant in the study completed 407 trials. The average number of hits was 151 and the average number of misses was 256. However, there were large variations in player performance. For example, in terms of the number of trials completed there was a standard deviation of approximately 146. Indeed the maximum number of completed trials was 810 and minimum was 255.

To look for clusters of players who performed at different levels of expertise, or who used different strategies, we calculated each player's percentage success rate, their average response time (as a gauge for shot difficulty), and their total number of hits. We then used this data in a multi-dimensional scaling routine based on a Sammon projection [38]. The results of our Sammon mapping are shown in Figure 3. As can be seen in this figure, players 3, 6, 8, 9 and 21 appear as a unique cluster in terms of their performance.

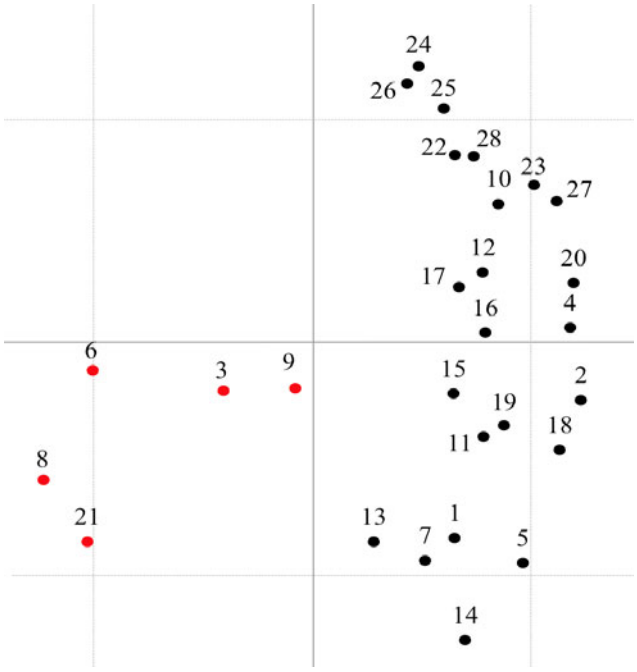


Fig. 3. Example Multi-dimensional scaling from the Sammon mapping indicating a distinct cluster of players (3,6,8,9,21)

While the non-linear projection associated with the Sammon mapping is difficult to correlate with the original variables it is extremely useful for the type of exploratory analysis we wanted to perform. Once we identified two possible player clusters we then used further interactive visualization software to analyze the players in terms of response times, success rates and the total number of hits (see Fig. 4).

TABLE I. NUMBER TRIALS PER PLAYER, AVERAGE PLAYER HIT RATES AND RESPONSE TIMES IN THE ALIENS GAME. FIVE PLAYERS (3,6,8,9,21) COMPLETED WELL ABOVE THE AVERAGE NUMBER OF HITS (151).

Player	Number Trials	% Hits	Average Response Time (pass no)	Total Hits
1	415	29	4.7	121
2	274	31	7.4	86
3	457	45	4.3	204
4	255	36	8.1	93
5	369	25	5.1	93
6	532	49	3.5	261
7	564	25	4.8	141
8	720	42	2.8	300
9	510	40	5.5	203
10	299	44	9.1	133
11	434	30	6.6	131
12	353	40	8.4	142
13	582	28	4.6	162
14	617	19	4.2	120
15	421	34	6.8	143
16	368	37	7.7	135
17	361	40	8.0	146
18	395	27	7.2	105
19	416	30	6.9	126
20	303	36	9.0	109
21	802	36	2.7	286
22	276	50	9.1	138
23	255	45	9.4	115
24	279	58	9.8	161
25	283	53	9.4	150
26	291	56	9.4	163
27	258	43	9.6	110
28	298	49	9.2	145

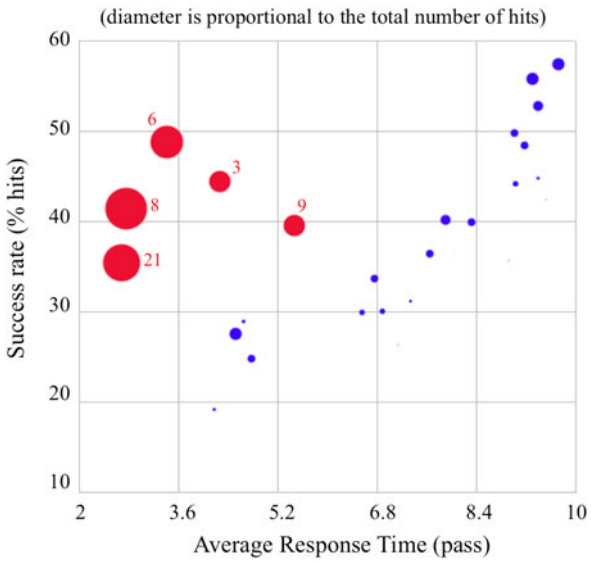


Fig. 4. Average response time versus the success rate for players in the Alien game. The diameter of points on the plot shows the relative number of hits during the game. The 5 players indicated are characterised by a high number of hits, low response time and unexpectedly high success rates.

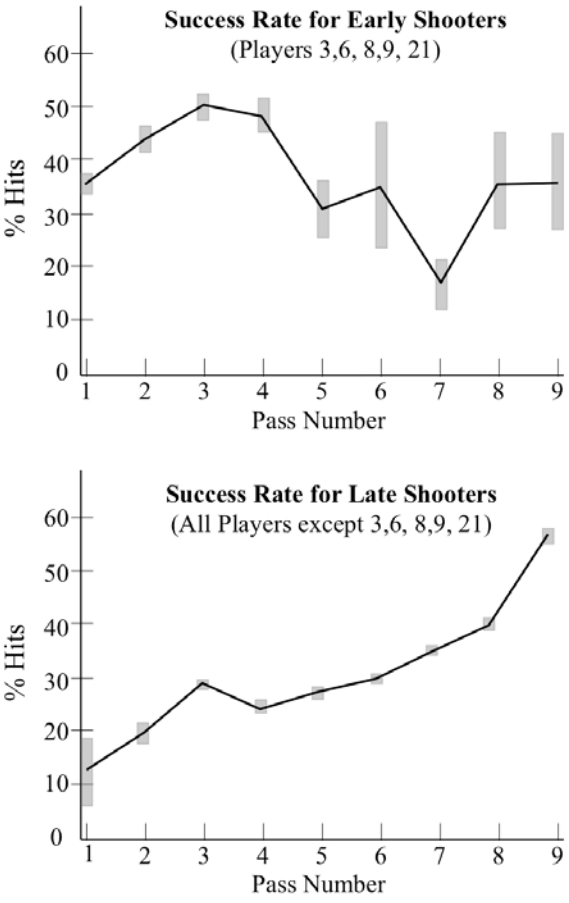


Fig. 5. Early versus late shooters in the Aliens game showing hit rate by pass for all shots taken by players 3,6,8,9,21 (top) with all other players (below). Notably, early shooters found shooting at early passes easier than late passes. Late shooters reflect expected difficulty, being successful at later passes.

To try and understand how more difficult, early shots could result in a higher probability of hits for some players we interviewed some of the identified group. They had discovered that on early passes, they could accurately time their shot by watching the approach of the alien to the edge of the screen. This provided in effect a low risk, high reward way to shoot early in the trial. The edge of the screen acted as a kind of 'gun sight'. It seems that four other players also identified the same strategy. The effect of this strategy is shown in Fig. 5.

Unfortunately this unintentional flaw in our Alien game design made it unsuitable for testing the hot hand phenomenon. The risk and reward for players of a hot hand game need to be balanced so that higher risk behavior from the player incurs lower levels of reward. As a result of this problem we designed an alternative game based on a simple perceptual challenge. This second game was called Buckets and is described in the next section.

III. EXPERIMENT 2: 'BUCKETS'

A. The Buckets Game

The Buckets game is based on a repeated perceptual challenge that requires players to decide which of four buckets is becoming darker (see Fig 6.). The goal of the game is to identify as many target buckets as possible in a fixed time period.

At the beginning of each trial, players view four buckets (rectangles). Each bucket is half filled with blue pixels (drops) that have been randomly positioned. This display is shown in Fig. 6. During a trial, the blue pixels are randomly re-positioned 10 times per second, creating a dynamic effect within every bucket similar to visual static. Over the course of a trial, one bucket (the target) accumulates additional blue pixels at a constant rate. Players can attempt to select the target at any time. A correct target selection is declared a hit, while an incorrect detection is declared a miss. Players are provided with clear visual and auditory feedback signaling the outcome of each trial.

The response time of the player in the Buckets game is equivalent to the pass number measured in the Aliens game. , The faster a player responds, the more difficult the task should become. Hence faster decisions allow more time for additional trials, however faster decisions are more risky and may be more likely to result in failure. This has been achieved by allowing more dark pixels (drops) to accumulate in the target bucket as the trial progresses. Drops accumulate at a constant rate in the target bucket, so as time progresses it becomes easier to distinguish from the three distracter buckets that do not accumulate more drops over time. In this way the Buckets Aliens games are analogous – a risk/reward strategy must be adopted in both games with the aim of finishing as many correct trials as possible within a fixed time period.

Despite these similarities, the games have an important difference. In the Buckets game a player has a 25% chance of simply guessing and still identifying the target correctly. Therefore, a player could attempt many trials and make many successes by simply guessing at the earliest possible time on

every trial. To counteract this strategy, incorrect decisions were followed by a brief penalty time period before the next trial began.

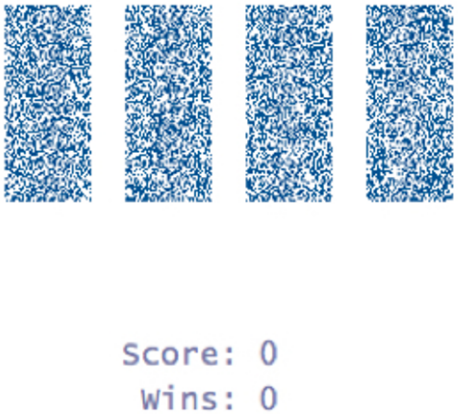


Fig. 6. The Buckets Screen showing the four buckets partially filled.

Once again the design of the game play emphasizes the need for both accuracy and speed in the players' responses. Waiting for the trial to be easy incurs a time cost, reducing the overall time remaining for subsequent attempts. Importantly, the game mechanics (i.e., rate of introduction of pixels, penalties, etc) were extensively piloted so that early attempts would provide roughly the same amount of correct decisions as later attempts over a long time period. A simple scoring mechanism keeps count of the number of wins or hits and provides feedback to the player (see Fig 6.).

Trials are separated by a brief period where no buckets are on the screen. New trials always begin unless time has run out. In the case where a trial is underway as time runs out, the trial continues until completion but no result is recorded. For all other trials the difficulty of the attempted shot (response time, where closer to 0 indicates the highest shot difficulty), and the outcome were recorded.

B. Methods

Twenty-four participants with normal or corrected to normal vision were recruited via posters placed at the University of Newcastle. In this game, each player was paid a set amount per correct response to help motivate them to make as many correct target selections as possible.

Before play, participants were shown two complete trials that did not require any response. This allowed them to view the total amount of change in the target over the course of each trial. All participants then played a 5-minutes long practice block, followed by four experimental blocks of 10 minutes each. Participants were encouraged to explore differing strategies during practice, and were only paid per correct response during the experimental blocks. Again the game goals were explained verbally and onscreen.

A complete trial uninterrupted by a player's response lasted 8000 ms (80 updates). Additional blue pixels were introduced at 1.875 pixels per update. Other game variables included a 300ms central fixation cross before each trial, 250ms pre- and post-fixation blank screens, and feedback after attempts (500ms for correct and 2150ms for incorrect attempts; the difference of 1650ms being the penalty for incorrect decisions). Participants indicated which rectangle they believed was the target by pressing one of four spatially-corresponding keys ('a', 's', ';', or '''), with each success being worth 1 point. At the end of each block, participants were given feedback on the number of correct decisions made for that block and their grand total. They were encouraged to use this feedback to monitor their performance.

The experiment was again run on IBM compatible computers using Windows XP and standard keyboards. Screen resolution was set to 1024 by 768 pixels on 17" CTR monitors. The experiment was coded in Actionscript 3.0 and run in Mozilla Firefox version 3.1 browsers for Windows. Participants wore Sennheiser headphones.

C. Results and Discussion

On average each participant completed approximately 370 trials with an average of 210 hits and 160 misses. There was a standard deviation in the number of trials of approximately 30. We note that the variation between players in terms of completed trials was much lower than the variable performance seen in the Aliens experiment. The key results for each player in the Buckets experiment are shown in TABLE II.

TABLE II.

AVERAGE PLAYER HIT RATES AND RESPONSE TIMES IN THE BUCKETS GAME.

Player	% Hits	Average Response Time (ms)	Total Hits
1	64	5.3	222
2	54	3.9	226
3	73	5.4	255
4	63	4.9	234
5	32	2.3	176
6	54	5.2	189
7	68	4.8	256
8	51	5.5	169
9	55	4.4	215
10	52	4.0	213
11	60	4.9	220
12	37	3.9	150
13	66	5.2	234
14	64	5.8	211
15	62	5.5	213
16	60	5.5	203
17	71	5.6	241
18	59	4.9	215
19	51	6.5	151
20	53	4.7	196
21	72	5.9	238
22	69	5.1	251
23	49	4.6	183
24	46	4.1	180

Once again we looked for clusters of players using the normalized results for each player's percentage success rate, their average response time and their total number of hits. We followed the same procedure as in the Aliens game and used these data in a multi-dimensional scaling routine based on a Sammon projection [38]. The results of our Sammon mapping for Experiment 2 are shown in Fig. 7. As can be seen in this figure, there appeared to be only one main cluster although players 5 and 19 appeared to be outliers. Table 2 highlights the results from these same two players.

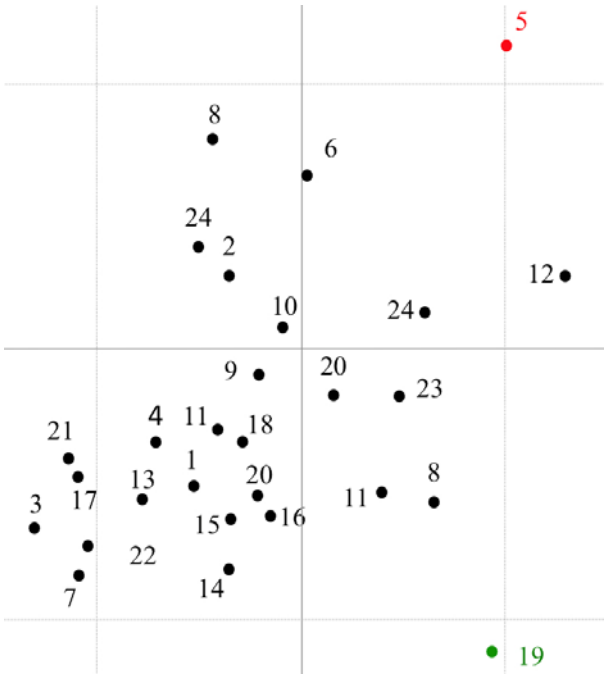


Fig. 7. Multi-dimensional scaling results from the Buckets game. Identifying two outliers (palyers 5, 19).

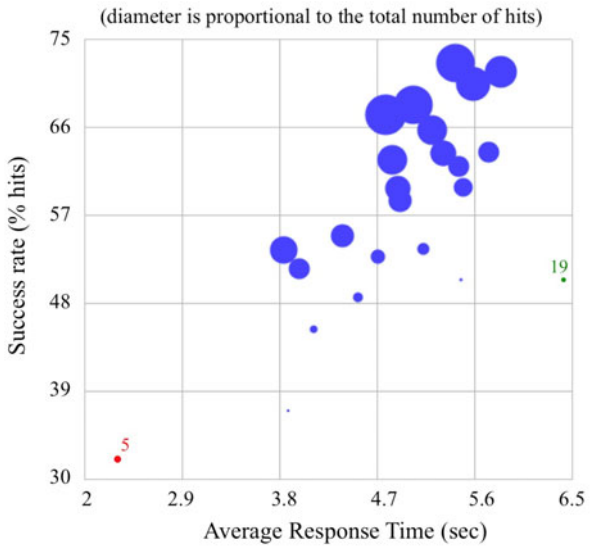


Fig. 8. Visualising player strategy in buckets game. Note that in comparison to Fig. 4 higher success rates are associated with slower response times.

Once more we followed the same procedure used with the Aliens game and employed interactive software to visualize the players in terms of their response times, success rates and total number of hits. Fig. 8 helps to highlight the two identified outliers. Player 5 seems to have shot very early, however as expected, she or he also had a low hit rate. That is this player took high risk but in doing so registered a low number of hits. Player 19 shot late but had a relatively low success rate. This may indicate poor aptitude to the task. The Pearson coefficient of correlation for average response time and percentage of hits using all players was 0.7. This satisfactory relationship between response time and success rates supports the use of the Buckets game in our hot hand study.

IV. GENERAL DISCUSSION

In this paper we have described the development and testing of two games, the 'Aliens' and 'Buckets' games. These games were specifically designed to study the Hot Hand phenomenon, which has been extensively studied in psychological research. These games offer for the first time a well-controlled testing environment for a phenomenon that was measured, up-till-now, off the laboratory and was therefore sensitive to a number of contextual variables (but see [39] for preliminary investigation in that direction). Analysis of the first game revealed certain biases in players' strategies that deemed it less appropriate for testing the Hot Hand. Therefore, a second experiment was developed with an eye on these biases. Indeed similar analysis on the results of the second game revealed it is robust to changes in players' strategies, and can therefore be used in the psychological arena to test the mechanisms that underlie the belief (and potential existence) of the Hot Hand.

The term Hot Hand marks the common belief, in basketball and other sports, that the probability of making a shot given that the player had just made the previous shot (i.e., the probability of a hit given a hit) is greater that the probability of making a shot given a miss on the previous shot-attempt. While strong belief in the hot hand is well documented, empirical evidence for hot hand is rather sparse. In their seminal study, Gillovich et al. [15] showed that even though both spectators and players strongly believed in the hot hand, professional basketball players were not more likely to make a shot if it was preceded by a successful attempt. Similarly, studies in other sports [e.g., 16, 17, 18, 19, 20] all failed to provide empirical support for the hot hand. However, we recount that many of these studies focused on field sports, where experimental control is minimal if not impossible. Studying the hot hand with specialized computer games, as we did here, allowed much better control of critical experimental factors. It was this intersection of performance in a task, and the difficulty of the task at hand that formed our departure point for the current study.

We developed two computer games that allow measuring both the performance of the players, and the difficulty level of each and every shot attempt. Both games featured challenges with binary outcomes, where players could either succeed or

fail on each trial. This type of binary challenge is essential for testing the hot hand.

Another important design feature for a hot hand game is that players have an average success rate of between 40-60% for the challenges. This should allow for both hot and cold streaks to be distinguished in the data. The Aliens game had an average success rate of 39% (std dev 9.9). The Buckets game had an average success rate within this range, of 57.5% (std dev 10.4). While these success rates are at the limit of what we would like they are both considered acceptable.

The most important criterion for our hot hand games is that players are rewarded appropriately for both efficiency and effectiveness in the repeated tasks they undertake. We tried to design both games so that there was a balance between risk and reward. We encouraged players to take risks (respond early) by rewarding them with more shot attempts. In setting up this reward structure we intended that higher risk would equate to lower success rates in the task. However, after collecting empirical data for the first, Aliens game we uncovered a serious design flaw. Some players had uncovered a 'cheat' in the game and were able to achieve high success rates when responding early in the game. This effectively made the game unsuitable for studying the hot hand. The alternative game called Buckets was developed and tested in the same manner. It was found to meet the requirement that fast response times relate to low success rates, thus making it acceptable for further study of the hot hand.

We plotted hit-rate data from both the Aliens and Buckets games as a function of difficulty (Figures 4 and 8, respectively) and furthermore visualized these data using Sammon projection (Figures 3 and 7) to identify clusters of players with similar and dissimilar strategies. The qualitative patterns in Figure 4 and 8 differ in a meaningful way; players in Figure 8 (Buckets) are roughly aligned along the main diagonal, suggesting that hit rate increased for players that were willing to wait longer, on average, before making a decision. Figure 4 (Aliens), in contrast, reveals some players that have responded very quickly yet were able to maintain a high level of performance. We referred to this sub-group of players earlier and suggested they have identified a 'cheat' in the game. We concluded that players in the first game may be divided to two groups based on their response strategies, whereas such division is unlikely to have happened in the second game.

Yet, the fact the players presumably used a single response strategy in one domain does not imply they may not differ in other aspects. In the remaining of the discussion we highlight interesting differences in players' performance and strategy that had been revealed by our analyses.

First, players differed in their competence level on both games. Figure 8 shows performance in the Buckets game, measured by % hit, as a function of difficulty (gauged by average response time). For a given level of task-difficulty, such as responses that were executed at around 4.7 seconds, on average, one player had a success rate of 53% while another had a success rate of 68%, with yet other players within this range. Clearly, for the same level of task difficulty different players could perform rather differently (by as much as [68-53] / 53 = 28%, in this example). Differences in player performance are not unexpected in games. Even as early as 1979 Atari recognized this and designed games such as Adventure [40] for the Atari 2600 to provide different difficulty levels.. More recently a number of games such as Max Payne [41] and Left 4 Dead [42] have incorporated techniques known as "challenge functions" [43] to dynamically adapt the difficulty of game play based on the current player performance.

Players also differed in the risk they were willing to take. Some players were willing to commit to a decision relatively quickly, responding by as early as 3.8sec, on average, while others had waited longer, some of them as long as 5.8sec (see Fig. 8 again). While fast responses clearly impacted performance by way of pushing hit rate down, these 'fast-to-respond' players seemed to have been willing to accept the risk associated with fast responses. The finding that the overall level of risk accepted by players showed large individual differences is commensurate with psychological research surrounding impulsivity and risk-taking [44, 45]. Indeed, a future avenue for research will be to critically assess the relationship between these psychological constructs and players' behaviour in our hot hand games. Of course the combination of risk-taking and difficulty is also an important consideration in the design of games. Indeed some attempts have already been made to dynamically adapt the game play difficulty by accounting for both player performance and their risk profile [46].

Finally, players may differ in the way they explore the game's environment. Some players may explore pay-offs across a range of difficulty levels, to test how to maximize gains, while others may settle on a given level of risk in an attempt to exploit known rewards. Hills, Todd, and Goldstone [47], as well as others, studied the trade-off between exploitation and exploration in mental strategies. We have addressed this issue elsewhere, in the more specific context of hot hand games [39]. In the current games, an exploration strategy may have allowed some players in the Aliens game to identify a 'cheat'. These players may have tried to respond across a range of latencies and discovered, either by chance or via systematic exploration, that early shots reward them with a high hit-rate, while also conserving time for additional shots. Critically, at least from the perspective of the hot-hand research, no such behavior was similarly rewarded in the latter, Buckets game.

V. CONCLUSION

To conclude, we developed and tested two games that allow assessing both performance and shot-difficulty in a hot hand challenge. If we assume there is variable difficulty in some sporting tasks them measures of sport performance, such as basketball shooting percentages, can sometimes be misleading. The novel contribution of the proposed games in that they provide a controlled testing environment, one that allows to accurately measure both performance outcomes (shooting percentage) as well as the difficulty of each shot.

Thus, we expect it to become a useful tool in the systematic exploration of the hot hand phenomenon. In this paper we focused not only on the evaluation of players' performance level, but also the evaluation of players' strategies, particularly in terms of risk taking. Players could have saved time by taking early shots with higher difficulty, or obtain higher accuracy rate on the expense of time if they were to wait until the trial became easier. Our analyses revealed individual differences across players in game-competence, risk taking, and possibly exploration-exploitation strategies. However, based on cluster analysis, the structure of the Buckets game makes it robust to these differences and therefore adequate as a platform for studying the elusive hot hand phenomenon.

The value of this work extends beyond the understanding of how players make decision in games and sporting contests. Many traditional business applications also rely on users making decisions, taking risks and adopting strategies. Consider applications of intra-day trading where market traders make many rapid decisions about how to trade stocks. For example, do stock traders develop 'hot hands', perhaps taking greater risks after a successful string of trades? More generally, what role does user strategy play on the efficiency or effectiveness of software designed to support business tasks? Is there an opportunity to improve the design of business software by gathering more empirical data and looking at user patterns? These are a few examples of open questions that form part of our larger study beyond sporting contests and computer games.

ACKNOWLEDGMENT

We would like to thank David Elliot from the Newcastle Cognition Lab for his invaluable work in coding the various versions of both the Aliens and Buckets games.

REFERENCES

[1] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and Foldit players, "Predicting protein structures with a multiplayer online game", Nature. Vol. 466 Issue 7307, p 756–760, 2010.

[2] L. von Ahn, and L. Dabbish, " Labeling images with a computer game", in CHI '04: Proc. 2004 Conf. Human Factors Comput. Syst., pp. 319–326, 2004.

[3] L. von Ahn, R. Liu, and M. Blum, "Peekaboom: a game for locating objects in images", in CHI '06: Proc. SIGCHI Conf. Human Factors Comp. Syst., pp. 55–64 , 2006.

[4] A. Derryberry, "Serious games: Online games for learning". Retrieved 10 August 2012 from Adobe Resources website: http://www.adobe.com/resources/elearning/pdfs/serious_games_wp.pdf

[5] A. Krotoski, "Serious fun with computer games", Nature. vol. 466 Issue 7307, p. 695, 2010.

[6] J. T. Hansberger, C. D. Schunn, and R. W. Holt, " Strategy variability: How too much of a good thing can hurt performance",. Memory & Cognition, vol. 34, pp. 1652–1666, 2006.

[7] D. Kirsh, and P. Maglio, "On distinguishing epistemic from pragmatic action", Cognitive Science, vol. 18, pp. 513–549, 1994.

[8] P. P. Maglio, M. J. Wenger, and A. M. Copeland, "Evidence for the role of self–priming in epistemic action: Expertise and the effective use of memory", Acta Psychologica, vol. 127, pp. 72–88, 2008.

[9] T. P. Alloway, M. Corley, and M. Ramscar, " Seeing ahead: Experience and language in spatial perspective", Memory & Cognition, vol. 34, pp. 380–386, 2006.

[10] J. Drury, C. Cocking, S. Reicher, A. Burton, D. Schofield, A. Hardwick, D. Graham, and P. Langston, "Cooperation versus competition in a mass emergency evacuation: A new laboratory simulation and a new theoretical model", Behavior Research Methods, vol. 41, pp. 957–970, 2009.

[11] A. Frey, J. Hartig, A. Ketzel, A. Zinkernagel, and H. Moosbrugger, "The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting", Computers in Human Behavior, vol. 23, pp. 2026–2039, 2007.

[12] G. Gunzelmann, and J. R. Anderson, "Location matters: Why target location impacts performance in orientation tasks", Memory & Cognition, vol. 34, pp. 41–59, 2006.

[13] G. A. Radvansky, and D. E. Copeland, "Walking through doorways causes forgetting: Situation models and experienced space", Memory & Cognition, vol. 34, pp. 1150–1156, 2006.

[14] M. Bar-Eli, S. Avugos, and R. Raab, (2006). Twenty years of "hot hand" research: Review and critique", Psychology of Sport and Exercise, vol. 7, pp. 525-553, 2006.

[15] T. Gilovitch, R. Vallone, and A. Tversky, "The hot hand in basketball: On the misperception of random sequences", Cognitive Psychology, vol. 17, pp. 295-314, 1985.

[16] P. D. Larkey, R. A. Smith, and J. B. Kadane, "It's okay to believe in the "hot hand", Chance, vol. 2, pp. 22-30, 1989.

[17] S. C. Albright, "A statistical analysis of hitting streaks in baseball." Journal of the American Statistical Association, vol. 88(424), pp. 1175-1183, 1993.

[18] R. D. Clarke, "Streakiness among professional golfers: Fact or fiction?" International Journal of Sports Psychology, vol. 34, pp. 63-79, 2003.

[19] R. D. Clarke, "An analysis of streaky performance on the LPGA tour. Perceptual and Motor Skills", vol. 97, pp. 365-370, 2003.

[20] R. D. Clarke, "Examination of hole-to-hole streakiness on the PGA tour". Perceptual and Motor Skills, vol. 100, pp. 806-814, 2005.

[21] R. L. Wardrop, "Simpson's Paradox and the Hot Hand in Basketball." The American Statistician, vol. 49(1), pp. 24-28, 1995.

[22] A. Wilke, and H. C. Barrett, "The hot hand phenomenon as a cognitive adaptation to clumped resources." Evolution and Human Behavior, vol. 30(3), pp. 161-169, 2009.

[23] A. Tversky and D. Kahneman, "Judgement under uncertainty: Heuristics and biases". Science, vol. 185(4157), pp. 1124-1131, 1974.

[24] A. Tversky and D. Kahneman "Belief in the law of small numbers". Psychological Bulletin, vol. 76 (2), pp. 105–110, 1971.

[25] G. Smith, G. Horseshoe pitchers' hot hands. Psychonomic Bulletin and Review, vol. 10 (3), 2003.

[26] R. M. Adams, "Momentum in the performance of professional tournament pocket billiards players." International Journal of Sport Psychology, vol. 26, pp. 580-587, 1996.

[27] R. Dorsey-Palmateer and G. Smith, "Bowlers' hot hands." The American Statistician, vol. 58(1), pp. 38-45. 2004.

[28] M. Raab, B. Gula, and G. Gigerenzer, "The Hot Hand Exists in Volleyball and Is Used for Allocation Decisions." Journal of Experimental Psychology: Applied, vol. 18(1), pp. 81-94, 2012.

[29] G. R. J. Hockey, "Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework." Biological Psychology, vol 45, pp. 73-93. 1997.

[30] D. Kahneman, *Attention and Effort*. New Jersey: Prentice-Hall, Inc 1973.

[31] N. Lavie, "Perceptual load as a necessary condition for selective attention." Journal of Experimental Psychology: Human Perception and Performance, vol 21(3), pp. 451-468. 1995.

[32] N. Lavie, "Distracted and confused?: Selective attention under load." TRENDS in Cognitive Sciences, vol 9(2), pp. 74-82. 2005.

[33] C. M. Farmer, A. K. Lund, R. E. Trempel, and E. R. Braver, (1997). "Fatal crashes of passenger vehicles before and after adding antilock braking systems." Accident Analysis and Prevention, vol 29, pp. 745-757. 1997.

[34] G. J. S. Wilde, *Target Risk* (1st ed.). Toronto, Ontario, Canada: PDE Publications. 1994.

[35] G. J. S. Wilde, D. Gerszke, and L. Paulozza, "Risk optimization training and transfer." Transportation Research Part F: Traffic Psychology and Behaviour, vol 1(1), 77-93. 1998.

[36] J. M. Rao, "Experts perception of autocorrelation: The hot hand fallacy among professional basketball players." Unpublished manuscript,

Department of Economics, University of California, San Diego, United States. 2009.
[37] J. W. Sammon "A nonlinear mapping for data structure analysis." IEEE Transactions on Computers vol. 18, pp. 401–409, 1969.
[38] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
[39] P. Williams, K. Nesbitt, A. Eidels, and D. Elliot, "Balancing Risk and Reward to Develop an Optimal Hot-Hand Game". Game Studies, vol. 11(1). ISSN:1604-7982, 2011.
[40] Atari Inc. Adventure. [Atari 2600], USA: Atari Inc. 1979.
[41] Valve Corporation. Left 4 Dead. [Windows], USA:Valve Corporation. 2008.
[42] Remedy Entertainment. Max Payne. [Windows], USA:Gathering Of Developers. 2001.
[43] P Demasi, and A Cruz, "Online Coevolution for Action Games." Proceedings of The 3rd International Conference on Intelligent Games And Simulation. London. pp. 113–120. 2002.
[44] M. Zuckerman, *Sensation seeking: Beyond the optimal level of arousal.* Hilldale, New Jersey: Erlbaum. 1979.
[45] M. Zuckerman, M. (2007). "The sensation seeking scale V (SSS-V): Still reliable and valid." Personality and Individual Differences, vol 43(5), pp. 1303-1305. 2007.
[46] G. Hawkins, K. Nesbitt, and S. Brown, "Dynamic Difficulty Balancing for Cautious Players and Risk Takers." International Journal of Computer Games Technology, vol. 2012, Article ID 625476, 10 pages, 2012. doi:10.1155/2012/625476.
[47] T. T. Hills, P. M. Todd, and R. L. Goldstone, "Search in external and internal spaces." Psychological Science, vol 19, pp.676-682. 2008.

Paul Williams is undertaking a PhD in Cognitive Psychology at the University of Newcastle, under the supervision of Dr. Ami Eidels. He is interested in developing online gaming platforms suitable for the investigation of how people change their behaviour in response to successes and failures. He is currently focused on refining a novel paradigm to study the behavioral phenomenon's known as the "hot hand" and "post-error slowing".
Paul.Williams@newcastle.edu.au

Dr Keith Nesbitt is a Senior Lecturer at the School of Design, Communication and IT, University of Newcastle. His research focuses on issues of perception and cognition applied to the areas of visualization, multi-sensory displays, virtual environments, computer games and conceptual modeling. You can access his website here: http://www.knesbitt.com
Keith.Nesbitt@newcastle.edu.au

Dr. Ami Eidels is a Lecturer in Cognitive Psychology, at the School of Psychology, University of Newcastle. He is also a principle investigator in the Newcastle Cognition Lab. His research focuses on visual perception and attention, combined with computational and mathematical modeling. You can access his lab's website here: http://newcl.org/eidels
Ami.Eidels@newcastle.edu.au

Mark Washburn is a former honours student in the Newcastle Cognition Lab, School of Psychology, University of Newcastle. Under the supervision of Dr. Ami Eidels he helped to design, develop, and test the Buckets game.
Mark.Washburn@uon.edu.au

Dr David Cornforth is with the School of Design, Communication and IT, University of Newcastle. His research interests are in health information systems, pattern recognition, artificial intelligence, multiagent simulation, and optimization. He is convenor of the Applied Informatics Research Group, University of Newcastle.
David.Cornforth@newcastle.edu.au

Chapter 3

In Chapter 2 we documented the Buckets game, a platform purpose built to explore the hot hand effect in terms of both speed and accuracy. I also outlined how the Buckets game design allowed us to draw a connection between the hot hand and post-error slowing, so that we might explore them simultaneously. Chapter 3 documents this simultaneous exploration of the hot hand effect and post-error slowing. Chapter 3 contains three components. The first component is a targeted literature review of post-error slowing, which briefly introduces our motivation to explore this effect using the Buckets game. The second component is the *Paper 3 Overview*, which highlights the unique contributions of Paper 3. This overview might be best read in conjunction with *Paper 3*, which is presented in full to conclude the chapter.

Post-error Slowing in Non Rapid-choice Tasks

Post-error slowing describes systematic increases in response time following an error in rapid choice tasks. Seminal findings of *post-error slowing* (Laming, 1968, 1979a, 1979b; Rabbitt, 1966,1969, 1979; Rabbitt & Rodgers, 1977; Rabbitt & Vyas, 1970, 1981) have been typically interpreted to suggest an increase in response caution is applied following errors. Recently however, some basic aspects of post-error slowing have come under scrutiny. These include how post-error slowing is best measured, and the relationship between post-error adjustments in response speed and accuracy. This relationship (or sometimes lack thereof) between post-error adjustments of response speed and accuracy has been used to support various causal explanations of the effect (e.g., Botvinick et al., 2001; Danielmeier & Ullsperger,

2011; Notebaert, Houtman, Opstal, Gevers, Fias, & Verguts, 2009). In plain words, it has been argued that the typical explanation of post-error slowing – increased caution following an error – should predict an increase in accuracy associated with post-error slowing. Because violations of this principle are regularly documented, the caution explanation has been scrutinised and alternate explanations such as the Orientating account (Notebaert et al, 2009), have been proposed. This material is discussed in Paper 3, so I will not cover it in great depth here. It suffices here to note that uncertainty surrounding the underlying causes of post-error slowing, and how best to measure it, has dominated recent literature.

This focus on the underlying causes of post-error slowing, and how it is best measured, means that little attention has been paid to the question of whether post-error slowing extends beyond tasks that require a rapid response after the presentation of simple stimuli. Yeung and Summerfield (2012) acknowledged this lacuna by noting that the current post-error literature was almost entirely focused on rapid choice tasks. They speculated that rapid choice benchmark findings might not scale up to goal driven and temporally extended tasks.

While this speculation remains untested empirically, some indirect evidence can be found that supports the speculations of Yeung and Summerfield (2012). For example, a standard finding in rapid choice tasks is that post-error slowing is largest for brief (or no) inter-trial intervals, and diminished or absent for inter-trial intervals greater than around 700ms (Danielmeier & Ullsperger, 2011; Jentzsch & Dudschig, 2009). This suggests that post-error slowing as currently demonstrated - while an important index of cognitive control – is not the result of participants making considered adjustments to their level of caution in response to errors. Yeung and Summerfield would describe this type of considered adjustment in caution to be a

meta-decision. If post-error slowing were the result of a meta-decision, allowing participants more time following errors should either not affect, or increase, post-error slowing. Whatever the underlying causes then, post-error slowing, as currently documented, seems a non-deliberate regularity of fast-paced sequential tasks that require a rapid response.

This is not to say that post-error slowing may not occur as a meta-decision in other tasks, but rather, these investigations (to our knowledge) have not been undertaken. Given that it is common to undertake tasks that require seemingly deliberate actions in the establishment and ongoing maintenance of speed-accuracy trade-offs (e.g., driving, daily tasks) or risk-reward trade-offs (e.g., investing, gambling, sports games), this omission seems puzzling. For example, while adopting more caution after a driving error seems intuitive (a-la the default explanation for post-error slowing), there is little evidence to support this position. The classical question posed by Rabbitt and Rodgers (1977) then; "What does a man do after he makes an error?" [1] (pp. 1) has been addressed, up until now, in a very narrow fashion. We tackle this lacuna in Paper 3, by utilising our cognitive game to explore post-error slowing and the hot hand simultaneously.

Paper 3 Overview

Paper 3 documents our simultaneous exploration of the hot hand and post-error slowing using the *Buckets game*. This paper represents the culmination of the development and application of our cognitive game paradigm, a process that spanned several years. The paper made several notable contributions to the literature. First and

[1] Recently, an insightful professor suggested 'asks a woman for help' as the best answer to this long-standing question.

foremost, we outlined the theoretical foundations for exploring the hot hand belief and post-error slowing simultaneously. Secondly, we documented that our novel cognitive game environment was suitable for this exploration. Lastly, our data provided several important empirical findings and thus provided a platform for further experimentation and theoretical development. Given that Paper 3 represents the culmination of several years of research and development, it is perhaps the achievement that I am most proud of in this thesis.

To demonstrate the viability of simultaneous exploration of the hot hand effect and post-error slowing, Paper 3 initially outlines the strong theoretical and empirical links between the two fields. We then noted that despite these strong similarities, the hot hand and post-error slowing had been studied over vastly differing timescales and environments. The discrepancy between the timescales and environments was of such magnitude that data from one domain could not easily inform the other; this meant that questions regarding the generalizability of findings remained in each. We therefore proposed the Buckets game as a means to better assess the similarities and differences between post-error slowing and the hot hand, as it provided a unique middle ground – an intermediate timescale and environment. Importantly, players of the Buckets game were paid or unpaid, which allowed to assess the impact of motivation – a variable highlighted as important across the hot hand and post-error slowing literature.

Summary and Transition

In Paper 3 our results showed some behavioural signatures consistent with the post-error slowing literature, such as a shift toward post-error slowing for paid players relative to unpaid players. This result was very important as it suggested that

irrespective of the environment, increased motivation is likely to result in an increased level of cognitive control. This observation is in line with the theoretical position of Botvinick and Braver (2015). As is generally the case however, it was the results that we could not predict based on pre-existing literature that were perhaps the most interesting. We found that unpaid players exhibited post-error *speeding* rather than slowing, and that paid participants showed neither post-error slowing nor post-error speeding. Our finding of post-error speeding for the unpaid group was especially surprising and rare.

This rare result was an especially important and novel contribution to the literature, as it provided the first empirical support for speculation that there may be substantial differences in post-error behaviour for goal driven and temporally extended tasks when compared to rapid choice tasks. We proposed our unpaid participants were generally unmotivated, and that rather than errors triggering a process in which cognitive control was recruited, errors actually decreased the level of cognitive control available. This might be considered *post-error recklessness*. For the group rewarded by monetary incentive we argued that the discouraging impact of errors was negated. In sum these results suggested an account of post-error speeding that rested on motivation (or more correctly, a lack of motivation). We noted the potential for this same explanation to account for other empirical findings of post-error speeding, such as those documented by Notebaert et al. (2009). Given the debate over post-error speeding and its relevance to exploring the causes of post-error slowing, future work could look to address this possibility.

With regards to the hot hand, we found the difficulty-accuracy trade-off resulted in a hot hand effect in our unpaid players. Estimated at approximately 5% improved accuracy given a hit, our unpaid players showed a hot hand effect larger

than any previous research we are aware of. This finding hinted the hot hand belief might well remain a fallacy at the professional level when motivation is high, but might be prevalent in amateur contexts when motivation is possibly lower. This could potentially explain the resilience of the hot hand belief in the face on contradictory evidence, as fans and players may have experienced the hot hand in amateur contexts.

Judgment and Decision Making, Vol. 11, No. 2, March 2016, pp. 174–184

Post-error recklessness and the hot hand

Paul Williams* Andrew Heathcote*† Keith Nesbitt* Ami Eidels*

Abstract

Although post-error slowing and the "hot hand" (streaks of good performance) are both types of sequential dependencies arising from the differential influence of success and failure, they have not previously been studied together. We bring together these two streams of research in a task where difficulty can be controlled by participants delaying their decisions, and where responses required a degree deliberation, and so are relatively slow. We compared performance of unpaid participants against paid participants who were rewarded differentially, with higher reward for better performance. In contrast to most previous results, we found no post-error slowing for paid or unpaid participants. For the unpaid group, we found post-error speeding and a hot hand, even though the hot hand is typically considered a fallacy. Our results suggest that the effect of success and failure on subsequent performance may differ substantially with task characteristics and demands. We also found payment affected post-error performance; financially rewarding successful performance led to a more cautious approach following errors, whereas unrewarded performance led to recklessness following errors.

Keywords: post-error slowing, hot hand, cognitive control, financial incentives

1 Introduction

The effects of recent outcomes on future performance have been the subject of considerable interest, mainly in two largely non-overlapping literatures about *post-error slowing* and the *hot hand*. Post-error slowing describes systematic increases in response time (RT) following an error in rapid choice tasks (Laming, 1968; Rabbitt, 1966a). The hot hand originated in sports, and describes an increase in the probability of success after previous success. The hot hand is often considered a fallacy as, despite the strong beliefs of spectators and players, the effect is not often empirically observed in professional sports (Gilovich, Vallone & Tversky, 1985; see also Avugos, Köppen, Czienskowski, Raab & Bar-Eli, 2013). Although the two phenomena are framed in terms of failure (post-error slowing) and success (the hot hand), both are measured by a difference between post-error and post-correct performance. From a measurement perspective, the key difference has been the primary dependent variable — RT for post-error slowing, and the probability of success for the hot hand. Recently, however, post-error slowing research has placed increased importance on the effect of errors on subsequent accuracy (e.g., Danielmeier & Ullsperger, 2011; Notebaert et al., 2009; Schroder & Moser, 2014). Hot hand research has also increasingly examined whether sports players attempt more difficult (e.g., quicker) shots following success, which may obscure improved performance if it is measured solely by accuracy (Bocskocsky, Ezekowitz & Stein, 2014; Rao, 2009). Thus, research on the hot hand and post-error slowing taps related questions.

Empirically, there are distinct similarities between recent hot hand findings and well-established regularities found in post-error slowing research. Rao (2009) used video analysis and found that basketball players attempted more difficult shots following a successful run. More recently, Bocskocsky, Ezekowitz and Stein (2014) employed enhanced tracking technology and found players on a "hot run" take more shots of higher difficulty, and perform at above expected performance levels if shot difficulty is taken into account. Although it is debateable whether the difficulty of complex actions such as basketball shots can be precisely quantified, the increased difficulty of basketball shots following success resembles performance in rapid-decision tasks where gradual speeding (analogous to more difficult shots) is observed over runs of correct responses that precede an error (Dudschig & Jentzsch, 2009; Laming, 1968; see Luce, 1986, for a review).

From a theoretical perspective, post-error slowing (Laming, 1968, 1979; Rabbitt, 1966a, 1966b, 1969; Rabbitt & Rodgers, 1977) was initially considered the result of an increase in caution following errors. The *caution explanation* also suggested that following success less caution is exercised, and response times get faster (Dudschig & Jentzsch, 2009; Jentzsch & Dudschig, 2009; Laming, 1968). Formal models of decision-making response-time (Dutilh et al., 2012a) and cognitive control (e.g., Botvinick, Braver, Barch, Carter & Cohen, 2001) have since established that increased

ARC-Discovery Project grants to AH and AE, ARC Professorial Fellowship to AH Keats Endowment Fund grant to PW and AE.

*The University of Newcastle School of Psychology, University of Newcastle, Callaghan, NSW, 2308, Australia. Email: paul.williams@newcastle.edu.au.

†The University of Tasmania

response times following errors can often be causally linked with a higher response criterion following instances of high response conflict, including errors.[1] The caution explanation aligns with the hot hand framework of Bocskocsky, Ezekowitz and Stein (2014), who noted that basketball players might be less cautious following successes. Hence, basketball players and experimental participants alike potentially employ less caution following success and more caution following errors. The level of caution adopted following success relative to failure is, therefore, central to both domains.

Despite these similarities, the hot hand and post-error slowing have typically been studied over greatly differing time scales and across very different environments. Post-error slowing research has been narrowly focused on simple and rapid choice tasks with high levels of experimental control (Yeung & Summerfield, 2012). In contrast, hot-hand research has mainly focused on uncontrolled sporting tasks (e.g., shooting a basketball) that unfold over longer (and often irregular) time scales. These narrow foci leave open questions in each field regarding the generalizability of findings. For example, Yeung and Summerfield (2012) questioned the degree to which the current body of post-error literature might scale up to explain decisions that are goal driven and temporally extended. Similarly, Bocskocsky, Ezekowitz and Stein (2014) found empirical evidence in support of theoretical speculation that basketball shooters attempt more difficult shots following success — yet the possibility that this finding generalizes as a behavioural regularity remains untested. In sum, each domain has had a narrow focus, and these foci are so widely separated that it is unclear whether post-error experimental findings might shed light on goal-driven behaviours in more complex environments (such as sporting performance), and vice versa.

To better assess the similarities and differences between post-error slowing and the hot hand, data are required that connect the two domains. Here we collected such data using the *Buckets game*, a computerized task, created by Williams, Nesbitt, Eidels, Washburn and Cornforth (2013), that utilises an intermediate time scale connecting post-error slowing and hot hand research. Participants were presented, on each trial, with four rectangular "buckets", each half-filled with randomly positioned pixels. Over time, one of the buckets (target) accrues more pixels and gradually become fuller, while the other buckets (distractors) remain half filled. The task of identifying the target bucket, therefore, becomes easier as the trial progresses. The defining features of this game are that it presents temporally extended decisions (trials lasted up to 8 seconds), and that players can elect to respond more quickly with less chance of being correct, or more slowly with a higher chance of being correct. That is, players self-selected the level of difficulty they were willing to assume for each attempt. The goal of the game is to maximise the number of correct decisions in a fixed time period. Hence, responding quickly offered the benefit of more attempts overall, but at the risk of lower accuracy. Williams et al. (2013) described in detail how the game's timing and incentive system were tuned. Players were explicitly informed that they control the difficulty of each attempt and that they can trade-off between difficulty and speed to maximize their overall performance. Because the speed-accuracy trade-off was explicit, the task slow-paced, and each individual attempt was embedded within the overarching global context of maximising correct decisions in a limited space of time, the task lent itself to deliberative post-error adjustments. In the language of Yeung and Summerfield (2012), the task encouraged meta-cognitive judgements.

With respect to post-error slowing, the Buckets game allows us to assess post-error adjustments in a relatively simple but goal-driven task that unfolds over up to 8 seconds. With respect to the hot hand, the Buckets game allows expansion of the recent work of Bocskocsky, Ezekowitz, and Stein (2014), who found professional basketball shooters attempted more risky or difficult shots after previous successes. We can assess in a controlled environment whether this finding reflects a systematic behavioural trend. Furthermore, if players systematically adopt more or less risk following success or failure, we can assess how this affects detection of the hot hand.

An important consideration in using the Buckets game is that participants are motivated to achieve its goals. Psychologists and economists hotly debate the benefits of financial incentives and how such incentives influence intrinsic and extrinsic motivation (e.g., Read, 2005; Camerer & Hogarth, 1999). Less controversial is the empirical finding that financial incentives do alter performance systematically in cognitive tasks (Botvinick & Braver, 2015; Camerer & Hogarth, 1999). For mundane laboratory tasks, financial incentives improve motivation and performance (Cameron, Banko & Pierce, 2001; Camerer & Hogarth, 1999; Kouneiher, Charron & Koechlin, 2009; Padmala & Pessoa, 2011). Further, monetary rewards seem to facilitate performance to a greater extent when incentives are contingent upon the level of performance (Bonner, Hastie, Sprinkle & Young, 2000). Botvinick and Braver (2015) described improvements in cognitive task performance due to financial incentives as a fundamental phenomena that links motivation to cognitive control. That is, financial incentives increase the level of cognitive control available for a task, which is turn improves performance.

Botvinick and Braver (2015) note that fluctuations in cog-

[1] Note in other instances, increased response times following errors have been linked to multiple causes (Dutilh, Forstmann, Vandekerckhove & Wagenmakers, 2013), the need to re-orient to the task following errors (Notebaert et al., 2009), or an increase of inhibition (Ridderinkhof, 2002).

nitive control linked to motivation are observed not only in overall performance but also at short, trial-by-trial, time scales. Indeed, post-error adjustments are typically considered a fundamental aspect of cognitive control (Botvinick, Braver, Barch, Carter & Cohen, 2001; Gehring & Fencsik, 2001; Ridderinkhof, Van Den Wildenberg, Wijnen & Burle, 2004). It is not surprising, then, that, like overall performance, post-error adjustments have been empirically linked to financial incentives and motivation. Sturmer, Nigbur, Schact and Sommer (2011) found that performance contingent on incentives led to an increase in post-error slowing, which is commensurate with findings of increased post-error slowing when financial rewards were tied to more accurate performance (Ullsperger & Szymanowski, 2004).

Motivation has also been of interest in the hot hand domain. For example, null results from experimental investigations (e.g., Gilden & Wilson, 1995) have been criticised because they were not collected from highly motivated participants typical of professional sporting settings (Smith, 2003). Thus, motivation and its effects on control and performance are of interest to both post-error slowing and the hot hand. Given the potential importance of motivation, we compared paid performance to unpaid performance in the Buckets Game. Payment was contingent on performance — participants received one point for each correct response, and higher overall scores received greater financial reward.

The post-error slowing literature suggests that participants may adopt a more cautious approach following errors, and the hot hand literature suggests that players adopt a more risky approach following success; we therefore expected, for both paid and unpaid players, post-error slowing and post-success speeding (which are equivalent results). Because the target became easier to identify over time, we expected this additional caution following errors to result in higher accuracy following errors and lower accuracy following success. Note this is the reverse of predictions based on belief in the Hot Hand. We expected financial incentives to exaggerate this post-error slowing and reversal of the hot hand. That is, we expected the performance-contingent incentives (higher financial reward for higher game scores) to enhance goal motivation and result in higher levels of cognitive control — observed as (1) overall improved performance and (2) increased post-error slowing.

As a caveat, we note that Bocskocsky, Ezekowitz and Stein (2014) reported basketball players took more risk following success with little or no reduction in accuracy. Therefore, it is possible in the Buckets game that post-error slowing and post-success speeding would not be associated with any appreciable change in accuracy. This result — more difficult attempts for no loss of accuracy — would indicate an overall increase in performance following success, consistent with Bocskocsky et al.'s view of the hot hand.

2 Method

2.1 Participants

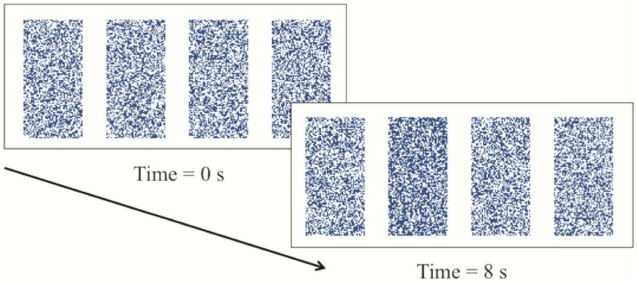
Sixty-seven undergraduates from the University of Newcastle, Australia, took part in the experiment, with 42 rewarded by course credit that was not contingent on performance. Of the 42 rewarded by course credit, 21 participated on campus in experimental testing rooms, while 21 participated online in their own time. At the beginning of the session for on-campus participants, we provided a verbal explanation of the game and encouraged them to remain motivated throughout the experiment. Despite these instructions, we found no differences between the on-campus and online sampling methods.[2] This is in line with findings that these two sampling methods produce equivalent results for both cognitive (Crump, McDonnell & Gureckis, 2013) and other psychological research (Gosling, Vazire, Srivastava & John, 2004). The remaining 25 players were undergraduate students, not limited to psychology, and recruited via posters placed around the campus. They participated in our testing rooms, and were paid $10 plus 5 cents per correct target-identification, with a maximum possible payment of $20. In addition to the standardised on-screen instructions, they received a verbal explanation of the game and the payment structure. We label the two groups in terms of reward: paid and unpaid.

2.2 The Buckets game

The Buckets game was coded in actionscript for Adobe Flash, an easily distributable platform that records response times with an adequate precision for our purposes (Reimers & Stewart, 2007). In the Buckets game, four 100x50 pixel rectangles ('buckets') were displayed on a computer screen, each with 50% of its pixels filled blue (blue dots). The location of blue pixels within buckets was randomly updated every 100ms, and one of the buckets was slowly filled with more blue pixels. The player was asked to identify this target (see Figure 1). The target received additional blue pixels at an average rate of 1.875 pixels per 100ms update. Players could select the target and hence terminate the trial at any time during the maximum trial duration of 80 updates (or equivalently, 8sec). A fixation-cross preceded trials and lasted 300ms. Visual (i.e., "CORRECT" or "INCORRECT") and auditory (i.e., cash register "ker ching", or incorrect buzz) feedback, lasting 500ms, was provided on the accuracy of each attempt, followed by a between-trial

[2] Two-tailed Bayesian independent samples t-tests, performed as per the analysis comparing paid and unpaid players below, and reported in favour of the alternate hypothesis, indicated no evidence for the hypothesis of a difference between on-campus and online sampling for post-error RT adjustments [traditional: $BF = 1.01$; robust method: $BF = 1.53$; matched: $BF = 0.62$)], or post-error accuracy adjustments [(traditional: $BF = 0.30$; robust: $BF = 0.89$); matched: $BF = 0.31$].

Figure 1: An example of the evolution of Buckets game stimuli. Initially all buckets have the same number of blue pixels. One bucket accumulates additional blue pixels over the course of an 8s trial (unless a response was given beforehand). Note players could terminate the trial at any stage by making their selection. Additional pixels were added, and the location of pixels within each bucket was randomly updated, every 100ms. The target, 2nd from left, has been accentuated for the purposes of demonstration.



white screen for 500ms. An additional 1,650ms of between-trial white screen was applied to incorrect attempts (i.e., a time-out penalty applied to balance the reward function). If a player had not responded already, a tone briefly sounded 6,000ms after the buckets appeared to make players aware that the end of each trial was approaching.

Players undertook five time-limited blocks, each separated by enforced breaks of minimum 30s duration. The first block was a 5 min practice that did not count in the final score, and players were encouraged to use this block to explore the relative benefits of making attempts at different time points throughout a trial. The final four blocks were each 10 mins in length. The total game score was the sum of correctly identified targets over the four 10 min blocks.

On-screen instructions indicated the aim of the game was to identify as many targets as possible within the time allocated. On-screen instructions also made explicit that faster, and so more difficult responses, allowed for more attempts overall, but at a higher risk of making errors. During play, a countdown clock indicated the number of seconds remaining in the block, and a counter indicated the number of correct decisions made during the current block. Between blocks, players were provided updates on their previous block performance, and overall performance.

2.3 Analyses

Post-error adjustments. There are several methods in the literature for measuring post-error adjustments. We used three — *traditional*, *robust*, and *matched*. Each method involves calculating a difference between post-error and post-correct performance. It is useful to note that post-error slowing (PES) measured in this way can also be considered post-correct speeding. The traditional method involves subtracting, for each participant, the mean RT of the post-error trials from the mean RT of the post-correct trials. Similarly for accuracy, it subtracts the conditional probability of a hit preceded by a hit from the conditional probability of a hit preceded by a miss.

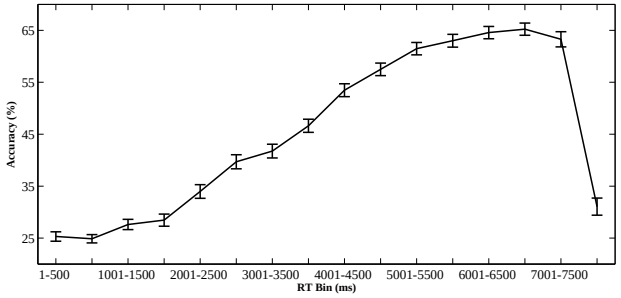
The other two measures address drawbacks of the simple global averaging used by the traditional method. One drawback is that short-term effects like post-error slowing can be confounded with long-term effects like fatigue, distraction, or boredom (Dutilh et al., 2012b). Dutilh et al. (2012b) proposed a solution that paired post-error trials with immediately preceding pre-error counterparts that are also post-correct trials. Pairwise differences are then calculated (i.e., [post-error RT] minus [pre-error, post-correct RT]), with the mean of the differences providing a robust measure of post-error RT adjustments. Dutilh and colleagues showed the robust method is able to differentiate true post-error adjustments from confounding long-term effects. Dutilh et al. (2013) employed the same type of pairs to calculate post-error accuracy adjustments. We describe these RT and accuracy based differences as "robust" measures.

A second drawback of the traditional method is that it can be confounded by systematic differences in the relative speed of correct and error responses (Hajcak & Simons, 2002). Consider a participant who slows down after all fast responses. This participant is not adjusting to errors, but is sensitive to the speed of their previous response. If errors are faster than correct responses — as is the case in the Buckets game — the traditional method of calculating post-error adjustments spuriously indicates post-error slowing. To counter such confounds, Hajcak and Simons paired each error response with a correct response closely matched on RT. We used such pairs[3] in the same way that pairs were used in the robust method, to calculate what we call "matched" measures based on both RT and accuracy.

Statistical comparisons. Null hypothesis tests cannot provide evidence in favour of the null, which is problematic because providing evidence for both null and alternate hypotheses is useful in assessing our results. Therefore, we performed all statistical comparisons using the Bayesian approach implemented in the BayesFactor package for R (Morey & Rouder, 2014), as called by JASP (Love et al., 2015) — the user friendly graphical interface for common statistical analyses. The Bayesian approach allows quantification of evidence in favor of either of the hypotheses, and each test produces a Bayes Factor (BF) that indicates

[3] In particular, we selected the closest matching but faster correct RT for odd errors (i.e., the first, the third, the fifth error, and so on, in terms of serial location), and the closest matching but slower correct RT for even errors. If a match within 30ms was not available this error was discarded from analysis. In the event of multiple identical RT matches, a random selection was made from those available. In the event that there were more errors than correct trials in a given data set, we began with the less common correct responses and searched for matching errors.

Figure 2: The probability of a correct decision by response time in the Buckets game. Error bars show the standard error of a proportion.



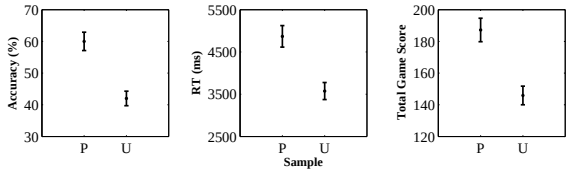
the factor by which prior beliefs should be changed by the data. We use the classification scheme proposed by Jeffreys (1961) to describe BF results. Unless otherwise specified we employed a default Cauchy prior width of r = 1 for effect size, as specified by Rouder et al. (2009) and Wetzels et al. (2009).

3 Results

We first checked whether accuracy improved as participants waited for more dots to accumulate in the target bucket. Accuracy by time for all attempts is shown in Figure 2. Accuracy increased as expected for 0–7,000ms, but then plateaued, and dropped steeply for responses slower than 7500ms. Errors that are slower than 7500ms included both non-attempts (failure to beat the deadline) as well as incorrect attempts. Given the high error rate, we suspected the looming deadline led to late guesses. Because it was impossible to identify and separate late guesses from "proper", non-guess attempts that resulted in errors (or hits, for that matter), we removed contributions to post-error adjustment calculations that relied on responses slower than 7500ms, which were 3% of all attempts, 25% of which were non-attempts. For example, for the robust method, if any e–2, e–1, e, or e+1 response was slower than 7500ms, where e indicates the trial index of an error, the pre- and post-error paired difference for this quartet of trials was removed from analysis. We also removed contributions relying on responses faster than 500ms, which were 8.8% of all attempts, 44% of which come from the 2 participants who are subsequently excluded. Based on players' self-report these very fast responses represented guesses.[4]

[4]Our findings were robust against variations in the exclusion criteria. To check, we re-ran our post-error analyses for accuracy and RT changes (as seen in Figures 4 and 5, and reported in corresponding text) for three different exclusion scenarios. Under scenario 1 no responses were excluded, and for scenario 2 only responses slower than 7500ms were excluded. For both scenarios we found an unchanged pattern of RT and accuracy post-error results, and statistical reliability increased for the critical RT results. Under scenario 3 only responses faster than 500ms were excluded. Here we again found an unchanged pattern of RT and accuracy post-error results, however, the statistical reliability of RT results decreased for the traditional (BF = 7.55) and matched (BF = 2.12) methods, but increased for the robust method (BF = 20.4). No participants were excluded under scenarios 1 and 2 whereas two participants were excluded under scenario 3.

Figure 3: Accuracy, mean RT, and total game score for paid (P) and unpaid (U) players. Error bars show the standard error of the mean.



We then confirmed that each participant had enough remaining responses to calculate post-error and hot-hand measures. For the traditional method we required that each participant contributed at least 20 errors and 20 correct responses. One player from the unpaid group failed to meet these criteria, having made many responses faster than 500ms. For the robust measure, we required at least 20 suitable pairs. One additional player from the unpaid group was excluded due to too many fast responses. For the matched measure, we also required 20 pairs, with no further exclusions required. This left 45 and 20 participants in the unpaid and paid groups, respectively.

Figure 3 demonstrates the efficacy of our payment manipulation, with higher accuracy, slower responding and higher overall game scores in the paid group. We used one-sided Bayesian independent samples t-tests to quantify the evidence for the hypotheses that the paid participants would be slower, more accurate, and accumulate higher overall scores. Here we report Bayes factors in favour of the alternate hypothesis. The Bayes factors were BF = 271, BF = 3780, and BF = 894 respectively, indicating that the observed data were much more likely under the alternative hypothesis that postulates an effect of payment than under the null hypothesis that postulates the absence of the effect. This is decisive evidence in each case. We conclude that paid players were more focused on achieving the game goals than players from the unpaid group.

3.1 Post-Error Analysis

Post-error response-times. For all methods, post-error adjustments were calculated on an individual basis. Figure 4 displays the results for post-error RT analysis and highlights two important results. Firstly, the direction of the difference between paid and unpaid participants was in line with expectations. Secondly, and surprisingly, no post-

Figure 4: Post-error (PE) RT adjustment in the Buckets game for paid (P) and unpaid (U) participants, for each of the traditional, robust, and matched measurement methods. The y-axis represents post-error RT adjustment. Above zero values indicate post-error slowing, or more caution following an error. Below zero values indicate post-error speeding, or less caution following an error. The errors bars indicate the standard error of the mean. The horizontal lines indicate the 95% credible interval for the mean.

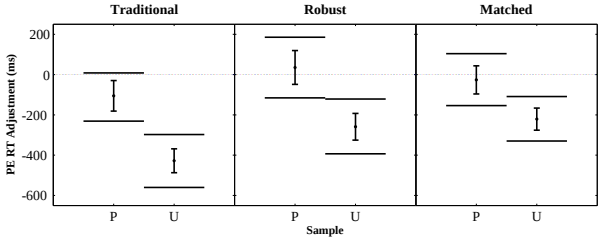
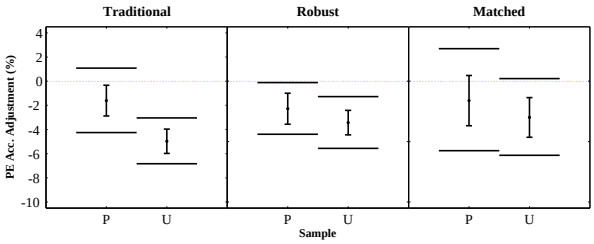


Figure 5: Post-error (PE) accuracy adjustment in the Buckets game for paid (P) and unpaid (P) players, for each of the traditional, robust, and matched measurement methods. The y-axis represents post-error accuracy adjustment. Above zero values indicate more accurate identification of the target following an error. Below zero values indicate more accurate identification of the target following success, or a hot hand. The errors bars indicate the standard error of the mean. The horizontal lines indicate the 95% credible intervals for the mean.



error slowing was observed for any of the groups and regardless of the method of calculation. Instead, considerable *post-error speeding* was documented for the unpaid group. This is supported by 95% credible intervals that indicate the unpaid group showed reliable post-error speeding for the traditional, robust, and matched methods. In contrast, the paid players showed no reliable post-error speeding for any method, and a near zero post-error RT adjustment for the robust and matched methods. One-sided Bayesian independent samples t-tests, reported in favour of the alternate hypotheses, confirmed that the paid sample showed less post-error speeding for the traditional, robust, and matched methods (traditional: BF = 43.8; robust: BF = 10.2; matched: BF = 3.32). According to Jeffreys (1961), this is very strong, strong, and substantial evidence respectively for the alternative hypothesis that postulates payment will lead to more post-error slowing (or less post-error speeding) than under the null hypothesis that postulates the absence of the effect.

Post-error accuracy. Figure 5 displays results for post-error accuracy adjustments. There was a tendency for lower accuracy following errors, or equivalently, higher accuracy following success, that is, a hot hand. Overall, this tendency ranged from 2–5%, but for the paid sample this tendency was smaller and less reliable. One-sided Bayesian independent samples t-tests, reported in favour of the alternate hypothesis, confirmed that the traditional method provided anecdotal evidence for the alternate hypothesis of a larger decrease in post-error accuracy for the unpaid group (BF = 2.53), whereas the robust (BF = 0.36) and matched (BF = 0.30) methods showed anecdotal and substantial evidence respectively for the null hypothesis of no difference between post-error accuracy adjustments for paid and unpaid players.

With the RT adjustments reported above, it seems that unpaid players become more cautious and accurate following success, or less cautious and accurate following errors.

Short- and long term effects of errors. Figure 4 suggested, for both paid and unpaid groups, that the traditional method indicated more post-error speeding than the other methods, for both the paid and unpaid groups. Because the traditional method captures both short-term and long-term sequential effects, whereas the other methods focus specifically on the short-term effects of errors, we used this difference to estimate the relative influence of short- and long-term effects in the Buckets game. A three (measure: traditional, robust, matched) by two (sample: paid, unpaid) Bayesian mixed model ANOVA indicated evidence for the main effects of measure and group with this two factor model maximizing the marginal probabilities relative to the null model of no effects, JASP estimating the BF~90,000. While no single factor model was supported, it can be instructive to assess these models to shed light on the relative influence of the two factors. In terms of the two effects, measure had an extremely strong influence relative to the null, BF~9,000, whereas group had a less pronounced effect, BF~10. Bayesian paired samples t-tests, reported in favor of the alternate hypothesis and with posterior model odds calculated using model priors that were adjusted for multiple comparisons[5], provided de-

[5]For k comparisons we set the prior probability of finding no difference in a single comparison (p) so that the probability of finding no difference in the set of k comparisons equals the total probability of finding one or more differences. That is, we solve $p^k = 1/2 \Rightarrow p = 2^{-1/k}$. In the present case where k=3, p=0.794. So if BF is the Bayes factor for Difference vs. No Difference for a particular comparison then the posterior odd (which can be conceived as a corrected Bayes Factor for the multiple comparisons) is (1

cisive evidence that the traditional method showed more post-error speeding (marginal mean = –267ms) than both the robust (marginal mean = –112ms, BF = 331) and matched (marginal mean = –124ms, BF = 1657) methods, and strong evidence for no difference between the robust and matched methods (BF = 0.03). Thus, the short-term effects of errors accounted for approximately half of the post-error speeding seen in the Buckets game.

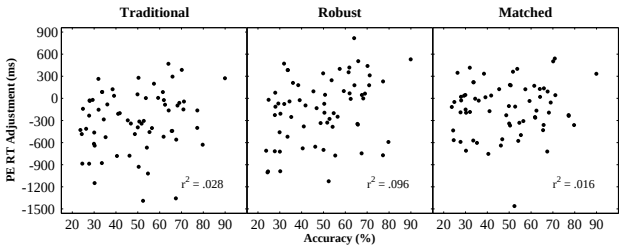
4 Discussion

We aimed to investigate sequential effects caused by the influence of previous-response success (or failure) on current performance. The Buckets game provided an intermediate time-scale and a carefully controlled environment so that both hot-hand and post-error statistics could be estimated from the same data. Players were either paid or unpaid, with payments structured to incentivise the Buckets game goal of maximising the number of correct target detections in a fixed time period. Past experimental investigations have typically found post-error slowing. In contrast, hot hand research has focused on professional sports settings. Although the hot hand is typically considered a fallacy (Avugos, Köppen, Czienskowski, Raab & Bar-Eli, 2013; Bar-Eli, Avugos & Raab, 2006), professional basketball players have been reported to take more difficult shots following success (Bocskocsky, Exekowitz & Stein, 2014; Rao, 2009), which would mask a hot-hand effect.

As expected, we found monetary rewards improved overall performance. We also found that financial incentives influenced post-error RT adjustments in the expected direction, toward post-error slowing. This is in line with previous findings that financial incentives improve performance in cognitive tasks (Kouneiher, Charron & Koechlin, 2009; Padmala & Pessoa, 2011; Bonner, Hastie, Sprinkle & Young, 2000; Camerer & Hogarth, 1990) and increase post-error slowing (Sturmer, Nigbur, Schact & Sommer, 2011; Ullsperger & Szymanowski, 2004). Our work provides an extension of these previous findings in that the shift we observed toward post-error slowing for paid players occurred in a novel and temporally extended task. This result was encouraging with regard to our primary theoretical investigation, it suggested that behaviour in the Buckets game — an intermediate environment between those typically used to study post-error slowing and the hot hand — showed behavioural signatures consistent with the post-error slowing literature. Thus, our data supported the broader position that increased motivation will result in an increased level of cognitive control, regardless of task (Botvinick & Braver, 2015).

p)/p X BF. For example, suppose that BF = 10 (i.e., data changes our belief by a factor of 10 in favor of a difference) then the posterior odds are 10(1 0.794)/0.794 = 2.6.

Figure 6: Post-error RT change by accuracy for the traditional (left), robust (middle), and matched (right) methods. R-squared indicates the proportion of variance in post-error RT changes accounted for by accuracy.



Importantly though, we found that unpaid players exhibited post-error speeding rather than slowing, and that paid participants showed neither post-error slowing nor post-error speeding. These results suggest the influence of the prior outcome may be quite different in an environment such as the Buckets game to those typically used to investigate post-error slowing or the hot hand. In regards to post-error slowing, our finding of post-error speeding for the unpaid group was especially surprising and rare.

Notebaert et al.'s (2009) *orienting account* of post-error slowing provides a potential reconciliation of this surprising result. Notebaert and colleagues proposed participants are surprised and distracted by errors when they are rare — the usual case in most post-error slowing research — and are slowed because they must reorient to the task after committing errors. Conversely, when success is rare, the orienting account predicts post-error speeding, a prediction that has been confirmed in some rapid-choice tasks when errors are more common than correct decisions (e.g., Houtman, Núñez Castellar & Notebaert 2012; Núñez Castellar, Kühn, Fias & Notebaert, 2010). Consistent with this account, in our data error rates were higher for unpaid participants (average rate of 58%) than they were for paid participants (average rate of 40%). Therefore, according to the orienting account, error rates for unpaid participants were in the region that might encourage post-error speeding, whereas error rates for paid participants were in the region that might encourage no post-error slowing.

The orienting account does not predict a difference in post-error behaviour based on the level of motivation. However, it makes another testable prediction, namely a positive relationship between the overall rate of errors and post-error RT adjustments, with post-error slowing increasing with increased accuracy. To test the orienting account we investigated the relationship between accuracy and post-error RT adjustments for all players in the two groups, paid and unpaid. Figure 6 shows that accuracy explains very little of the variance in any of the three measures of post-error RT adjustments. Bayesian tests of correlation, using a Beta prior

width of 1 and reported in favor of the alternate hypothesis, confirmed that for both the traditional (BF = 0.37) and matched methods (BF = 0.25), there was evidence for the null hypothesis of no relationship between accuracy rate and magnitude of post-error adjustments. For the robust method (BF = 3.34), there was evidence in favor of a relationship. A potential reason for a lack of orienting effects is that the Buckets game had a minimum 1,300ms inter-trial-interval and a possible 8,000ms trial time. These longer time scales may have negated the impact of re-orientation. In any case, the orienting account cannot explain post-error speeding in the Buckets game.

With the orienting account excluded, a lack of post-error slowing for paid participants and the post-error speeding observed for unpaid participants suggest there are substantial differences between post-error behaviour in the Buckets game and in typical rapid-choice tasks. Future work could explore the specific task demands that are responsible for this lack of post-error slowing. To this end, it is useful to note that the Buckets game — while a novel intermediate step between rapid choice and sporting environments — is related to two other paradigms. First, it is related to the expanded judgement task developed by Irwin, Smith and Mayfield (1956), and in particular the information-controlled expanded-judgement tasks used by Brown, Steyvers and Wagenmakers (2009), and Hawkins, Brown, Steyvers and Wagenmakers, (2012). In these tasks, evidence for a target item among distractors is accumulated stochastically on screen in discrete time steps. As in the Buckets game, information toward the correct decision accumulated slowly over time, and the longer a participant waited before responding, the more likely they were to correctly identify the target. These tasks closely resemble the temporally extended nature of the Buckets game. Second, the goal driven structure of the Buckets game, in which players were asked to maximise the number of successes within a fixed time period, is related to rapid choice tasks used to investigate reward-rate optimization (e.g., Bogacz, Hu, Holmes & Cohen, 2010; Simen et al., 2009). It would be interesting, therefore, to examine post-error effects in these paradigms. In any event, a lack of post-error slowing in the Buckets game provides the empirical evidence in support of the speculations of Yeung and Summerfield (2012); there may be substantial differences in post-error behaviour for goal driven and temporally extended tasks when compared to rapid choice tasks.

Given the differences between paid and unpaid post-error performance, our data support an account of post-error speeding in the Buckets game that rests on participant motivation. We propose our unpaid participants were generally unmotivated, and rather than recruiting cognitive control, errors further decreased the level of cognitive control available. In other words, we propose that in the Buckets game environment — which we note had relatively low success rates — unpaid participants were discouraged by errors and consequently made less cautious responses, whereas success encouraged them to try harder. This might be considered "post-error recklessness". For the group rewarded by monetary incentive however, cognitive control was enhanced — as evidenced by better overall performance — and the discouraging impact of errors was negated, explaining why we observed post-error speeding for unpaid, but not paid, participants. It may have been that there were individual differences in the motivating effect of financial incentives, so that some paid participants were motivated to increase caution after errors, but some were discouraged by them as in the unpaid group, so that on average there was no post-error slowing. Future research might directly measure motivation in order to check whether it correlates with the level of post-error slowing. Future work may also consider whether similar mechanisms contribute to post-error speeding observed in rapid choice when error rates are very high.

With regards to the hot hand, unlike professional basketball where post-success increases in shot difficulty may mask the hot hand (Bocskocsky, Exekowitz & Stein, 2014; Rao, 2009), we found the difficulty-accuracy trade-off was most likely a major cause of us finding a hot hand effect in our unpaid players. This hot hand effect was absent for paid players. Estimated at approximately 5% by the traditional measure, our unpaid players showed a hot hand effect closer in size to that reported in hot hand beliefs (Gilovich, Vallone & Tversky, 1985) than any previous research we are aware of. This finding hints at reconciliation between hot hand beliefs and empirical data that rests on motivation and cognitive control. Specifically, when player motivation is low, a decrease in cognitive control may follow errors, and an increase in cognitive control may follow success. In this way, success breeds success. Critically, the hot hand may well remain a fallacy at the professional level when motivation is high, but fans and players may have experienced the hot hand themselves in amateur contexts where motivation is lower — hence the resilient nature of the belief. Future research might examine whether similar post-error recklessness occurs in the amateur sport context, where motivation may be lower and repeated errors may discourage players, whereas success may provide encouragement to take more care, and hence be more accurate. This would be commensurate with the findings that golfers (Cotton & Price, 2006) and tennis players (Klaassen & Magnus, 2001) with little competitive experience were more likely to demonstrate the hot hand than those with more competitive experience.

5 Conclusion

Our simultaneous investigation of post-error slowing and the hot hand revealed a surprising pattern of results that both supported and challenged existing theories. The take home message from our work is that caution is required when

applying our current understanding of how errors and success influence behaviour in novel contexts. We confirmed that motivation and cognitive control are central considerations when exploring the effect of previous outcomes. This conclusion is commensurate with the speculations of Yeung and Summerfield (2012); there may be substantial task-dependent differences in post-error behaviour for temporally extended and goal driven tasks. We conclude that post-error slowing should not necessarily be considered a general phenomenon in decision-making, but rather one that is pervasive in tasks that require a rapid response without much deliberation. Although replication and extension of our work to amateur sport is required, our work also has the potential to increase our understanding of hot hand beliefs. Although likley a fallacy in professional sports, the hot hand may be observed in contexts that encourage post-error recklessness.

References

Avugos, S., Köppen, J., Czienskowski, U., Raab, M., & Bar-Eli, M. (2013). The "hot hand" reconsidered: A meta-analytic approach. *Psychology of Sport and Exercise, 14*(1), 21–27. http://dx.doi.org/10.1016/j.psychsport.2012.07.005

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise, 7*(6), 525–553. http://dx.doi.org/10.1016/j.psychsport.2006.03.001

Bocskocsky, A., Ezekowitz, J., & Stein, C. (2014). *The hot hand: A new approach to an old "fallacy"*. Paper presented at the MIT Sloan Sports Analytics Conference, Boston.

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed-accuracy trade-off that maximizes reward rate? *Quarterly Journal of Experimental Psychology (Hove), 63*(5), 863–891. http://dx.doi.org/10.1080/17470210903091643

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research, 12*, 19–64.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual Review of Psychology, 66*, 83–113. http://dx.doi.org/10.1146/annurev-psych-010814-015044

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.

Brown, S., Steyvers, M., & Wagenmakers, E.-J. (2009). Observing evidence accumulation during multi-alternative decisions. *Journal of Mathematical Psychology, 53*(6), 453–462. http://dx.doi.org/10.1016/j.jmp.2009.09.002

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst, 24*(1), 1–44.

Camerer, C. F., & Hogarth, R. M. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1), 7–42. http://dx.doi.org/10.1023/a:1007850605129

Cotton, C., & Price, J. (2006). The hot hand, competitive experience, and performance differences by gender. Retrieved from Social Science Research Network website: http://dx.doi.org/10.2139/ssrn.933677

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE, 8*(3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology, 2*(233), 1–9. http://dx.doi.org/10.3389/fpsyg.2011.00233

Dudschig, C., & Jentzsch, I. (2009). Speeding before and slowing after errors: Is it all just strategy? *Brain Research, 1296*, 56–62. http://dx.doi.org/10.1016/j.brainres.2009.08.009

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E. J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging, 28*(1), 64–76. http://dx.doi.org/10.1037/a0029875

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012a). Testing theories of post-error slowing. *Attention, Perception, and Psychophysics, 74*(2), 454–465. http://dx.doi.org/10.3758/s13414-011-0243-2

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012b). How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology, 56*(3), 208–216. http://dx.doi.org/10.1016/j.jmp.2012.04.001

Gehring, W. J., & Fencsik, D. E. (2001). Functions of the medial frontal cortex in the processing of conflict and errors. *The Journal of Neuroscience, 21*(23), 9430–9437.

Gilden, D. L., & Wilson, S. G. (1995). Streaks in skilled performance. *Psychonomic Bulletin & Review, 2*(2), 260–265.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314. http://dx.doi.org/10.1016/0010-0285(85)90010-6

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104. http://dx.doi.org/10.1037/0003-066x.59.2.93

Hajcak, G., & Simons, R. F. (2002). Error-related brain activity in obsessive-compulsive undergraduates. *Psychiatry Research, 110*(1), 63–72.

Hawkins, G., Brown, S., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review, 19*(2), 339–348. http://dx.doi.org/10.3758/s13423-012-0216-z

Houtman, F., Núñez Castellar, E. P., & Notebaert, W. (2012). Orienting to errors with and without immediate feedback. *Journal of Cognitive Psychology, 24*(3), 278–285.

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 51*(4), 261–268. http://dx.doi.org/10.1037/h0041911

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *The Quarterly Journal of Experimental Psychology, 62*(2), 209-218. http://dx.doi.org/10.1080/17470210802240655

Jeffreys, H. (1961). *Theory of Probability.* Oxford: Oxford University Press.

Klassen, F. J. G. M., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association, 96*, 500–509.

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience, 12*(7), 939–945. http://dx.doi.org/10.1038/nn.2321

Laming, D. (1968). *Information theory of choice-reaction times.* New York: Academic Press.

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica, 43*, 199–224.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropman, D., Verhagen, A. J., ... Wagenmakers, E.-J. (2015). JASP (Version 0.7) [Computer Software].

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

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition, 111*(2), 275–279.

Núñez Castellar, E. P., Kühn, S., Fias, W., & Notebaert, W. (2010). Outcome expectancy and not accuracy determines posterror slowing: ERP support. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 270–278.

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of Cognitive Neuroscience, 23*(11), 3419–3432. http://dx.doi.org/10.1162/jocn_a_00011

Rabbitt, P. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology 71*(2), 264–272.

Rabbitt, P. (1966b). Error correction time without external error signals. *Nature, 212*(5060), 438–438.

Rabbitt, P. (1969). Psychological refractory delay and response-stimulus interval duration in serial, choice-response tasks. *Acta Psychologica, 30*(0), 195–219. http://dx.doi.org/10.1016/0001-6918(69)90051-1

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology, 29*(4), 727–743. http://dx.doi.org/10.1080/14640747708400645

Rao, J. M. (2009). *Experts' perceptions of autocorrelation: The hot hand fallacy among professional basketball players.* http://www.justinmrao.com/playersbeliefs.pdf. UC San Diego.

Read, D. (2005). Monetary incentives, what are they good for? *Journal of Economic Methodology, 12*(2), 265–276. http://dx.doi.org/10.1080/13501780500086180

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods, 39*(3), 365–370. http://dx.doi.org/10.3758/BF03193004

Ridderinkhof, K. R., van den Wildenberg, W. P. M., Wijnen, J., & Burle, B. (2004). Response inhibition in conflict tasks is revealed in delta plots. In M. Postner (Ed.), Attention (pp. 369–377). New York: Guillford Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16, 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Morey, R. D., & Rouder, J. N. (2014). BayesFactor 0.9.6. Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

Schroder, H. S., & Moser, J. S. (2014). Improving the study of error monitoring with consideration of behavioral performance measures. *Frontiers in Human Neuroscience, 8*(MAR). http://dx.doi.org/10.3389/fnhum.2014.00178

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1865–1897. http://dx.doi.org/10.1037/a0016926

Smith, G. (2003). Horseshoe pitchers' hot hands. *Psychonomic Bulletin and Review, 10*(3), 753–758.

Sturmer, B., Nigbur, R., Schacht, A., & Sommer, W. (2011). Reward and punishment effects on error processing and conflict control. *Frontiers in Psychology, 2*, 335. http://

dx.doi.org/10.3389/fpsyg.2011.00335

Ullsperger, M., & Szymanowski, E. (2004). ERP Correlates of error relevenace. In M. Ullsperger & M. Falkstein (Eds.), *Errors, Conflicts, and the Brain. Current opinions on Performance Monitoring* (pp. 171–184). Leipzig: MPI for Human Cognitive and Brain Sciences.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin and Review,* 16, 752–760. http://dx.doi.org/10.3758/PBR.16.4.752

Williams, P., Nesbitt, K., Eidels, A., Washburn, M., & Cornforth, D. (2013). Evaluating Player Strategies in the Design of a Hot Hand Game. *GSTF Journal on Computing (JoC), 3*(2), 1–11. http://dx.doi.org/10.7603/s40601-013-0006-0

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321. http://dx.doi.org/10.1098/rstb.2011.0416

Section 2

Chapters 4, 5, and 6

Chapter 4

In Section 1 the development of a novel cognitive paradigm was outlined. This development culminated in a cognitive game that we referred to as the Buckets game. In Chapter 3 I documented the application of the Buckets game to explore the theoretical and empirical links between the hot hand and post-error slowing. Also in Chapter 3, I documented the comparison of three methods for calculating post-error slowing. This allowed us to compare the contributions of errors to that of global effects such as fatigue or boredom. These methods utilised in Chapter 3, for calculating post-error slowing and hot hand effects in empirical data, could be readily implemented to test sequential dependence in other psychological paradigms and inform researchers about the mechanisms of cognitive control.

In Section 2, I take this approach and focus on a well-established paradigm – the emotional Stroop task. We use the emotional Stroop task to explore sequential effects in clinically oriented samples. Chapter 4 provides the launching pad for this exploration and contains two components. The first component is the *Paper 4 Overview*, which is broken into two sub-sections, described further below. This overview might be best read in conjunction with *Paper 4*, which is presented in full to conclude the chapter.

Paper 4 Overview

This *Paper 4 Overview* has 2 sub-sections. I recommend Paper 4 be read before, or in conjunction with, this component. The first sub-section highlights the unique contributions of *Paper 4*. The second sub-section is a *Summary and Transition* section, which as the name suggests, summarises the contributions of

Paper 4 and sets the stage for Chapter 5. Perhaps most importantly for the continuity of thesis, the transition section makes a strong case for using the emotional Stroop task to explore the sequential effects, particularly in populations with depression. This positions Section 2 as a natural and clinically oriented extension to Section 1.

Paper 4 provides a methodological review and best practice guideline for the implementation of the emotional Stroop task. In the emotional Stroop task, the participant names the print-color of singly presented words. The words are typically either emotional (e.g., SAD) or non-emotional (e.g., SAP), and this simple task yields results of theoretical and practical consequence. The typically slower response time for emotional words compared to non-emotional words, known as the emotional Stroop effect (ESE), documents that participants are sensitive to the emotional valence of stimuli even when it is completely irrelevant to the task at hand. McKenna and Sharma (2004) argue that at least part of the effect is a consequence of carry-over effects, from emotionally charged words on preceding trials, and thus lends itself naturally to the analysis techniques used for in Section 1. Paper 4 is accompanied by an online video, which can be accessed here:

http://www.jove.com/video/53720/the-emotional-stroop-task-assessing-cognitive-performance-under

The work is novel in that it systematically reviews the influence of various emotional Stroop task methodologies on the ESE. Such a review is timely because the emotional Stroop task has become an immensely popular method for probing emotion and anxiety with both patient and non-patient populations. Indeed, we follow this precise path in Papers 5 and 6. While the prevalence of the ESE is not disputed in the

literature, the magnitude of the effect varies considerably, making a systematic review and set of standardised recommendations, as provided by Paper 4, timely.

As we identify in Paper 4, substantial variability in the ESE is evident in studies testing well-defined population groups for computer presentations. This effect size variability is problematic because it prevents the possibility of the emotional Stroop task being used to explore individual differences. In other words, despite the overwhelming popularity of the emotional Stroop task, the clinical potential of the paradigm may yet remain untapped because effect size variability means pathology cannot be explored and diagnosed at the level of the individual. Thus, one purpose of our work was to present a clear and standardised set of procedures that would remove confounding and unwanted variability. As an additional bonus, removing this unwanted variability would improve the reliability of comparisons and conclusions made at the group level.

Of course, one specific methodology would not meet the needs of the diverse range of research undertaken via the emotional Stroop task. Even still, a standardised set of methodologies would be valuable, with the final design dependent on the specific area of interest. We take this approach ourselves in Chapter 5, in which we investigate sequential Stroop effects proposed by McKenna and Sharma (2004). Specifically, McKenna and Sharma proposed partitioning of the ESE into two components: (1) the influence of the emotional content on the current trial, or *fast effect*, and (2) the impact of emotional content on the subsequent trial, or *slow effect*. As we note in Paper 4, this partitioning of the ESE requires a mixed presentation design. In a mixed presentation design, the emotional and non-emotional words are intermixed within a block. I return to the advantages of this design when introducing Paper 5.

Summary and Transition

Paper 4 provides a methodological review and best practice guideline for the implementation of the emotional Stroop task. This review is a valuable addition because the magnitude of the ESE varies considerably, making a systematic review and set of standardised recommendations a timely. We take this approach ourselves in Chapter 5, where we explore partitioning of the ESE into fast (current trial) and slow (subsequent trial) components, which is an example of a sequential effect. As we note in Paper 4, exploring sequential effects requires an emotional Stroop task with a mixed presentation design. We use this type of emotional Stroop task in both Chapters 5 and 6, using clinically oriented samples. Specifically, we measured participants for depression symptoms and use variants of a mixed presentation emotional Stroop task to investigate both the fast and slow ESE (Chapter 5), and post-error adjustments (Chapter 6).

This work is novel because even though the emotional Stroop task has been heavily employed to study the effects of emotional stimuli on cognition for anxious and depressed populations (see Williams, Mathews, & MacLeod, 1996, for a review), the emotional Stroop task has not been applied to explore the impact of emotional stimuli on sequential effects. While some work had been undertaken in this area, ultimately this work showed inconsistent and ambiguous effects. I address this previous work more completely in Chapter 6, and so I will not pursue it further here. It suffices to note that little headway has been made in using the emotional Stroop paradigm to study sequential effects.

This lack of research for sequential effects using the emotional Stroop task is surprising given that recent neuropsychological studies suggest that cognitive deficits in depression - particularly those related to executive functions such as planning and

error detection - may be partially due to abnormal responses to negative feedback or oversensitivity to errors. For example, Beats et al. (1996) found a rapid deterioration of performance for elderly depressed patients once a mistake was committed and observed. Elliott et al. (1996, 1997) extended these findings by showing that the abnormal response to negative feedback was specific to depression and correlated with depression severity. False and negative feedback has also been found to disrupt performance for depressed patients in visual discrimination tasks (Murphy et al, 2003). Pizzagalli, Peccoralo, Davidson, and Cohen (2006) noted the importance of these results when documenting the importance of rostral anterior cingulate cortex (ACC) function in depression. The rostral ACC is a region implicated in both error detection and the evaluation of the emotional significance of events. Pizzagalli et al. showed that for participants with low (but not high) levels of depression symptoms, resting activation within this region predicted post-error adjustments.

In summary, there is considerable evidence to suggest (1) errors and negative feedback may influence the cognitive performance of those with depression, and (2) that the region implicated in the evaluation of emotional content is also implicated in the regulation of post-error adjustments. These findings provided substantial motivation to use the sequential methodologies and techniques I explored in Section 1 of this thesis, and apply them to emotional Stroop task, as outlined to begin Section 2. Section 2 of this thesis is therefore a natural extension of the work I outlined in Section 1.

Video Article

The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content

Moshe Shay Ben-Haim[1], Paul Williams[2], Zachary Howard[2], Yaniv Mama[3], Ami Eidels[2], Daniel Algom[1]

[1]School of Psychological Sciences, Tel-Aviv University

[2]School of Psychology, University of Newcastle

[3]Department of Behavioral Sciences, Ariel University

Correspondence to: Moshe Shay Ben-Haim at shay.mbh@gmail.com

URL: http://www.jove.com/video/53720
DOI: doi:10.3791/53720

Keywords: Emotional Stroop, selective attention, sustained effect, habituation, emotion, anxiety

Date Published: 3/29/2016

Citation: Ben-Haim, M.S., Williams, P., Howard, Z., Mama, Y., Eidels, A., Algom, D. The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content. *J. Vis. Exp.* (), e53720, doi:10.3791/53720 (2016).

Abstract

The emotional Stroop effect (ESE) is the result of longer naming latencies to ink colors of emotion words than to ink colors of neutral words. The difference shows that people are affected by the emotional content conveyed by the carrier words even though they are irrelevant to the color-naming task at hand. The ESE has been widely deployed with patient populations, as well as with non-selected populations, because the emotion words can be selected to match the tested pathology. The ESE is a powerful tool, yet it is vulnerable to various threats to its validity. This reportrefers to potential sources of confounding and includes a modal experiment that provides the means to control for them.The most prevalent threat to the validity of existing ESE studies is sustained effects and habituation wrought about by repeated exposure to emotion stimuli. Consequently, the order of exposure to emotion and neutral stimuli is of utmost importance. We show that in the standard design, only one specific order produces the ESE.

Video Link

The video component of this article can be found at http://www.jove.com/video/53720/

Introduction

Modern life is replete with emotion and stress. Who has avoided the emergency room or (witnessing) a traffic accident? In order to perform efficiently under such stressful situations, it is important to preserve one's composure by focusing on the relevant stimuli. However, research has shown that the emotional valence of the stimulus can affect attention, in particular modulate the speed of processing.In the laboratory, one of the most popular paradigms to study the effect of negative stimuli on performance is the emotional Stroop task. The typical finding is that it takes people longer to name the ink color of emotion words than that of neutral words,the Emotional Stroop Effect (ESE). There are several accounts that attempt to explain the observed slowdown attributing attention[3], freezing[2], or mood[3],however it is still a matter of current debate.

The experimental setup of the emotional Stroop task is well known. Words in color are presented singlyfor view and the participant's task is to name the ink color of each word as quickly and accurately as possible. The words come from two categories of different valence. The first category includes negative words (*e.g.*, DEATH) or words related to a specific psychopathology (*e.g.*, GERMS with obsessive-compulsive patients or BATTLE with post-traumatic stress disorder patients). The second category includes neutral words (*e.g.*, CHAIR). The ESE is the difference in color-naming latency between the emotional and the neutral words. The stimuli can be presented in a single block with emotion and neutral words intermixed in a random fashion or in two separate blocks defined by word category. The slowdown with emotion words is usually more pronounced when the ESE is derived in the blocked design[4, 5]. Therefore, the block design has become the method of choice for researches of the ESE and it is the method applied in this protocol too(see **Figure 1** for an illustration of the emotional Stroop experimental setup).

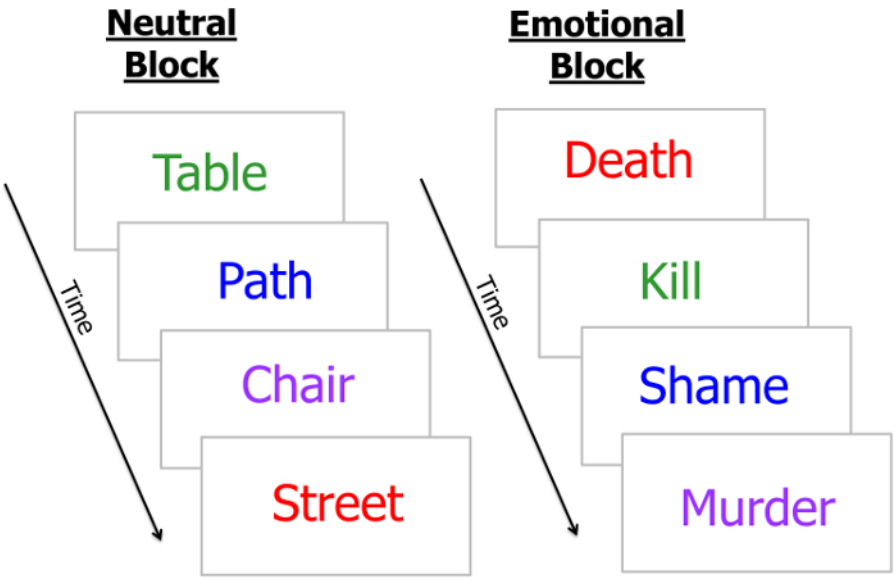


Figure 1: The Experimental Array in the Emotional Stroop Task: The participant's task is to name color in which the word appears. Please click here to view a larger version of this figure.

It is important to distinguish the ESE from its older namesake, the classic Stroop effect (SE)[6]. In the classic Stroop task, *colorwords* such as RED or GREEN are presented in various ink colors and the participant's task is to name the ink color of the words. Despite the shared task - to identify the ink color of words - the SE and the ESE differ. Because the words are color words, the stimuli in the classic Stroop task divide into congruent (the word naming its ink color) and incongruent (word and color conflict) categories.The SE is defined as the difference in color naming performance between congruent and incongruent stimuli. Because the quality of congruity does not apply to stimuli in the ESE - the word CANCER in blue is neither more nor less congruent than the word LECTURE in brown - the SE is not defined in the environment of the ESE. The ESE documents the effect of the emotional valence of the stimuli on performance.

The ESE, just like the SE, has generated voluminous research. In fact, the ESE rivals its namesake in sheer output of experimental studies with both patient and non-patient populations (for a review, see[1, 7]). The task has been employed with a gamut of pathologies from generalized anxiety (*e.g.*[8, 9]) to trait anxiety (*e.g.*[10, 11]) to obsessive-compulsive disorders (*e.g.*,[12, 13]) to depression ([9, 14]) to social phobia ([15, 16]) topost-traumatic stress disorders (*e.g.*[17, 18]). The ESE has also been studied with unselected populations (*e.g*[2, 3, 19, 20]), although the effects in heathy participants are not always observed and are often less pronounced. At least a portion of ESE's popularity is attributable to its objective nature as it is not based on self-report and is not intrusive. Furthermore, the emotion words can be selected to tap the specific pathology or current concern of the patient.

Below, we portray the steps required to design and perform an emotional Stroop experiment. Our purpose in this report is to describe in detail an ESE experiment with needed controls. The most important feature of this design is the control it provides against various threats to validity. The main threat treated in this design is that of habituation. Adopting these procedures renders the ESE a valid and reliable means of assessing attention under emotion.

Protocol

The protocol follows the guidelines of Tel-Aviv University Helsinki human research ethics committee.

1. Word Selection and Matching

1. Create lists of words for each word category of interest. For example, make lists of generally emotional words (*e.g.*,HATE, POX), concern relevance words (*e.g.*, RAPE,VICTIM), and neutral words that are preferably orthographic neighbors[21] of the emotional/concern relevant words(*e.g.*,GATE,BOX). Create lists with as many words as possible as not all of these words will be used; a smaller subset of matched words will later be selected.
    1. When deemed necessary, verify the valence/emotionality/arousal of the words by a questionnaire rating the words on a ratingscale (*e.g.*, from 1 - 7). Attempt to select words that are at the extreme end of the scale. If comparing emotional words with positive words, attempt to include words that are matched on (absolute) emotionality and arousal scores.

2. Write for each word on the lists, its length in characters as well as its estimated frequency in the appropriate language (for English words, use log hyperspace analogue to language (HAL) frequency[22]) in order to match the word lists on lexical factors.
NOTE: This is important as these variables can also affect color naming latency (*e.g.*[21, 23]). Arguably, the most important lexical attributes to control are word *frequency*, *length*, and if possible *orthographic neighborhood* (which naturally intersect with length).

	Emotional			Neutral	
Word	Length	Frequency	Word	Length	Frequency
hate	4	10.7	gate	4	9.7
dead	4	11.2	dear	4	10.29
poor	4	10.9	pool	4	9.7
snake	5	8.6	shake	5	8.6
gloom	5	8.1	bloom	5	8.2
bomb	4	9.64	comb	4	7.39
pox	3	7.1	box	3	12.1
Average	4.1	9.4	Average	4.1	9.4

Table1: Example of a List of Words Matched on Length, Orthographic Neighborhood, and Average Frequency.

3. Select words by adding pairs each time that match in length and frequency, and that are orthographic neighbors of one another (for example replace one of the emotional word letters with another to form a neutral orthographic neighbor, (see **Table 1**).
    1. If a complete match is not possible, balance the bias of frequency in the subsequent pairs of words by adding a pair of words with a small gap in frequency in the opposite direction of the gap formed by the previous pair of words. Ensure that the final lists of matched words contains 20 - 50 words in each word category in order to have a minimum of 20 trials in a block to collect sufficient data and avoid repetition of words and hence habituation[3].
    NOTE: If many words are difficult to obtain, it is possible to introduce some repetition of words. However try to keep repetitions to a minimum as it may dilute the ESE due to habituation[3].

4. Check that the final list is matched as much as possible on all lexical variables (*e.g.* word frequency, word length). Perform a statistical validation (student's t-test) to confirm that the final lists do not differ significantly from each other on frequency.

2. Preparation of Experimental Design

1. Select the design: Blocked (the stimuli are presented in separate blocks of trials defined by word valence) or mixed (the stimuli are presented in one block in which emotion and neutral words are intermixed in the same list).
    1. Choose a blocked design if seeking to test global effects at the level of the word category.
    NOTE: Effect sizes are usually higher in a blocked design than in a mixed presentation[4, 5]. Therefore, blocked design is often the favored method of presentation.
    2. Choose a mixed design if seeking to decompose the ESE at the individual word level and for "fast" and "slow" effects, asa negative item can affect not only its own naming (the ESE) but also that of the immediately following item. Consequently, the former is dubbed 'fast effect,' the latter 'slow effect'[24, 25].

2. In a blocked experimental design, select a fixed or counterbalanced/randomized order of blocks.
    1. For a counterbalanced order present the two possible order of blocks to different groups of participants:
    Group A: *Neutral Block - Emotional Block*
    Group B: *Emotional Block - Neutral Block*
    NOTE: In this balanced design, an ESE is expected to emerge only in the group performing in the first order due to sustained effects in the group performing in the second order (see **Figure 4**)[3]. We therefore suggest using a fixed presentation of blocks.
    2. For a fixed order of blocks present first the neutral block then the emotional block, and if desired present an additional new neutral block flowing the emotional block in order to examine sustained effects: *Neutral Block 1 - Emotional Block -- Neutral Block 2.*
    NOTE: In this setup, two effects can be observed. The first is the canonical ESE, calculated as the difference in performance between the emotional block and the first neutral block. Asecond is asustained effect, obtained by subtracting the mean latency of thesecond neutral block from that of the first neutral block. A positive difference indicates the presence of sustained effects brought about by the emotional block.
        1. In order to rule out confounding through training or fatigue, advisably perform an auxiliary experiment with several blocks ofneutral items only.
        NOTE: With three blocks of 40 neutral words, no residual fatigue is expected (see **Figure 3**)[3]. If there are more than 40 words per block, or more experimental blocks, it may be necessary to control for effects of fatigue by counterbalancing block order (but check for order-of-blocks effects in the statistical analysis).

3. Experimental Programing and Randomization

1. Choose a computer software or programing language to serve as a vehicle to present the stimuli and measure the participant's responses.
    1. Optionally, use the commercially available DirectRT software which is relatively easy to deploy and reliable. See Supplemental Code File for a DirectRT executable excel file as an example for a typical programed ESE experiment.Some additional software packages are SuperLab and E-prime, which are also suitable alternatives for governing the experiment.

2. Select the method of responding: manual or oral. Both types of responding are appropriate.

1. For keyboard activated responding,use a longer training session in order for the participant to learn the mapping of keys onto the ink colors (of about 20 - 40 trials).
2. For vocal responding use ashorter training session (of 4 - 8 trials). Set the reaction time to be measured from the outset of the stimulus to the first phoneme said.
   NOTE: Vocal responses pose difficulties at the stage of deciphering the responses (one must listen to the recordings and classify errors). However, a new algorithm that mechanically classifies the vocal recordings can render actual human classification gratuitous[26]. An advantage of oral responses is that the resulting auditory files can be further analyzed (via other dedicated programs) for vocal parameters associated with emotion[3].

3. Choose the ink-colors to be assigned to the words (*e.g.*, the colors blue, green, red, and purple). Use easily discriminable colors against a well contrasted background (*e.g.*, white or grey). If key-press responding is used, use no more than 4 colors, so that the mapping of colors to the keys is easily mastered.
   NOTE: If vocal responding is used, it is possible to use many more colors. Recall though that with a computer algorithm to classify the data, voice-identification errors increase with the number of colors (= responses).
4. Use an easily legible font, and size for the words.
5. Since the preferred approach is to use no repetition of words, assign to each word a single color randomly by the computer program for every participant.
6. Present each word singly around the middle of the computer screen.Optionally, introduce a small amount of spatial uncertainty, so that each word is presented in a random different position approximately 50 pixels around the center (*e.g.*, see Supplemental Code File). This is done in order to discourage participants focusing on a small section of the word (thus circumventing reading).
7. Prepare a short training block to familiarize the participant with the task and the stimuli. For vocal responding, a few trials using the word 'example' in each of the experimental colors may suffice (with computerized identification of data, a whole session is neededto train the voice-identification algorithm); for key-press responding, as many as 20 - 40 trials may be necessary to master the mapping of the colors.
8. Following training, present the experimental blocks (*e.g.*, three blocks of neutral-emotional and neutral words). Introduce short breaks between successive blocks of trials (of say, 30-60 seconds each). Do not allow the next block to begin before the designated interval for the break has elapsed as participants tend to rush through the experiment in order to finish it quickly.
9. Prepare the task instructions. At the start of each block, present the following instruction 'Respond to the ink color of the word as quickly and accurately as possible.' Avoid mentioning word reading, or that the word should be ignored as this request may artificially augment reading the words (see ironic process theory, *e.g.*[27]).
10. Specify the inter-trial or inter-stimulus interval (ISI) between experimental trials of a stimulus and the successive stimulus. Typically use an ISI of 500 msec (in blocked presentation).
    NOTE: Short ISI may promote shorter responses than longer ISI and carry over-effects with emotion items have been reported for ISI up to 1,000 msec[24].
11. Optional: Add an anxiety questionnaire at the end of the experiment or in a separate session. This provides for a baseline measure of the participant's anxiety, such as the state -trait anxiety inventory (STAI)[28].

4. Subject Selection and Preparation

1. Once programing is completed, recruit participants preferably from the same age-group and background. Participants can enroll for course credit, for pay, or on a voluntary basis.
2. Make sure participants are native speakers of the language used in the study, and do not have any attention deficits or color blindness.
3. Guide the participant to a quiet room in front of a computer. Explain the task instructions and ask the participant to read additional instructions written on the computer screen.

5. Data and Statistical Analysis

1. As a rule, perform reaction time analyses with respect to correct responses only. Also, exclude extreme responses.
   1. Typically, exclude responses faster or smaller than 2.5 SD around the mean. Nevertheless, be careful not to discard more than 5% of the data. Error rates are typically small, but it is still advisable to compare them across conditions and rule out a speed-accuracy tradeoff.

2. With a fixed presentation order, perform planned comparisons (or a student's t-test, if only two blocks are administered) with a designated statistical software such as SPSS or STATISTICA (here instructions are given for STATISTICA). For testing of the ESE, compare the emotion block with the first neutral block. For sustained effects compare the second neutral block (which followed the emotional block) with the first neutral block.
   1. Perform planned comparisons by choosing -Statistics->Advanced Linear/Non-Linear Models->General Linear Models->More results -> Planed comps.

3. When the order of blocks is counterbalanced across participants, perform an ANOVA with Block Valence as a within-subjectfactor and Order of Blocks as a between-subjectfactor.
   1. Perform ANOVA bychoosing -Statistics->Advanced Linear/Non-Linear Models->General Linear Models.

Representative Results

When blocks follow the neutral-emotion-neutral sequence (*e.g.*[3]), a large ESE of 34 msec is observed via slower responses in the emotion block (mean of 791 msec) than in the first neutral block (mean of 757 msec; see **Figure 2**). The same group of participants wasalso fairly sluggish to name the ink color in the second set of (other) neutral words (mean of 778 msec). The 21 msec difference in performance between the two blocks with neutral itemsdocuments the presence of the sustained effect of exposure to negative emotional stimuli.

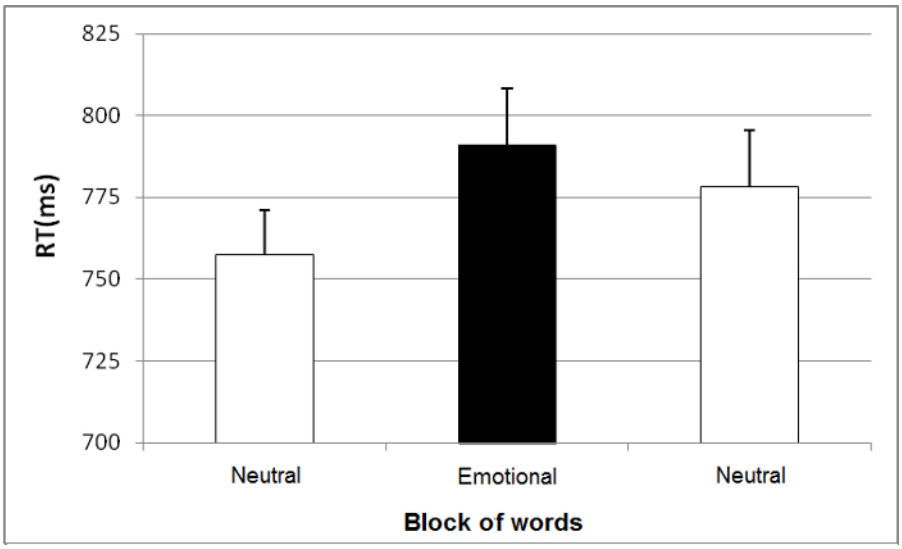


Figure 2: Mean RTs to Name the Ink-color of Singly Presented Words in Three Blocks of Trials with Neutral, Emotion, and More Neutral Items. The blocks with neutral items entail different matched words. Vocal responses were used in this experiment. The error bars depict one standarderror around the mean. Please click here to view a larger version of this figure.

In order to verify that there is no flagging of attention or fatigue (especially with a fixed order of blocks), an auxiliary experiment entailing solely blocks of neutral items can be performed.If there is no difference in performance across successive blocksof trials, one can assume that such effects are minimal (see **Figure 3**).

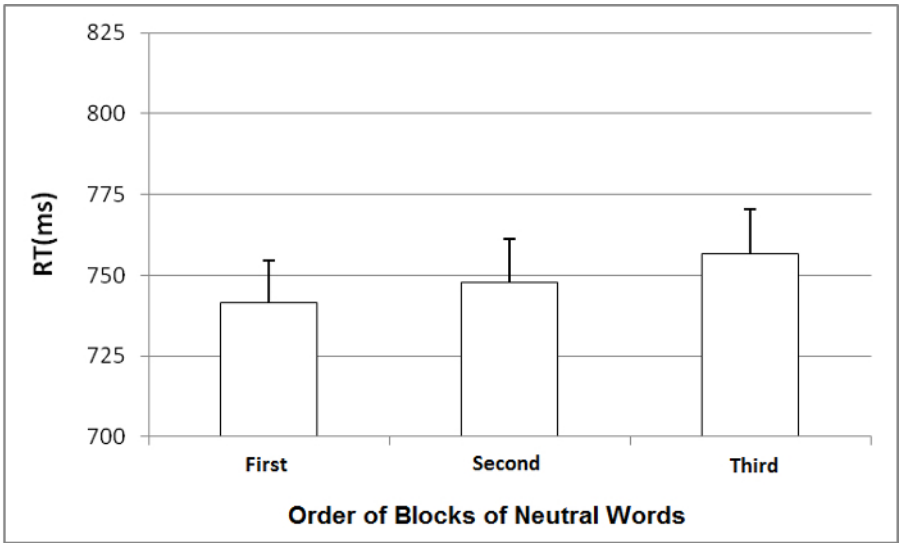


Figure 3: Mean RTs to Name the Ink-color of Singly Presented Words in Three Blocks of Trials with Neutral Items. Each block entails a different set of matched words. Vocal responses were used in this experiment. The error bars depict one standard error around the mean. Please click here to view a larger version of this figure.

In the typical ESE study in the literature with only two blocks of trials (emotion, neutral), an ESE is only expected to emerge in the group of participants performing first in the neutral block.There is not an ESE in the reversed order of blocks. This should result in an interaction of Block Valence and Block Order in the pertinent ANOVA. Of course, this standard design is not suited to test sustained effects or fatigue and habituation. Therefore, we suggest using designs with a minimum of three blocks, neutral-emotion-neutral [**Figure 4**].

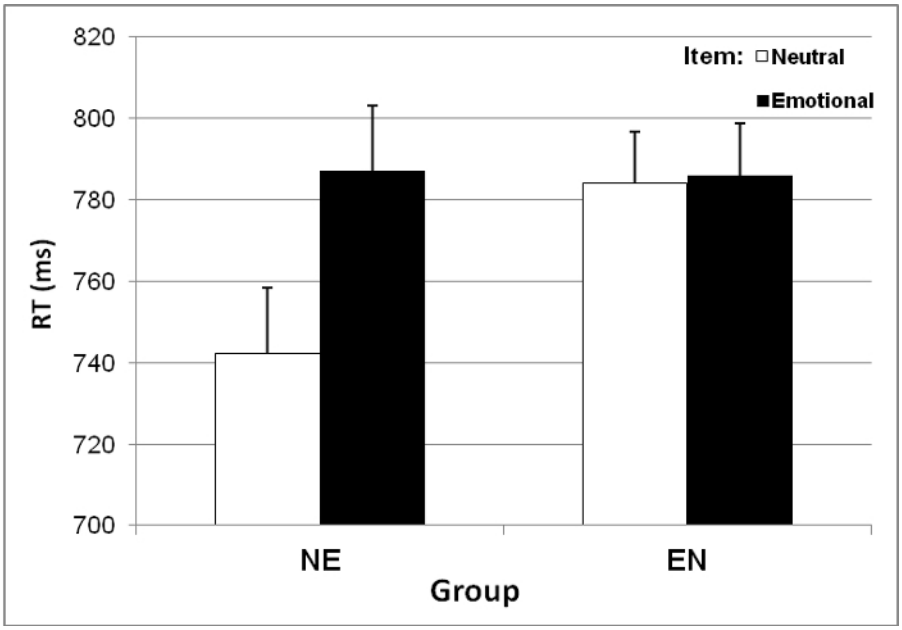


Figure 4: Mean RTs to Name the Ink Color of Emotion and Neutral Words Presented in Blocks of Different Order. In the Neutral-then-Emotional (NE) group, the block of neutral words preceded the block of emotion words. In the Emotional-then-Neutral (EN) group, the block of emotion words preceded that of neutral words.Vocal responses were used in this experiment.The error bars depict one standarderror around the mean.This figure has been modified from[3]. Please click here to view a larger version of this figure.

Supplemental Code File. Please click here to download this file.

Discussion

The ESE comprises a very simple task: The participant names the ink color of singly presented words. This simple task yields results of both pragmatic and theoretical consequence. The ESE documents the fact that people are sensitive to the emotional valence entailed in stimuli although this feature is completely irrelevant to the task at hand.

The ESE hasevolved into an immensely popular method for probing emotion and anxiety with both patient and non-patient populations[1, 74]. Its appeal can be attributed to its potential as an objective (computer-based)diagnostic tool, free of potential patient-therapist interaction bias. The emotion words can be selected to match the specific pathology or current concern of the patient. Furthermore, the tool is not intrusive nor self-report based. The efficiency of the ESE is firmly established at the group level, but it has yet to be demonstrated at the individual level[29, 30]. Further studies are needed in order to assess the reliabilityof the individual patient's ESE and its relation with other known measures of anxiety and other computer based paradigms such as the dot probe[31]. Also, despite the prevalence of the ESE, the precise magnitude of the effect is moot, with reported effect sizes ranging from -1 to 400 msec[1]. This is partly due to the use of different settings (*e.g.*, computer/ cards/ oral) or to the specific pathology group tested. However, substantial variability is still evident in studies testing well-defined population groups in similar test settings. One purpose of the current protocol is to present a clear and standardized procedure byway of removing confounding and unwanted variability.This can be achieved by employing criticalstepssuch as avoiding word repetitions, bypassinghabituation, and allowing for proper lexical control. Following these guidelinesshould helpresearchers collect valid data,draw unbiased conclusions,improve reliability, and aid with comparisonsacross various emotional Stroop studies.

The importance of the critical steps granted, there are many possibilities for variation. Given the possibility of different research questions,the current protocol may not be optimal for some and departures from the current protocol are possible. For example, if a researcher wishes to examine the effect of vocal emotional interference, modifications of the protocol may be necessary Researchers should decide their preferred method of administration to fit their experimental needs. Variations also apply to the number of blocks and wordcategoriesto individual word selection (*e.g.*, controlling for additional lexical variables such as number of syllables) to determining the number of trials (words) within blocks toemploying a mixed or blocked design to introducing fixed/randomized/counterbalanced order of blocks to choosing a vocal or keypress responding to choosing the colors or specifying the inter-trial and block intervals. The advantages and disadvantages of each of these considerations are addressed in the relevant protocol steps.Most can be fitted to one's needs and individual preference.

Disclosures

The authors have nothing to disclose.

Acknowledgements

The authors have no acknowledgements.

References

1. Williams, J.M.G., Mathews, A., & MacLeod, C. The emotional stroop task and psychopathology. *Psychol.Bull.* **120** (1), 3-24, (1996).
2. Algom, D., Chajut, E., & Lev, S. A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *J.. Exp.. Psychol..Gen.* **133** (3), 323-338, (2004).
3. Ben-Haim, M.S., Mama, Y., Icht, M., & Algom, D. Is the emotional Stroop task a special case of mood induction? Evidence from sustained effects of attention under emotion. *Atten..Percept..Psycho.* **76** (1), 81-97, (2014).
4. Holle, C., Neely, J.H., & Heimberg, R.G. The effects of blocked versus random presentation and semantic relatedness of stimulus words on response to a modified Stroop task among social phobics. *Cognitive..Ther.. Res..* **21** (6), 681-697, (1997).
5. Richards, A., French, C.C., Johnson, W., Naparstek, J., & Williams, J. Effects of Mood Manipulation and Anxiety on Performance of an Emotional Stroop Task. *Br.J.Psychol.* **83**, 479-491 (1992).
6. Stroop, J.R. Studies of Interference in Serial Verbal Reactions (Reprinted from Journal Experimental-Psychology, Vol 18, Pg 643-662, 1935). *J.. Exp..Psychol..Gen.* **121** (1), 15-23 (1992).
7. Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J., & van IJzendoorn, M.H. Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychol.Bull.* **133** (1), 1-24, (2007).
8. Mathews, A., Mogg, K., Kentish, J., & Eysenck, M. Effect of Psychological Treatment on Cognitive Bias in Generalized Anxiety Disorder. *Behav.Res.Ther.* **33** (3), 293-303, (1995).
9. Mogg, K., & Bradley, B.P. Attentional bias in generalized anxiety disorder versus depressive disorder. *Cognitive..Ther.. Res..* **29** (1), 29-45, (2005).
10. Mogg, K., Kentish, J., & Bradley, B.P. Effects of Anxiety and Awareness on Color-Identification Latencies for Emotional Words. *Behav.Res.Ther.* **31** (6), 559-567, (1993).
11. Rutherford, E.M., MacLeod, C., & Campbell, L.W. Negative selectivity effects and emotional selectivity effects in anxiety: Differential attentional correlates of state and trait variables. *Cognition..Emotion.* **18** (5), 711-720, (2004).
12. Foa, E.B., Ilai, D., Mccarthy, P.R., Shoyer, B., & Murdock, T. Information-Processing in Obsessive-Compulsive Disorder. *Cognitive..Ther.. Res.* **17** (2), 173-189, (1993).
13. Mcnally, R.J., Riemann, B.C., Louro, C.E., Lukach, B.M., & Kim, E. Cognitive Processing of Emotional Information in Panic Disorder. *Behav.Res.Ther.* **30** (2), 143-149, (1992).
14. Mitterschiffthaler, M.T.*, et al.* Neural basis of the emotional Stroop interference effect in major depression. *Psychol.Med.* **38** (2), 247-256, (2008).
15. Amir, N., Freshman, M., & Foa, E. Enhanced Stroop interference for threat in social phobia. *J.Anxiety Disord.* **16** (1), 1-9, (2002).
16. Andersson, G., Westoo, J., Johansson, L., & Carlbring, P. Cognitive bias via the Internet: A comparison of web-based and standard emotional Stroop tasks in social phobia. *Cognitive Behaviour Therapy.* **35** (1), 55-62, (2006).
17. Paunovic, N., Lundh, L.G., & Ost, L.G. Attentional and memory bias for emotional information in crime victims with acute posttraumatic stress disorder (PTSD). *J.Anxiety Disord.* **16** (6), 675-692, (2002).
18. Constans, J.I., McCloskey, M.S., Vasterling, J.J., Brailey, K., & Mathews, A. Suppression of attentional bias in PTSD. *J.Abnorm.Psychol.* **113** (2), 315-323, (2004).
19. Mama, Y., Ben-Haim, M.S., & Algom, D. When emotion does and does not impair performance: A Garner theory of the emotional Stroop effect. *Cognition..Emotion.* **27** (4), 589-602, (2013).
20. Mckenna, F.P., & Sharma, D. Intrusive Cognitions - an Investigation of the Emotional Stroop Task. *J..Exp..Psychol..Learn.* **21** (6), 1595-1607, (1995).
21. Grainger, J., Oregan, J.K., Jacobs, A.M., & Segui, J. On the Role of Competing Word Units in Visual Word Recognition - the Neighborhood Frequency Effect. *Percept.Psychophys.* **45** (3), 189-195, (1989).
22. Balota, D.A.*, et al.* The English Lexicon Project. *Behav..Res..Method.s..* **39** (3), 445-459, (2007).
23. Larsen, R.J., Mercer, K.A., & Balota, D.A. Lexical characteristics of words used in emotional Stroop experiments. *Emotion.* **6** (1), 62-72, (2006).
24. McKenna, F.P., & Sharma, D. Reversing the emotional stroop effect reveals that it is not what it seems: The role of fast and slow components. *J..Exp..Psychol..Learn..* **30** (2), 382-392, (2004).
25. Frings, C., Englert, J., Wentura, D., & Bermeitinger, C. Decomposing the emotional Stroop effect. *Q.J.Exp.Psychol.* **63** (1), 42-49, (2010).
26. Eidels, A., Ryan, K., Williams, P., & Algom, D. Depth of Processing in the Stroop Task Evidence From a Novel Forced-Reading Condition. *Exp..Psychol.* **61** (5), 385-393, (2014).
27. Wegner, D.M. Ironic Processes of Mental Control. *Psychol.Rev.* **101** (1), 34-52, (1994).
28. Kendall, P.C., Finch, A.J., Auerbach, S.M., Hooke, J.F., & Mikulka, P.J. State-Trait Anxiety Inventory - Systematic Evaluation. *J.Consult.Clin.Psychol.* **44** (3), 406-412, (1976).
29. Dresler, T.*, et al.* Reliability of the emotional Stroop task: An investigation of patients with panic disorder. *J.Psychiatr.Res.* **46** (9), 1243-1248, (2012).
30. Eide, P., Kemp, A., Silberstein, R.B., Nathan, P.J., & Stough, C. Test-retest reliability of the emotional stroop task: Examining the paradox of measurement change. *J.Psychol.* **136** (5), 514-520 (2002).
31. Macleod, C., Mathews, A., & Tata, P. Attentional Bias in Emotional Disorders. *J.Abnorm.Psychol.* **95** (1), 15-20, (1986).

Chapter 5

In Chapter 4 I detailed a systematic review of the emotional Stroop task and outlined a core set of methodologies that could be adjusted to suit specific research objectives. In Chapter 5 I build on this understanding of the emotional Stroop task to develop a modified version of the task. In Paper 5 we outline this task and present it to participants psychometrically measured for depression symptoms using the Beck Depression Inventory (BDI-II; Beck, Steer & Brown, 1996). Of most relevance to the current thesis, we explore the slow emotional Stroop effect (ESE). The slow ESE is a sequential effect, as it estimates the influence of the emotional valence of stimuli on the response speed of the subsequent trial. Chapter 5 focuses on the evaluation of this slow ESE, and has two components. The first component is the *Paper 5 Overview*. I recommend Paper 5 be read before, or in conjunction with this section. *Paper 5* is presented in full to conclude the chapter.

Paper 5 Overview

In Paper 5 we tested whether the processing of emotional stimuli was obligatory, non-obligatory, or task dependent by applying a novel emotional Stroop task; the *forced-processing* emotional Stroop task. In this novel forced-processing task, participants identified the colour and the emotional valence of words (i.e., emotional or non-emotional). The unique design and instructions of this task (explained in detail in Paper 5) forced participants to read, and thus engage, with the emotional content of every item. The mean response speed for each participant from this task are then compared to their performance in a control emotional Stroop task, in which participants identified the colour of words and performed font discrimination

(i.e., italic or not italic). By comparing results across these tasks within subjects, we provided a powerful means discriminating between three alternative views of emotional processing in the emotional Stoop task: obligatory, non-obligatory, and task dependent. Critically for the current body of work, Experiment 2 of Paper 5 allowed an evaluation of sequential effects in the emotional Stroop task. McKenna and Sharma (2004) refer to these effects as fast and slow ESE's.

In framing the argument for fast and slow ESE's, McKenna and Sharma (2004) challenged current theoretical interpretations of performance on the emotional Stroop task. Critical to their argument was the fact *blocked designs*, in which each block either contains all emotional or all non-emotional trials, revealed a larger ESE than *mixed designs* in which emotional and neutral items were mixed within a block (Holle, Neely, & Heimberg, 1997; Phaf, & Kan. 2007). This methodology-dependent difference in the size of the ESE was highlighted in Chapter 4. McKenna and Sharma proposed the difference in ESE for blocked- and mixed-presentation designs may be accounted for by emotionally charged words slowing responses not only on the current trial, but also on subsequent trials. The interference to colour processing from the emotional content of the presented word could then be considered as a *fast effect* (slowdown on the current trial) or *slow effect* (slow down on subsequent trials). For example, in the sequence of stimuli 'SACK'… 'SAD'… 'PAD', a fast effect would describe slower responding on the word 'SAD', whereas a slow effect would describe slower responding on the word 'PAD' (due to the carry-over effect of the preceding, emotionally-charged word 'SAD'). They argued that the literature had previously presented data that was an amalgamation of these fast and slow processes, and this may have limited our understanding of the impact of emotional stimuli.

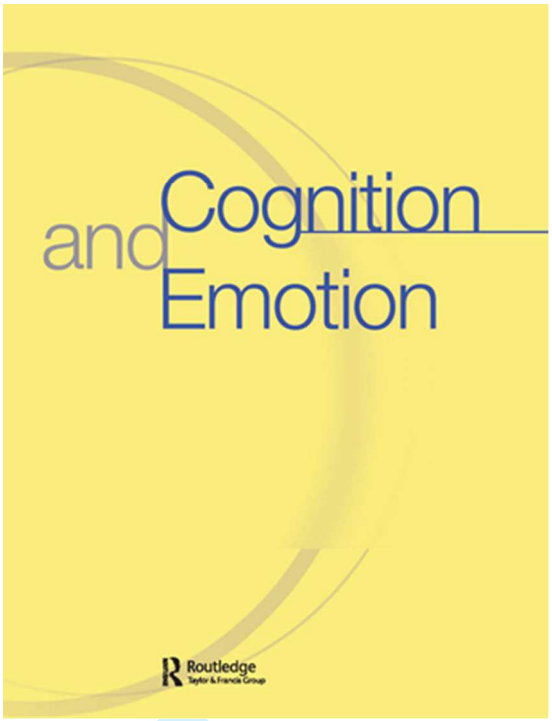
McKenna and Sharma (2014) explored fast and slow effects using a mixed design with pseudorandom trial sequences. The method they used to calculate fast and slow ESE's was designed to eliminate possible confounding influences on results. This is perhaps most easily explained by considering fast effects. Although fast effects were akin to the traditional ESE (i.e., they measure the influence of emotional stimuli on processing for the current trial), McKenna and Sharma calculated the fast ESE by comparing response times for emotional and non-emotional items, *but only for those trials following non-emotional items.* This means that the fast ESE calculation was free of any contaminating slow ESE effect. For the slow ESE, they examined the response time *of only neutral items* that were preceded by emotional, or non-emotional words. Given the conditioning used in these calculations, the measurement of fast and slow effects holds many similarities to the robust measurement method for post-error effects, which I outlined in Chapter 3 of this thesis. Given my previous experience, this placed me in a unique position to perform and interpret the analyses related to these sequential effects.

In their original study, McKenna and Sharma (2014) successfully demonstrated separate fast and slow ESE's, however, their choice of pseudorandom sequencing had the inherent limitation that participants could conceivably predict sequences and anticipate emotional items, causing differences in responses and subsequent biased trends

Summary and Transition

In Paper 5, we outline a fully randomised methodology for presenting stimuli and measuring fast and slow ESE's. This methodology is novel in that it does not rely on pseudorandom sequences. Secondly, while there have been numerous replications of the findings of McKenna and Sharma (2014) regarding slow effects in recent literature (Cane, Sharma, & Albery, 2009; Frings, Englert, Wentura, & Bermeitinger, 2010; Phaf, & Kan, 2007; Waters, Sayette, Franken, & Schwartz, 2005; Wyble, Sharma, & Bowman, 2005), we found no reliable slow ESE in our study. It is possible that our lack of a slow ESE is linked to the unique nature of our emotional Stroop tasks, with both the forced and control tasks requiring dual decisions (i.e., print colour and one other decision).

In summary, with regards to sequential effects, we documented two important and novel results in Paper 5. Firstly, we documented a reliable, fully randomised experimental and statistical methodology for partitioning fast and slow ESE's. This methodology can be used in the future to explore differences in clinical and non-clinical populations. Secondly, we showed that the slow ESE might not be as robust as had been assumed previously, as it did not generalize to a non-standard emotional Stroop design.



An Investigation into How Emotional Words Affect Processing in the Emotional Stroop Task

Journal:	*Cognition and Emotion*
Manuscript ID	CEM-FA 326.16
Manuscript Type:	Full Article
Date Submitted by the Author:	17-Jul-2016
Complete List of Authors:	Ross, Rachel; University of Newcastle, Williams, Paul; University of Newcastle Eidels, Ami; University of Newcastle
Keywords:	Attention, emotional Stroop task, emotional Stroop effect, disengagement, forced-processing

An Investigation into How Emotional Words Affect Processing in the Emotional Stroop Task

Running head: How Do Emotional Words Affect Processing

Rachel Ross, Author, Rachel.Ross20@gmail.com, University of Newcastle, Australia

Paul Williams, Author, Paul.Williams@newcastle.edu.au, University of Newcastle, Australia

Ami, Eidles, Author, Ami.Eidels@newcastle.edu.au, University of Newcastle, Australia

Corresponding author:

Ami Eidels

School of Psychology

University of Newcastle

Ami.Eidels@newcastle.edu.au

+61 – 2 – 492 17089

HOW DO EMOTIONAL WORDS AFFECT PROCESSING

Abstract

The emotional Stroop effect (ESE) is calculated as the difference in reaction time between classifying the print color of emotional (e.g., SAD) and non-emotional (PAD) words. Since participants focus on color and ignore the emotional content, the existence of ESE demonstrates an automatic attentional bias towards emotional stimuli. We tested whether processing emotional stimuli is obligatory, non-obligatory or task dependent. Fifty-five participants across two experiments completed a control task and a novel forced-processing task. In the forced-processing task participants were asked to classify the color of words and their emotional valence, the latter forcing participants to process word meaning. Results demonstrated an inverse ESE in the forced task but not in the control. We concluded that emotional processing does not occur on all trials, supporting a non-obligatory view of processing, and that the ESE is driven by stimuli disengagement.

Keywords: *Attention, emotional Stroop task, emotional Stroop effect, disengagement, forced-processing*

HOW DO EMOTIONAL WORDS AFFECT PROCESSING

Human interactions rely heavily on understanding our own emotions and the emotions of others. Processing emotions accurately is vital, such as responding to a friendly or an aggressive facial expression (Richards, French, Clader, Webb, & Rox, 2002; Fox, Russo, & Georgiou, 2005). Quick and accurate processing of emotional stimuli is so important that additional attention may be allocated to that stimulus, creating an attentional bias. In the emotional Stroop task, participants respond to the color of emotional or non-emotional words whilst ignoring the words' meaning. Responses to the print color of emotional words are typically slower than responses to the color of non-emotional words (Williams, et al., 1996). The difference between emotional and non-emotional items is termed the emotional Stroop effect (ESE), offering a measure of attentional bias towards emotional words. Despite decades of research using ESE data the underlying cognitive processes remain poorly understood.

This paper will evaluate the Emotional Stroop task as a measure of attention bias and attempt to explain the cognitive mechanisms that underlie emotional processing. We will offer a novel method to aid a theoretical resolution regarding the claimed automaticity of emotional processing and the process by which attention bias occurs.

Attentional Bias Theory and the Emotional Stroop Task

Attentional bias is a heightened sensitivity to, and/or preoccupation with, threatening stimuli in the environment (Wyble, Sharma, & Bowman, 2005). Attention bias is claimed to occur across three stages; attentional shift, engagement and disengagement. The Emotional Stroop Task (EST) is, by far, the most frequently applied experimental paradigm for measuring attentional bias towards emotional stimuli (Williams, Mathews & MacLeod, 1996). Participants are asked to classify the print color of emotional and neutral words, such as BAD and BED printed in green or red, whilst refraining from reading the actual word. The difference in reaction time between emotional words and neutral words offers an emotional

Stroop effect (ESE). A positive ESE marks slower responses on emotional trials, and the slowed effect is commonly referred to as the *interference* of the emotional word on processing, describing an attentional bias towards emotional stimuli (Algom, Chajut, & Lev, 2004; Williams el al., 1996).

The EST has successfully established that both clinical and non-clinical populations respond to the color of emotional words slower than neutral words (Williams el al., 1996). It has been proposed that interference occurs at the disengagement stage of attentional processing. Specifically, Verges and Estes (2008) suggest difficulty disengaging from emotional items plays a key role in the ESE. Several attempts have been made to understand this attention bias for emotional stimuli. Experiments typically demonstrate that the presence of an emotional or threatening stimulus, which is unrelated to performance (i.e., an emotional distractor), leads to a delay in response time.

Underlying Mechanisms of Emotional Processing

The cognitive processes that underlie the ESE are still unclear despite decades of application of the EST. We explore two key theoretical propositions. Firstly, we explore whether or not attention bias towards emotional stimuli in the EST is an automatic or obligatory process. Because of the pervasive nature of the ESE, this long-standing assumption has, up until now, been accepted with little consideration. Secondly, we explore the proposition that changes in emotional processing is a task dependent phenomenon; the response-relevance hypothesis (Verges & Estes, 2008).

Automaticity

Automaticity is a complex concept and requires the processes under scrutiny to meet several criteria including non-consciousness, unintentional, involuntary, obligatory and effortless (Wyble et al., 2005). Since there is no consensus concerning the definition of an automatic process (see Tzelgov, 2002), we focus on one quintessential aspect of emotional processes in the EST, namely the nature of processing emotional content. In the EST, processing the emotional content of items is irrelevant for task performance but occurs nonetheless. Processing the emotional content interferes with colour naming, despite the fact that the task requires to focus on the latter and ignore the former. Thus, emotional processing in the emotional Stroop task has been explained in terms of the *obligatory nature of emotional processing* (sometimes referred to interchangeably as the automaticity account, Phaf, & Kan, 2007).

Automatic emotional processing is intuitive from an evolutionary perspective as directing resources towards processing and responding to salient emotional information is adaptive for safety and survival (Carretié, 2014). In the EST, traditional views of obligatory processing propose an all-or-nothing view whereby all attention directed towards an item is the result of competing resources from a limited capacity reservoir (Williams el al., 1996). Contrary to this all-or-nothing view, others have proposed automaticity as a gradient that develops with learning (MacLeod, 1991). Here, each dimension varies in automation with more automatic processes causing interference onto less automatic processes. According to this account, in the EST, emotional processing is more automatic and thus interferes with the less automatic process of color classification.

Key evidence for the automaticity view of emotional processing involves experiments with subliminal presentations. This method presents stimuli for a time duration too short for conscious perception so that conscious processing cannot influence or mask the automaticity of word processing (Fox, 1993; Macleod & Hagan, 1992; Mogg, Bradley, Williams el al.,

1993). Phaf and Kan (2007) conducted a meta-analysis on the effect sizes of 70 studies investigating the automaticity of the emotional Stroop effect. Examining studies using subliminal (referred to as suboptimal) presentations, the review found all studies had effect sizes close to zero, with none approaching significance. Phaf and Kan concluded that "no study contained convincing evidence for automaticity in the emotional Stroop at suboptimal presentation" (p. 190). Phaf and Kan suggested that the ESE is the result of a slow disengagement of emotional words and concluded with two observations, (i) that the automatic view of processing emotional words still lacks clarity, and (ii) that a more diluted version of the automaticity view is slowly being accommodated in the literature.

Alternative Models to Automaticity

Aligning with Phaf and Kan's (2007) latter conclusion, Algom, Chajut, and Lev (2004) proposed a generic slowdown hypothesis that challenged the very nature of the EST. Algom et al. (2004) proposed that the ESE does not demonstrate an automatic attention bias nor it is a true Stroop effect; rather it demonstrates a generic slowdown caused by processing emotional stimuli. The authors proposed that a defining feature of all Stroop tasks is the existence of a logical relationship between the target dimension and the task irrelevant dimension. For example, Eidels, Townsend and Algom (2010) asked participants in a standard (i.e., not emotional) Stroop task to classify the print colors (red, green) of color words (RED, GREEN). Clearly, the combinations of color and word form logically congruent or incongruent combinations. In the EST, emotional and non-emotional word lack the logical semantic congruence or incongruence of the original Stroop stimuli (e.g., word RED in red print-color vs GREEN in red), and therefore need not be considered a Stroop effect.

Algom et al. (2004) instead provided a novel interpretation of the ESE, proposing that an automatic system captures threatening stimuli and prioritizes it above ongoing activity (color naming) in a freezing response. Therefore, the interference of emotional words was

better explained by a reaction to threatening stimuli, rather than a Stroop effect. Algom et al. provided support for the generic slowdown hypothesis via a series of experiments entailing reading and lexical decision tasks. From the results of these tasks they concluded the ESE measures an early pre-attentive (fast and automatic process) freezing of activity, and that this generic cognitive slowing did not constitute an automatic attentional bias for threatening information. According to this cognitive-slowing account, the threat word is processed by a threat detection system that interferes with *all* ongoing cognitive processes when a threat is detected.

Conversely, several studies demonstrated that the ESE is evident with non-threatening words such as positive valence or self-relevance items (Martin, Williams, & Clark, 1991; Mogg, & Marden, 1990; Segal et al., 1995; Williams el al., 1996). This is a serious challenge to Algom, et al.'s (2004) theory, as a slowdown could occur in the absence of biologically relevant threats.

Response-Relevance Hypothesis

Estes and Verges (2008) challenged Algom et al. (2004) generic slowdown theory by proposing that the slowed response to emotional stimuli was not generic but task dependent. Estes and Verges suggested that negative stimuli only elicited a slowed response when the emotionality of the stimuli was *irrelevant* for response. These researchers proposed that interference occurs due to a delay in disengagement with emotional stimuli rather than a generic slowdown. Theoretically, they suggested that adaptive behaviour often requires a rapid response to threatening stimuli such as fleeing or fighting rather than freezing. This task-dependent response-relevance hypothesis posits that a slowdown effect will occur only if disengagement from the stimuli is necessary for task performance. When the stimulus valence is response-relevant, disengagement is not necessary and negative stimuli will speed

the response time. Estes and Verges' hypothesis aligns with Phaf and Kan's (2007) earlier suggestion of a disengagement-related explanation for the ESE.

Estes and Verges (2008) attempted to discriminate between Algom et al. (2004) motor suppression theory and their own response-relevance hypothesis. Their experiment compared participants' performance in a task requiring disengagement (lexical decision task) and a task not requiring disengagement (valence judgment task). The findings supported their main hypothesis. Negative words elicited greater interference when disengagement was necessary for task completion compared to the valence task that required no disengagement. They concluded that the delay was not best explained a generalized slowdown, but rather by the response-relevance hypothesis. Additionally, Estes and Verges' theory does not imply a threat-related cause of the delay, thus evading the theoretical flaw of the generalized motor theory (Algom, Chajut, & Lev, 2004). The proposition that disengagement plays a role in the ESE has prompted more flexible explorations of underlying mechanisms of the ESE.

Assumptions of the Emotional Stroop Task

Alternative models of attentional bias have prompted further enquiry into ESE analysis and interpretation. The ESE is typically calculated by subtracting the mean response time of the emotional condition from the mean response time of the neutral condition. An implicit assumption of this method of calculation is that participants process all words - despite their detrimental effect on performance - and therefore processing of each and every word is obligatory. However, this implicit assumption is open to scrutiny given a difference between two collections of means does not have the explanatory power to determine whether emotional processing occurs on each trial. Processing may not be obligatory on each and every trial, yet ESE can still be observed.

Individuals could process emotional content on some trials and successfully ignore the word's content on other trials. EST data may in fact be a combination of these two processes that have not been partitioned out (Eidels, Ryan, Williams & Algom, 2014), or perhaps a combination of shallower and deeper level of processing (Craik & Lockhart, 1972). Thus, participants in the emotional Stroop task could either (i) process the emotional content of all items to the same extent, the obligatory view; or (ii) process some items deeply and other items in more shallow way (or not process their emotional content at all), the non-obligatory view. In sum, there is emerging evidence that challenges both the all-or-nothing and the gradient view of automaticity in the EST. There is also a lack of available methods to directly examine the existence and extent of automaticity. A novel task is outlined below that can discriminate between these options.

The Forced-Processing Task

The forced-reading Stroop task was developed to explore whether reading in the classic (non-emotional) Stroop task is obligatory (Eidels, Ryan, Williams, & Algom, 2014). In this task, participants were presented with color words (RED, GREEN) and their orthographic neighbors (e.g., ROD, GREED) and had to respond to the print color but only if the word was a color word. Contrary to the classic Stroop tasks, participants were forced to read and engage with every item. This forced-reading task was then compared to a control or classic version. Eidels et al. proposed that if reading occurs on every trial, an obligatory view, the magnitude of the observed Stroop effect should be the same in the forced and control tasks. However, if reading does not occur in every trial, the non-obligatory view, the forced-reading task should reveal a larger Stroop effect. An enhanced Stroop effect was recorded in the forced-reading task compared to the classic task, supporting a non-obligatory view of reading in the Stroop.

The current study adapts the forced-processing task from the classic to the emotional Stroop milieu. In the novel forced-processing task, participants must respond to both the ink color and the emotional content of words (emotional or non-emotional). Participants are forced to

read and engage with the emotional content of every item. This task is then compared with a control emotional Stroop task. The control task also involves two decisions, ink color and whether or not any of the letters are in italics, yet does not require judgment about emotional content.

The forced-processing design also allows for an exploration of Verge and Estes (2008) response-relevance hypothesis. The control emotional Stroop task we develop here demonstrates a lexical-decision variant whereby emotional processing is not required for task performance. Performance in this task requires *disengagement* with the emotional content. Alternatively, the forced processing task involves an emotional-valence decision-making task whereby emotional processing is required for performance; disengagement with emotional content is *not* required for task performance. Based on the response-relevance hypothesis, we would expect emotional items to be slower compared to non-emotional items in the control task (disengagement necessary) and emotional items to be faster than non-emotional items in the forced-processing task (disengagement not necessary). The comparison of performance on these two tasks offers a definitive test of Verges and Estes (2008) response-relevance hypothesis.

In sum, the goal of the current study is to discriminate between the obligatory view, non-obligatory view and response-relevance hypothesis of emotional processing in the emotional Stroop task. As shown in Figure 1, each theory of interest predicts a different pattern of results. The obligatory view predicts and additive (or nil) observed difference between the forced-reading task and the control task (left panel). Alternatively, the forced task might demonstrate a different ESE compared to the control task, suggesting not all items are processed in the control task, and thus supporting the non-obligatory view of emotional processing (middle panel). Lastly, a delay in disengagement may be driving the ESE. Disengagement is necessary in the control task, as emotional processing is not required for task performance. On the other hand, in the forced task, disengagement is not necessary. If the delay-in-disengagement theory holds, we might expect an

ESE in the control task and a reversed (i.e., negative) ESE in the forced task, as emotional processing may facilitate processing latencies (right panel). Experiment 1 tests which of the predicted patterns of interaction depicted in Figure 1 is supported by data. Experiment 2 then replicates Experiment 1 with minor methodological changes, which allow exploration of slow vs fast effects (Sharma & McKenna, 2004; discussed below) and offers some clarification regarding the role of implicit and explicit processing.

Experiment 1

Method

Participants and Design

37 undergraduate psychology students were recruited through an online experimental management system (8 males, 29 females: *Mage* = 23.3 years, *SD* = 6.1, age range: 18 - 48). Participants volunteered to take part and were compensated with course credits. All had normal or corrected to normal vision and English as their first language. The experiment was a within-subject 2 (task: control, forced) x 3 (word condition: positive, negative and non-emotional) design with reaction time and accuracy as dependent variables. All participants completed both tasks on separate days and all auxiliary questionaries.

Apparatus and Materials

The stimulus set comprised 24 words, six positive emotional words (e.g., BETTER, GLAD), six negative emotional words (BITTER, SAD) and 12 uncategorised non-emotional words (BUTTER, PAD), see Table 1 for the complete list. Non-emotional and emotional words were matched on frequency, length, and were orthographic neighbours where possible (BETTER, BUTTER). This ensured participants could not rely on local cues to respond, forcing them to read the entire word. Items were piloted at an earlier stage to ensure accurate categorisation into emotionally positive, emotionally negative and non-emotional conditions.

Post-experiment, participants completed a word identification questionnaire to further ensure word conditions were correctly categorised, revealing 97.75% correctly identified, with no participants excluded for poor accuracy.

(Table 1 about here)

Words were printed in red or green color (RGB values of 220, 0, 0, and 0, 170, 0, for red and green, respectively). Words appeared in uppercase Arial font, bold and size 30. Participants sat 60cm from the 17" monitor so stimuli occupied a visual angle of up to 4.77 degrees. Stimuli were presented using Python™, which also recorded response times to the 1ms. On each trial, presentation of a fixation cross for 500ms was followed by a blank screen for 500ms, then followed by the stimulus (a single word in color) for a maximum of 4000ms. The presentations of stimuli were response terminated.

At the end of the second session, participants filled in a paper-and-pencil word identification questionnaire comprising all 24 experimental words. Participants identified how they responded to each word during the experiment: neutral, emotionally positive or emotionally negative. This further validated the pilot study and was used to confirm the correct interpretation of word emotionality. Participants completed the Beck Depression Inventory®–II (BDI-II), a psychometric assessment measuring depressive symptoms (Beck, Steer & Brown, 1996), and the Depression, Anxiety and Stress Scale (DASS 42) (Lovibond, & Lovibond, 1995). Both are self-administered, paper and pencil questionnaires consisting of 21 and 42-items for the BDI-II and DASS 42, respectively. Participants identified how they felt by checking one item out of four statements. BDI-II and DASS demonstrated strong validity and reliability in both nonclinical and clinical populations (Sharp & Lipsky, 2002; Storch, Roberti & Roth, 2004; Crawford, & Henry, 2003).

Procedure

Each participant performed both the control emotional Stroop task and the forced-processing emotional Stroop task on separate sessions. Task order was counterbalanced and separated by a minimum of 24 hours. On a given trial, a single word was presented in the centre of a white background. The tasks used the exact same stimuli, but differed in instructions. In the forced-processing task, participants were asked to classify the print color of the word as red or green, but use different keys (even for the same color) depending on the emotional valence of the word. Thus, they had to process the color as well as the emotionality of the item. Response keys were designated "emotional red", "emotional green", "non-emotional red" and "non-emotional green". The control task involved a different role that also required scanning of the letters but did *not* require processing of the emotional content; participants identified the color of the word and whether or not it contained an italic letter. The four response options were "italic red", "italic green", "non-italic red" and "non-italic green".

Button responses were made on a Cedrus response pad and button locations were counterbalanced. Experimental testing was conducted in a quiet, dimly lit room with air-conditioning. Participants were given task instructions followed by 12 practice trials with automated responses. This was followed by 20 practice trials with participant participation and feedback and then 20 practice trials without feedback. Data were collected in the subsequent experimental blocks, with a forced one-minute break between blocks. On conclusion of the second session, participants completed the item classification questionnaire, the BDI-II and DASS.

Data Analysis

Response speed and accuracy for negative and positive trials were practically the same, so we collapsed them into one category – emotional items. A 2 by 2 within-subjects

analysis of variance (ANOVA) was conducted on the factors task (control, forced) and emotional condition (emotional, non-emotional). Paired-samples t-tests were conducted on emotional condition (emotional, non-emotional) for both tasks. Bonferroni adjusted alpha level were applied where appropriate to account for familywise error rate. The dependent measure was reaction time (RT). Data satisfied the ANOVA assumptions with independent observations, normally distributed residuals and homoscedasticity.

Results and Discussion

Exclusion Criteria

Accuracy rate across participants was high (M = 94%, SD = 3.17%) with no participants excluded due to poor accuracy. Practice trials, non-responses, and error trials were excluded. Trials with RTs below 200ms and trials slower than 2.5 standard deviations below the mean RT were excluded from further analysis. Results were similar when we excluded responses 2.5 and 3 standard deviations from the mean, either way. Analyses of the word identification questionnaire revealed word valence (emotional, non-emotional) was identified incorrectly for only 2.25% of words.

Accuracy Analysis

Accuracy analysis revealed accuracy did not differ significantly across tasks and conditions. Accuracy was slightly poorer on emotional (M = 93%, SD = 4.26%) than non-emotional (M = 94%, SD = 3.04%) items, F (1, 36) =1.19, p = .284. Accuracy was slightly poorer in the forced task (M = 93%, SD =4.26%) than the control task (M = 94%, SD = 3.04%), F (1, 37) = 0.42, p = .521. Accuracy analysis ruled out the possibility of a response time-accuracy trade-off, so further analyses focus only on RTs.

Emotional Stroop Data

Mean RTs are presented in Figure 2 and can be compared with the predictions of the three theories illustrated in Figure 1. There was a significant ESE in the control task and a significant but reversed ESE in the forced task. These results are consistent with the response-relevance hypothesis of emotional processing (right panel of Figure 1). We performed a 2 (task: control, forced) x 2 (condition: emotional, non-emotional) within-subjects ANOVA on participants mean RTs. The control task RTs were significantly slower (M = 962ms, SD =201) than the forced task (M = 926, SD = 161), F (1, 36) = 5.25, p = .028. Across tasks, RTs were significantly slower for the non-emotional (M = 951, SD = 177) compared to the emotional condition (M = 937, SD = 177), F (1, 36) = 4.73, p = .036. Consistent with predictions of the response-relevance hypothesis, a significant interaction effect was evident between task and condition, F (1, 36) = 31.21, p < .001. Results for individual participants are detailed in Table A1 (Appendix).

Post-hoc paired-samples t-tests further examined the ESE for each task, with Bonferroni adjusted alpha level of .025 to account for familywise error rate. For the control task, emotional items (M = 970, SD = 203) were significantly slower than non-emotional items (M = 954, SD = 201) with a positive ESE of 16.3ms, t (36) = 3.96, p< .001. For the forced task, non-emotional items (M = 948, SD = 164) were significantly slower than emotional items (M = 905, SD =164) with a reversed ESE of -42.9ms, t (36) = -4.017, p < .001. Thus, we see a significant ESE in the control task and a significant but reversed ESE in the forced task. The reversed Stroop effect in the forced task is consistent with the delayed disengagement hypothesis (compare results in Figure 2 with theory predictions in Figure 1). Disengagement was required in the control task, slowing emotional processing, and emotional processing was facilitated by a lack of disengagement in the forced processing task.

Experiment 1 produced a significant ESE in the control task and significant but reversed ESE in the forced task, supporting a disengagement view of processing. Two processes, explicit and implicit, may offer an alternative explanation of Experiment 1 data. We now discuss these processes and then offer Experiment 2 to resolve the ambiguity.

Experiment 1 data supports Verges and Estes's (2008) response-relevance hypothesis. However, the results of Experiment 1 could be explained by differences between explicit and implicit emotional processing rather than delayed disengagement. The forced-processing task required explicit processing whereas the control task did not. In their discussion, Verges and Estes proposed that the processing interference, observed in emotional items, may only occur when valence is processed explicitly. Furthermore, increasing the salience of the emotional content of items, in the forced processing task, may be causing a significant change in emotional processing rather than change being driven by a task effect. Therefore, the pattern of results in Experiment 1 could be accounted for by a change in salience of emotional content.

Experiment 2 is aimed to resolve this ambiguity. Experiment 1 does not discriminate between an interference effect caused by a difference in processing (implicit, explicit) or type of task (lexical, valence). In Experiment 1, the forced-processing task involves a valence decision-making task, which requires explicit emotional processing. On the other hand, any emotional processing undertaken on the control task is implicit. Thus singling out the cause of interference is not possible. To shed light on this alternative interpretation, we developed a second experiment involving two *implicit* emotional processing tasks. Rather than a valence task we added a lexical-decsion task for which emotional processing is implicit. The primary methodological difference is this forced task involves asking participants to differentiate between words and non-words rather than judging emotional valence. This ensures

participants process word content without explicit emotional processing. Additional non-words were included in list of words used for both tasks.

With minimal alterations to the method, Experiment 2 also allowed exploration of Sharma and McKenna's (2004) slow effect. In their highly influential paper, Sharma and McKenna hypothesised that interference to color processing could come from the emotional content of the presented word (fast effect) but also from emotional words on *previous* trials (slow effect). Considering that blocked trials of emotional items revealed a larger Stroop effect than mixed trials, Sharma and McKenna proposed interference caused by an emotionally charged word was carried across trials. This carry-over-across-trials effect, or slow effect, was contrasted to the *fast* effect, which is the concurrent interference caused by the presented item. For example, in a sequence of colored words such as PAINT, DEPRESSION, and SAD, slow response latencies for the last item could be due to its emotional content (fast effect), or due to the emotional content of the item that precedes it (the slow, carry-over effect), or both. Previously, the literature has failed to isolate these two processes and had presented data that was potentially an amalgamation of both processes.

Conditional response-time analyses were conducted in Experiment 2 to explore slow (effect of previous items) and fast (effect of currently presented item) effects. We compared RTs of non-emotional trials that followed an emotional item (e.g., the sequence SAD, SAT) with RTs of non-emotional trials that followed a non-emotional item (PAINT, SAT). To isolate the effect of the previous trial, we controlled for the condition of the current item, using only non-emotional items. Items were then grouped into conditions by the identity of the *previous* item. For example, an emotional item effect was measured as the RT to the word SAT when preceded by the word SAD. A non-emotional item was measured as the RT of SAT when preceded by PAINT. This design, and analysis, can draw out naturally occurring sequences of interest without relying on a pseudorandom order for the sequences to occur.

Experiment 2

Method

Participants

A new sample of 22 participants from the same pool of Newcastle University students were recruited (5 male, 17 female: *Mage* = 22.45 years, *SD* = 6.43, age range: 18-46).

Apparatus and Materials

The stimulus set comprised the same 24 words from Experiment 1 with the addition of 24 matching non-words (see, earlier for Table 1). The non-word stimuli comprised pseudo-word anagrams of the existing word stimuli. For example, BETTER, BITTER, and BUTTER were matched by BETERT, BETIRT, and BETURT, respectively. Anagram variants of the word stimuli were used to reduce variability between the word and non-word stimuli. These non-words were pronounceable to prevent participants from responding based on the illegal orthography of non-pronounceable non-word letter-strings, and instead forcing participants to read each word or non-word in full. Where possible, the first letter of each non-word remained the same as its word counterpart so that participants were unable to respond based on the local cue of the initial letters of each string. The orthographic Levenshtein distances of non-words from their word counterparts were matched to the orthographic distances of other non-words of equal length to reduce variability between non-word stimuli (Yarkoni, Balota & Yap, 2008).

Procedure

Participants completed the two tasks, forced-processing task and control task, on separate days. The procedure was similar to Experiment 1 with a single variation in the forced processing task: participants were instructed to identify whether the stimulus presented was a word (BETTER) or non-word (BETERT), and classify the color (red or green). Hence,

participants responded to stimuli by pressing one of four pre-assigned buttons, indicating if the stimulus was a red word, a green word, a red non-word, or a green non-word. The stimulus list was similar for both tasks. Apart from the addition of non-words, the control task procedure remained identical to Experiment 1; respond to font (italic, non-italic) and color (red, green).

Data Analysis and Accuracy Analysis

Analyses were the same as Experiment 1. Accuracy for all participants was high (M = 93.31%, SD = 5.29%). No participants were excluded from analyses due to poor accuracy. Accuracy was significantly lower on non-emotional (M = 92.14%, SD = 5.79) than emotional (M = 94.48%, SD = 4.79) words, F (1, 17) = 19.15, $p < 0.001$. There was no significant difference in accuracy between the forced and control tasks, F (1, 17) = 0.08, p = 0.784. There was no evidence of an RT-accuracy trade-off.

Results and Discussion

Emotional Stroop Data

Results are presented in Figure 3 and were consistent with Experiment 1, with a reversed ESE in the forced task and a positive, though non-significant ESE in the control task. We performed a 2 (task: control, forced) x 2 (condition: emotional, non-emotional) within-subjects ANOVA on participants' mean RT. RTs were significantly slower in the control task (M = 1,081ms, SD = 358) compared to the forced task (M = 937, SD = 205) F (1, 21) = 12.51, p =.002. There was no RT difference between the emotional (M =1,009, SD = 262) and non-emotional conditions (M = 1,009, SD = 265), F (1, 21) =.002, p = .964. The interaction effect across tasks and conditions was significant, F (1, 21) = .13.06, p = .002. The pattern of interaction (Figure 3) reflected that of Experiment 1 (cf Figure 2). Results for individual participants are detailed in Table A2 (Appendix).

Post-hoc paired-samples t-tests further examined the ESE for each task, with Bonferroni adjusted alpha level of .025. In the control task, emotional items (M =1,088, SD = 274) were significantly slower than non-emotional items (M =1,074, SD = 279) with an ESE of 14.2 ms, t (21) =2.55, p = 0.02. For the forced task, the effect reversed, with emotional items significantly faster (M = 929, SD =160) than non-emotional items (M = 944, SD = 156), documenting a reversed ESE of -13.9ms, t (21) = -3.13, p = .005. Thus, the results of Experiment 2 replicated those of Experiment 1, with a significant ESE in the control task and a significant but reversed ESE in the forced task. Consist results across experiments gives evidence of a response relevance explanation rather than an alternative influence of explicit processing.

Slow Emotional Stroop Effect

The effects reported hitherto are the so-called "fast effects", in which trials are categorized as emotional or non-emotional based on the identity of the currently presented item. To examine slow emotional Stroop effects, analyses were repeated with all categorisations into emotional or non-emotional now conditioned of the identity of the previous, rather than current trial (i.e., mean RTs of current non-emotional trials, conditioned on the identity of the previous trial). A 2 (task: control, forced) x 2 (condition: emotional, non-emotional) within-subjects ANOVA found a significant RT increase in the control task (M = 1081, SD = 277) compared with the forced-processing task (M = 936, SD = 158), F (1, 21) = 12.25, p = .002. The main effect of condition was not significant with practically the same mean RT for non-emotional (M = 1009, SD = 207) and emotional items (M = 1008, SD = 201), F (1, 21) = .103, p = .752, and a minimal interaction between task and condition, F (1, 21) = 2.96, p = .1. Slow effects were thus analysed but results were not significant and will not be explored in detail.

General Discussion

Obligatory vs Non-Obligatory Processing in the Emotional Stroop Task

In two experiments with different participants and slight procedural differences we observed a positive ESE in the control task that does not force emotional processing, and a reversed ESE in a novel task that forces emotional processing. These results support our hypothesis that emotional processing is not obligatory in the ordinary (control) emotional Stroop task. If it were, the effect should have been the same regardless of whether emotional processing is forced or not. With variation to the task instructions, that forced participants to process all items, a qualitatively different ESE was identified. This suggests that, rather than an automatic process, participants process items differently when the task forces emotional processing. The change in instructions across tasks produced reversed ESEs, whereby participants processed emotional items faster when emotional processing is forced or required for performance (forced task). When emotional processing is not required (control task), emotional items slowed performance.

An increase in emotional processing was evident in the forced task compared to the control task, offering support for the non-obligatory view (cf Figure 1). In order to directly explore obligatory processing in the emotional Stroop task, a novel task was implemented that ensured participants process all items. This task was then compared to the control emotional Stroop. The obligatory view of processing proposes that every word is processed in the control task, thus the two tasks should manifest a similar ESE. Alternatively, the non-obligatory view supposes that not *all* words are processed in this task. Results demonstrated a larger ESE in the forced task compared to the control task, ruling out the obligatory view.

Delayed Disengagement in the Emotional Stroop

The negative ESE recorded when emotional processing was required, and positive effect when it was not required (i.e., the interaction between task and emotional condition)

suggests that processing is affected by a disengagement process. When participants are required to respond to font and color, and therefore emotional content is not relevant, disengagement with emotional processing hinders performance slowing the response for emotional items. Alternatively, when participants are required to processes emotional content, without disengagement, emotional responses are facilitated and are faster than neutral words. This suggests that processing of emotional content alone does not slow down performance, and could even speed up response times. A delay only occurs when there is need to disengage with emotional processing. The theoretical implications of these findings are discussed below.

Verges and Estes (2008) proposed that a fleeing or fighting response can occur, alternative to a freeze response. In this process, selective responding is occurring dependent on whether emotional processing is relevant for performance. Verges and Estes argue that in a flee or fight response, the motor system is prepared for action thus emotional responding induces a faster reaction. Emotional words only slow performance if emotional processing is irrelevant for performance as disengaging with emotional processing is required.

Verges and Estes (2008) also suggested that, rather than driven by task relevance, faster responses for negative words may be evident in *any* task for which valence is processed explicitly. In Experiment 1, only the forced task which involved explicit emotional processing caused a facilitation for emotional items. In Experiment 2 we exchanged the explicit for an implicit emotional processing task.

Focusing on Experiment 2 data, we can compare a lexical decision task with emotional processing (control task), to a lexical decision task with implicit emotional processing (forced-task). In this latter task, participants are asked to identify if the item is a word or non-word. Here participants are forced to read the entire word but not respond to the valence; thus implicitly processing the emotional content of the word. Interestingly, in this

forced task responses to emotional items were faster compared to neutral items despite the implicit processing. The results across both experiments demonstrate that the when disengagement is necessary emotional items facilitated responding in both implicitly- and explicitly-processed emotional trials. Results suggest that processing emotional words implicitly or explicitly can lead to a facilitation effect when disengagement is not required. This suggests that the effect is driven by engagement and disengagement rather than explicit emotional processing.

Attention Bias and Depression

Inferences derived from ESE analysis are not limited to attentional bias but applicable to other cognitive phenomena. Since initially proposed by Beck (1979), it has been established that differences in attention bias towards emotional content form a valid index for exploring depression (Bradley, Mogg, Fella, & Hamilton, 1998; Okon-Singer, Tzelgov, & Henik, 2007). This link between attention bias and psychopathology underpinned Beck's (1979) negative schematic theory of depression. The current conceptualisation of the relationship between attentional bias and psychopathology is bidirectional. Not only is attentional bias a by-product of emotional disorders but the bias plays a role in forming and also maintaining emotional disorders (Williams el al., 1996). Negative experiences in early childhood could lead to the development of dysfunctional beliefs; generally inactive, these beliefs can become activated by a negative life event and trigger a depressive episode. The cyclical role of attention bias in emotional disorders highlights the necessity to understand and measure the extent to which emotional attention bias presents in depression. Considering this theroetical foundation, attentional biases using the EST have been found within several clinical populations; however, the extent to which depressed populations display a true ESE remains contentious (Mogg & Bradley, 2005; Gotlib & McCann, 1984; Hills & Knowles, 1991; Williams el al., 1996).

Correlational analysis was conducted on BDI-II scores and ESE for both experiments. In Experiment 1, a very weak nonsignificant linear relationship was found between BDI-II scores and ESE in the control and forced task, r (37) = -.244, p = .145, and r (37) = -.065, p = .704, respectively. Considering the lack of significance and the problems associated with null hypothesis testing, Bayes Factor regression analyses were conducted (cf. Rouder & Morey, 2011). This calculation examined the odds ratio of the null hypothesis (no relationship exists) over the alternative model (correlation exists). Bayes Factor values greater than one indicate higher likelihood for the null- over the alternative model. Bayes Factor regression analysis for BDI-II scores and the ESE was completed for both control and forced tasks, Bayes Factor = 1.332, Bayes Factor = 2.964, respectively. For both tasks then, the results suggest that the null model (no relationship) was more likely than the alternative model (relationship exists). Identical analysis was conducted for Experiment 2. again showing a very weak nonsignificant linear relationship between BDI-II scores and ESE in both the control and forced tasks, r (22) = 0.16, p = .491, and r (22) = -.132, p = .559, respectively. Bayes Factor regression analysis was again in favour of the null for both regression analysis, Bayes Factor = 2.178, Bayes Factor = 2.288, respectively. It is possible that the range of depression scores within this sample was limited, lacking sufficiently large (clinically-depressed) scores to reveal an effect. Use of a clinical population would best clarify the relationship between depression and ESE.

In addition to the theoretical implications discussed so far, Williams, Mathews and McLeod (1996) suggested a theoretical link between ESEs and rumination in emotionally disturbed populations, which was not the focus of this study. Rumination is the persistent and focused attention on negative emotions and it is suggested to increase the resting activation level for emotional processing (Nolen-Hoeksema, 2000). Recent literature has suggested a link between delayed disengagement and rumination (Koster, De Raedt, Goeleven, Franck, & Crombez, 2005). Trait rumination and delayed disengagement have not explicitly been

examined using the emotional Stroop paradigm. It is possible that the failure to disengage from emotional stimuli in the EST reflects a ruminative process. Future research into the link between emotional disengagement and rumination may offer insight into the cognitive mechanisms that underlie rumination.

Probability-Mixture Model for Emotional Processing

Theoretical and practical implications of this study are significant. Emotional Stoop data has been understood to illustrate an interference of emotional items on processing. Assumptions have been made about the EST, that participants are processing emotional content on all items, despite it irrelevance, and emotional processing infers with performance. We have demonstrated that when emotional processing is forced, slowed interference does not occur. Rather, emotional processing facilitates performance. Thus we propose that emotional processing in the EST does not occur on each item and that it does not always interfere with performance. The study supports a non-obligatory view of emotional processing. We demonstrated that the typical ESE is likely derived from partial (or shallow) word processing rather than obligatory processing. Additionally, our results offer supportive evidence for Verges and Estes (2008) task-relevance hypothesis. ESE is likely derived from partial word processing, and may also be driven by a disengagement effect.

The emotional Stroop task claims to measure attentional bias towards emotional stimuli in our environment. The finding that some but not all emotional stimuli draw attentional bias is significant as it undermines the foundations the ESE analysis. Two consequences are evident. Firstly, inferences made based on the ESE, both theoretical and applied, ought to be considered with caution. Secondly, changes to the control emotional Stroop task are necessary to ensure consistency in item processing depth and frequency. Without methodological modifications the common ESE analysis, a comparison of means is fundamentally flawed.

A simple possible account of the observed results is a probability-mixture model. This model holds that emotional stimuli are processes on some but not all trials. Any observation from the emotional Stroop task is drawn from either a probability distribution of processing emotional content (probability of processing = *p*) or from a distribution of not processing (probability of not processing = 1-*p).* Emotional Stroop data come from a mixture of these two probability distributions. Increasing the probability of processing on a given trial should increase the observed effect. The probability of processing was increased in this study by forcing participants to process every trial. This forced task produced a significant reversed ESE thus supporting the non-obligatory view of emotional processing in the emotional Stroop task.

Conclusions

People have a compulsion to process emotions; so readily that some believe the process to be automatic. Due to the urgency for emotional processing, for social interactions and survival, additional resources can be directed towards the item, creating an attention bias. The ESE attempts to quantify this attentional bias towards emotional stimuli. Decades of inferences about human behaviour have been derived from two assumptions; that emotional processing is automatic and emotional words slow processing. The current study offers a twofold development in our understanding. Firstly, it challenges a fundamental assumption made in the ESE analysis, that participants process the emotional content of every word regardless of instructions not to. The current ESE analysis lacks the inferential power to determine frequency of processing and thus a conclusion that processing is obligatory may be misguided. This study offers evidence that participants process the emotional content of some but not all items in the emotional Stroop. Secondly, it provides further evidence that the ESE is impacted by the disengagement of emotional stimuli. In sum, we have demonstrated that emotional processing in the EST does not occur on every item (at least not to its full extent)

and it does not always interfere with performance. These results undermine our fundamental understanding of the emotional Stroop effect and offer new directions for future research.

References

Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a stroop effect. *Journal of Experimental Psychology: General, 133* (3), 323-338. Retrieved from http://dx.doi.org/10.1037/0096-3445.133.3.323

Bradley, B. P., Mogg, K., Falla, S. J., & Hamilton, L. R. (1998). Attentional bias for threatening facial expressions in anxiety: Manipulation of stimulus duration. *Cognition & Emotion, 12* (6), 737-753. doi:10.1080/026999398379411

Carretié, L. (2014). Exogenous (automatic) attention to emotional stimuli: a review. *Cognitive, Affective, & Behavioral Neuroscience, 14* (4), 1228-1258. Doi: 10.3758/s13415-014-0270-2

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11*(6), 671-684. doi: 10.1037/h0084237

Crawford, J. R., & Henry, J. D. (2003). The depression anxiety stress scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42* (2), 111-131. doi: 10.1348/014466503321903544

Eidels, A., Townsend, J. T., & Algom, D. (2010). Comparing perception of Stroop stimuli in focused versus divided attention paradigms: Evidence for dramatic processing differences. *Cognition, 114*(2), 129-150. doi:10.1016/j.cognition.2009.08.008

Eidels, A., Williams & Ryan, K. (2014). Depth of processing in the Stroop task: evidence from a novel forced-reading condition. *Experimental Psychology, 61* (5). Retrieved from http://dx.doi.org/10.1027/1618-3169/a000259

Estes, Z., & Verges, M. (2008). Freeze or flee?: Negative stimuli elicit selective responding. *Cognition, 108*, 557-565. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027708000760

Fox, E. (1993). Attentional bias in anxiety: Selective or not? *Behaviour Research and Therapy, 13* (5), 487-493. doi: 10.1037/0033-2909.133.1.1

Fox, E., Russo, R., & Georgiou, G. A. (2005). Anxiety modulates the degree of attentive resources required to process emotional faces. *Cognitive, Affective, & Behavioral Neuroscience, 5*(4), 396-404. doi: 0.3758/CABN.5.4.396

Gotlib, I. H., & McCane, C. D. (1984). Construct accessibility and depression: An examination of cognitive and affective factors. *Journal of Personality and Social Psychology, 47*(2), 427-439. doi: 10.1037/0022-3514.47.2.427

Hill, A. B., & Knowles, T. H. (1991). Depression and the "emotional" Stroop effect. *Personality and individual differences, 12*(5), 481-485. Retrieved from http://www.sciencedirect.com/science/article/pii/019188699190066K

Koster, E. H., De Raedt, R., Goeleven, E., Franck, E., & Crombez, G. (2005). Mood-congruent attentional bias in dysphoria: maintained attention to and impaired disengagement from negative information. *Emotion, 5*(4), 446. Retrieved from http://dx.doi.org/10.1037/1528-3542.5.4.446

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behaviour Research and Therapy, 33*(3), 335-343. doi:10.1016/0005-7967(94)00075-U

MacLeod, C., (1991). Half a century of research on the stroop effect: An integrative review, *Psychological Bullitin, 109(2),* 163-203. Retrieved from http://dx.doi.org/10.1037/0033-2909.109.2.163

MacLeod, C., & Hagan, R. (1992). Individual differences in the selective processing of threatening information, and emotional responses to a stressful life event. *Behaviour Research and Therapy, 30,* 151-161. Retrieved from http://www.sciencedirect.com/science/article/pii/0005796792901387

Martin, M., Williams, R. M., & Clark, D. M. (1991). Does anxiety lead to selective processing of threat-related information? *Behaviour Research and Therapy, 29*(2), 147-160. Retrieved from http://www.sciencedirect.com/science/article/pii/0005796791900433

Mogg, K., & Bradley, B. P. (2005). Attentional bias in generalized anxiety disorder versus depressive disorder. *Cognitive Therapy and Research, 29*(1), 29-45. doi: 10.1007/s10608-005-1646-y

Mogg, K., Bradley, B. P., Williams, R., & Mathews, A. M. (1993). Subliminal processing of emotional information in anxiety and depression. *Journal of Abnormal Psychology, 102* (2), 304-311. Retrieved from http://webs.wofford.edu/steinmetzkr/Teaching/Psy310Lab/Mogg.pdf

Mogg, K., & Marden, B. (1990). Processing of emotional information in anxious subjects. *British Journal of Clinical Psychology, 29*(2), 227-229. doi: 10.1111/j.2044-8260.1990.tb00874.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology, 109* (3), 504-511. doi: 10.1037/0021-843X.109.3.504

Okon-Singer, H., Tzelgov, J., & Henik, A. (2007). Distinguishing between automaticity and attention in the processing of emotionally-significant stimuli. *Emotion, 7*, 147-157. doi: 10.1037/1528-3542.7.1.147

Phaf, H. R., & Kan, K. J. (2007). The automaticity of emotional Stroop: A meta-analysis. *Journal of Behaviour Therapy and Experimental Psychiatry, 38*, 184-199. doi:10.1016/j.jbtep.2006.10.008

Richards, A., French, C. C., Clader, A. J., Webb, B., & Rox, R. (2002). Anxiety-related bias in the classification of emotionally ambiguous facial expressions. *Emotion, 2*(3), 273-287. doi: 10.1037//1528-3542.2.3.273

Segal, Z. V., Gemar, M., Truchon, C., Guirguis, M., & Horowitz, L., M. (1995). A priming methodology for studying self-representation in major depressive disorder. *Journal of Abnormal Psychology, 104*(1), 205-213. doi: 10.1037/0021-843X.104.1.205

Sharma, D., & McKenna, F. P. (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 382-392. doi: 10.1037/0278-7393.30.2.382

Tzelgov, J. (2002). Trading automatic/nonautomatic for unconscious/conscious. *Behavioral and Brain Sciences, 25*(03), 356-357. doi: http://dx.doi.org/10.1017/S0140525X02500066

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1), 3-24. Retrieved from http://brainimaging.waisman.wisc.edu/~perlman/papers/stickiness/WilliamsEmoStroop1996.pdf

Wyble, B., Sharma, D., & Bowman, H. (2005). Modelling the slow emotional Stroop effect: Suppression of cognitive control. *Progress in Neural Processing, 16*, 291. Retrieved from https://www.cs.kent.ac.uk/pubs/2005/2011/content.pdf

Appendix

Individual Subjects Results

Table A1

Individual Subjects Results Experiment 1

Participant	ESE Forced-Processing Task	ESE Control Task	BDI-II score	DASS Depression scores
1	13.77	-114.7	3	2
2	-19	-28.5	23	23
3	12.88	-6.86	20	14
4	-13.8	-16.1	12	1
5	51.4	-64.06	12	5
6	31.04	-34.51	10	5
7	-7.41	-77.05	10	1
8	76.2	-72.52	4	1
9	12.19	-109.6	26	21
10	-18.07	-118.77	2	0
11	14.2	-20.36	37	26
12	46.4	75.7	8	5
13	5.19	-73.46	1	0
14	-43.1	-54.44	31	23
15	28.2	154.65	4	0
16	21.31	-60.1	14	9
17	0.38	-94.46	11	3
18	20	68.17	5	0
19	64.5	-124.6	14	6
20	-12.9	-71.98	15	7
21	65.4	60.6	15	11
22	19.7	24.67	5	2
23	28.7	-106.38	5	4
24	6.51	-11.4	5	5

25	3.73	-19.98	0	1
26	6.68	-73.97	26	23
27	10.01	-136.93	8	4
28	18.42	-87.46	3	1
29	26.23	-74.65	7	3
30	18.89	-7.99	11	3
31	2.6	-53.17	17	10
32	52.35	20.53	25	16
33	22.5	28.43	5	0
34	26.18	-59.51	3	0
35	-9.82	-25.75	40	22
36	11.66	-127.61	31	23
37	10.5	-93.9	11	4

Table A2

Individual Subjects Results Experiment 2

Participant	ESE Forced-Processing Task	ESE Control Task	BDI-II score	DASS Depression score
1	-34.81	-21.1	22	16
2	-2.56	15.82	18	12
3	-44.85	48.27	30	26
4	-23.24	37.72	18	5
5	-16.99	-1.6	21	24
6	6.7	-0.4	31	22
7	-18.43	10.8	9	32
8	-27.3	20.2	5	8
9	-23.62	16.2	31	39
10	23.81	29	3	1
11	-5.23	29.92	17	30
12	-28.72	64.9	21	25
13	-37.1	6.6	3	3
14	-26.93	-8.48	31	39
15	-18.17	34.8	12	15
16	-1.22	-11.74	21	24
17	-26.7	-7.9	6	3
18	-31.57	25.56	3	4
19	40.3	-29.7	4	11
20	-4.34	53.5	34	35
21	12.18	28.33	21	28
22	-17.18	-28.3	19	17

Table 1

List of stimulus items ordered by condition for Experiment 1 and 2

Non-emotional	Positive emotional	Negative emotional	Non-word	Non-word	Non-word
ADDRESSED	IMPRESSED	DEPRESSED	ASEDDERDS	ISEMPERDS	DESEPERDS
DEPICTED			DICTEDEP		
BUTTER	BETTER	BITTER	BETURT	BETERT	BETIRT
LETTER			LETERT		
FERRY	MERRY	MISERY	FYRER	MYRER	MYRIES
MINISTRY			MYSTRINI		
LATELY	LOVELY	LONELY	LELATY	LELOVY	LELONY
DAILY			DILYA		
PAD	GLAD	SAD	DAP	GALD	DAS
SAT			ATS		
TENURE	PLEASURE	FAILURE	TERUNE	PERAUSEL	FERAULI
PAINT			PITAN		

Note. Both words and non-words were matched on frequency and length, where possible, to ensure participant could not rely on local cues to respond.

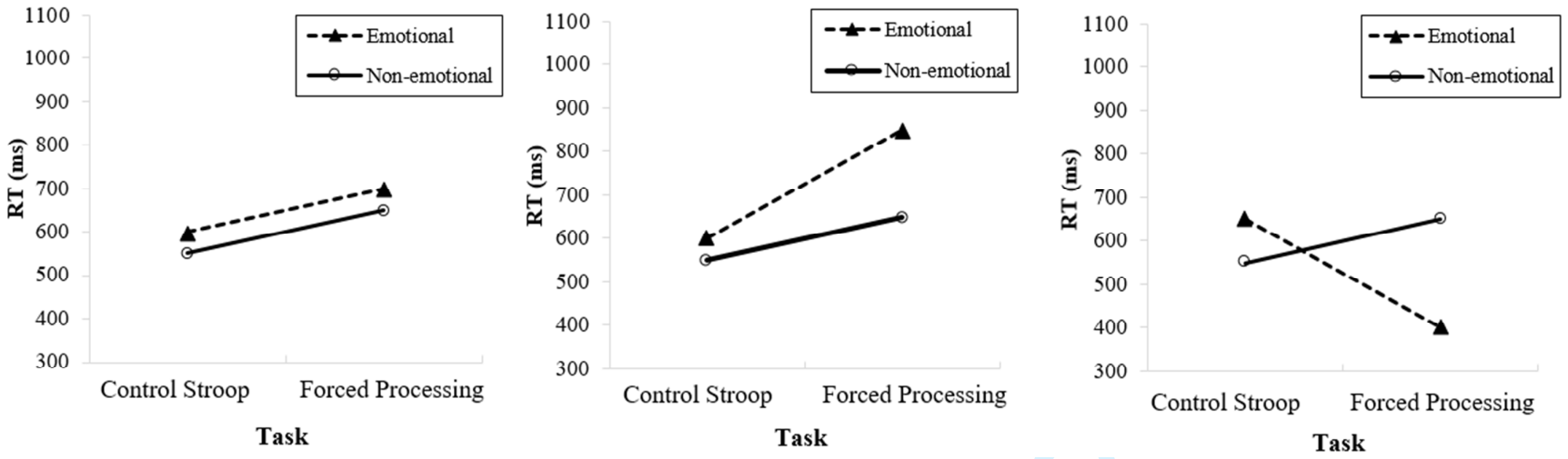


Figure 1. Mean reaction-time predictions for the obligatory view (left-hand panel), non-obligatory view (middle panel), and response-relevance hypothesis (right panel) of emotional processing in the forced-processing and control Stroop tasks.

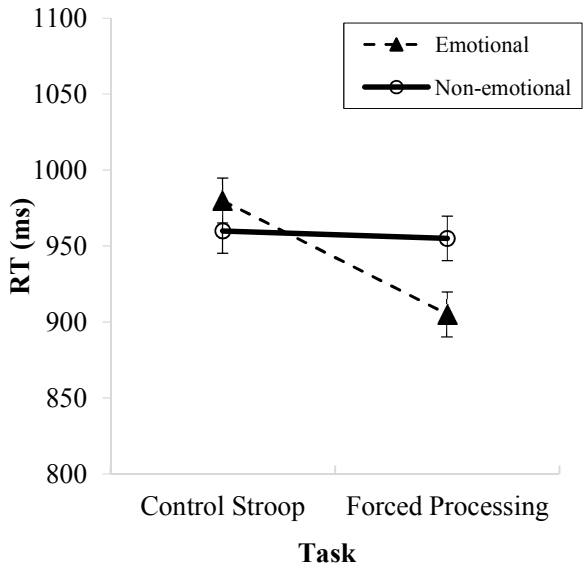


Figure 2. Results of Experiment 1: Mean reaction time (RT) for correct classification of print color as a function of condition (emotional, non-emotional) and task (control emotional Stroop, forced-processing emotional Stoop). Error bars represent one standard error around the mean.

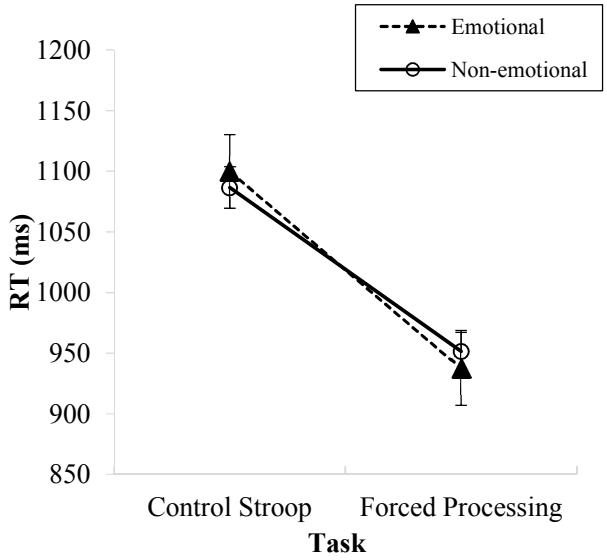


Figure 3. Results of Experiment 2: Mean reaction time (RT) for correct identification of print color as a function of condition (emotional, non-emotional) and task (control Stroop, forced-processing). Error bars represent one standard error around the mean.

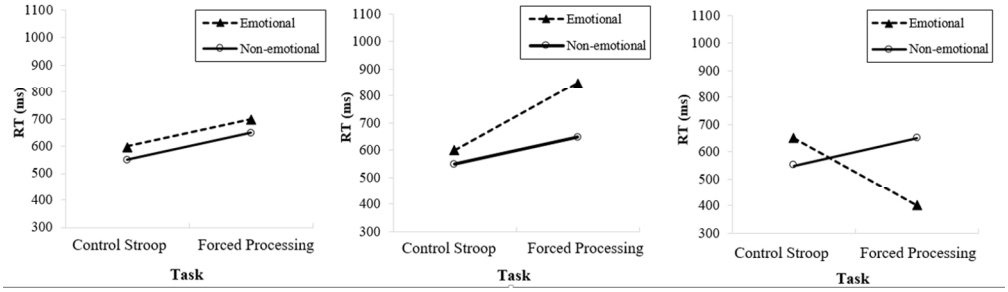


Figure 1. Mean reaction-time predictions for the obligatory view (left-hand panel), non-obligatory view (middle panel), and response-relevance hypothesis (right panel) of emotional processing in the forced-processing and control Stroop tasks.

343x102mm (96 x 96 DPI)

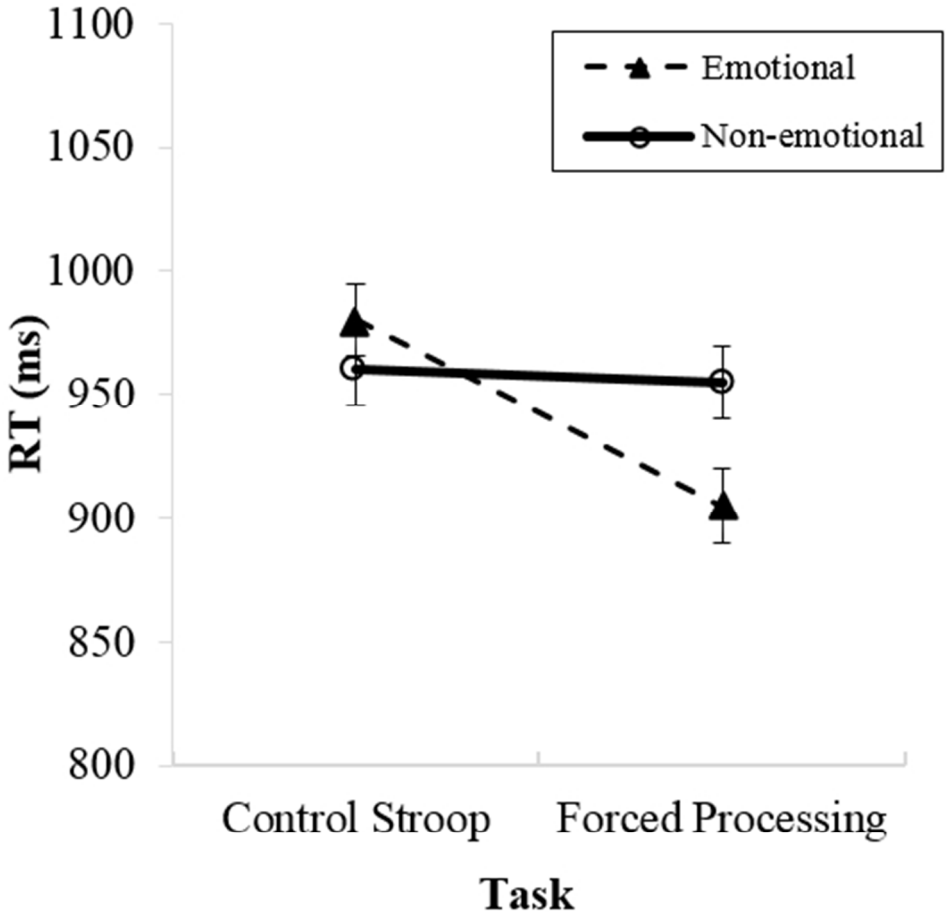


Figure 2. Results of Experiment 1: Mean reaction time (RT) for correct classification of print color as a function of condition (emotional, non-emotional) and task (control emotional Stroop, forced-processing emotional Stoop). Error bars represent one standard error around the mean.

138x132mm (96 x 96 DPI)

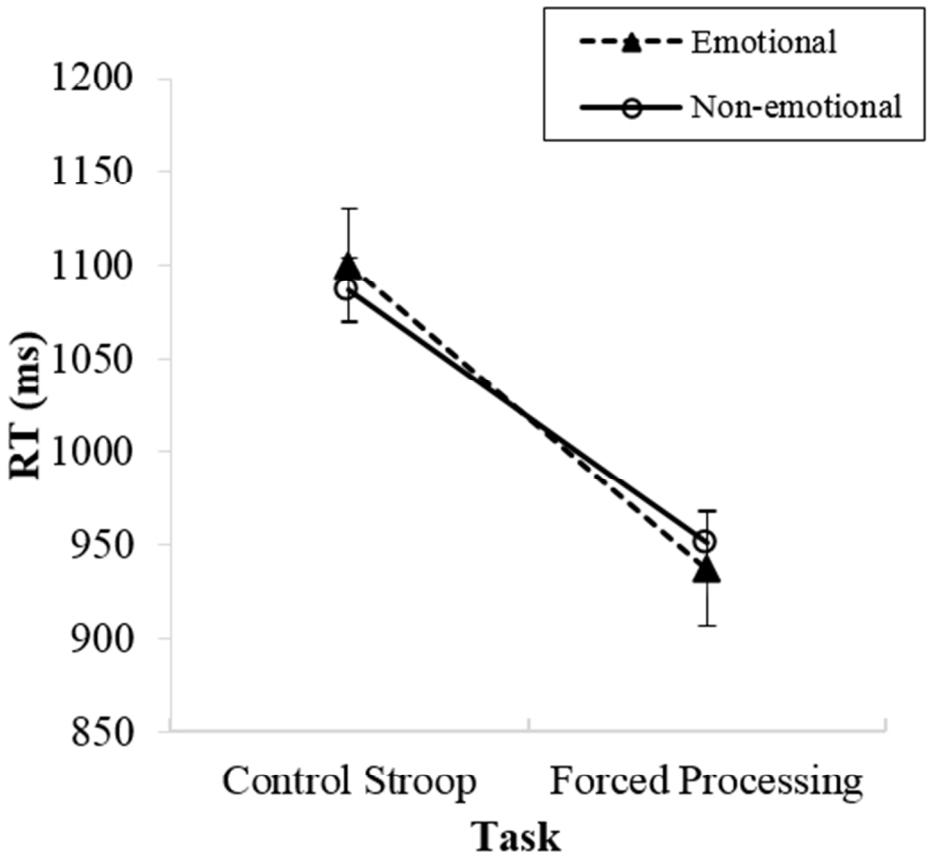


Figure 3. Results of Experiment 2: Mean reaction time (RT) for correct identification of print color as a function of condition (emotional, non-emotional) and task (control Stroop, forced-processing). Error bars represent one standard error around the mean.

146x130mm (96 x 96 DPI)

Chapter 6

In Section 1, I outlined the development and application of *The Buckets game*, which we used to explore the theoretical and empirical links between the hot hand and post-error slowing. I also documented and compared three methods for calculating post-error slowing. The use of multiple methods was novel and constructive as it allowed us to compare various contributions to post-error effects. That is, we were able to assess the local contributions of errors independently of global contributions such as fatigue or boredom.

In Section 2, so far, I have outlined a methodological review and best practice guideline for the implementation of the emotional Stroop task, and then built on this understanding to develop modified versions of the task that we presented to participants assessed for depression symptoms. We used these modified versions to discriminate between three alternative views of emotional processing in the emotional Stoop task: obligatory, non-obligatory, and task dependent. We also explored, critically in terms of this thesis, fast and slow emotional Stroop effects (ESEs).

In Chapter 6 I document the use of two of the post-error slowing measurement methods outlined in Section 1 - the *traditional method* and the *robust method* - with a version of the classic emotional Stroop task outlined to begin Section 2. The current body of work thus comes full circle and forms a coherent whole. Chapter 6 contains three components. The first component is a targeted literature review of *post-error slowing and depression*, which briefly introduces our motivation to explore this effect using the emotional Stroop task. The *Paper 6 Overview* follows this literature review and highlights the unique contributions of Paper 6. This overview might be best read in conjunction with *Paper 6*, which is presented in full to conclude the chapter.

Post-error Slowing and Depression

Since the seminal findings of *post-error slowing* (Laming, 1968, 1979a, 1979b; Rabbitt, 1966, 1969, 1979; Rabbitt & Rodgers, 1977; Rabbitt & Vyas, 1970, 1981), references to post-error slowing have appeared in over 2100 peer-reviewed articles[1]. Because of its pervasive nature in fast-paced choice tasks, post-error slowing has provided an important benchmark in testing theories of cognitive control (e.g., Botvinick, Braver, Brach, Carter, & Cohen, 2001; Gehring, Goss, Coles, Meyer, & Donchin 1993; Holroyd & Coles, 2002 ; Yeung, Botvinick, & Cohen, 2004). The assumption that post-error slowing is intrinsically linked to cognitive control and adaptive behaviour - has led to a proliferation of clinical studies in which the effect is used to survey the functional aberrations of psychiatric diagnoses. These include depression (Compton, Lin, Vargas, and Quandt, 2008), attention deficit hyperactivity disorder (Shiels, Tamm, & Epstein, 2012; van Meel, Heslenfeld, & Oosterlaan, 2007), borderline personality disorder (de Bruijn et al., 2006), autism spectrum disorders (Bogte, Flamma, van der Meere, van Engeland, 2007; Thakkar, et. al., 2008), schizophrenia (Carter, MacDonald, Ross, & Stenger, 2001; Kerns et. al., 2005), as well as other anxiety related disorders (e.g., Proudfit, Inzlicht, & Mennin, 2013; Hajcak, McDonald, Simons, 2003).

As discussed in Section 1 Chapter 3, however, variation in the relationship between adjustments in response speed and accuracy has brought the caution, or cognitive control explanation of post-error slowing into question. According to models that explain post-error slowing as the result of an increase in caution

[1] A search using the keyword phrase "post-error slowing" was performed using Google Scholar on the 26/9/2016, with returns limited to individual articles for an unlimited time range.

following errors (Botvinick, Braver, Brach, Carter, & Cohen, 2001), slowing after an error should be associated with a proportional increase in accuracy. However, examples of post-error slowing associated with no increase in accuracy, or a decrease in post-error accuracy, are commonly found and have been used to support alternate explanations of the effect (e.g., Notebaert et al., 2009; see Danielmeier & Ullsperger, 2011, for a review). This uncertainty has brought into question the ever-expanding pool of applied research that assumes post-error slowing is a benchmark of cognitive control.

Inconsistent findings in clinically focused studies have further fuelled uncertainty in the use of post-error slowing as a behavioural benchmark. For example, few, if any studies have found clear and unambiguous differences for depressed populations when compared to normal populations for post-error slowing, or other behavioural indicators of cognitive control (Paulus, 2015; Pizzagalli, Peccoralo, Davidson, & Cohen, 2006; Saunders & Jentzsch, 2014). For depressed populations in particular, the lack of a consistent difference or differences in the behavioural markers of cognitive control is baffling, because neurophysiological studies have documented unambiguous differences in an executive or cognitive control system centred on the anterior cingulate cortex (ACC) and prefrontal cortex (PFC). This system is implicated in information processing and responding when adaptive behaviour is required, such as error monitoring and correction, and response inhibition (Dehaene, Posner, & Tucker, 1994; Miller & Cohen, 2001). This network functions abnormally in depression (Davidson et al., 2002). With respect to the ACC in particular, neurophysiological studies have documented structural (Ballmaier et al., 2004), neurochemical (Auer et al., 2000; Rosenberg et al., 2004), and functional (Beauregard

et al., 1998; George et al., 1997; Kumari et al., 2003) differences for depressed populations.

Interestingly, while there are no consistently documented differences in the behavioural markers of cognitive control for depressed individuals, there is clear and replicable evidence of differences in other cognitive functions. Experimental tasks exploring cognitive biases have provided strong support for the suggestion that depression is marked by negative automatic thoughts and biases in attention and memory (Mathews & MacLeod, 2005). Pertinently, unlike anxiety, depression is not characterized by a high level of alertness in the processing of negative or threatening material. Rather, once engaged with negative material (be it emotional stimuli, error feedback, or distracting automatic negative thoughts), depressed individuals seem to struggle with disengaging. That is, once they are engaged with negative material, depressed individuals struggle to re-orient to goal directed behaviour (for a full review, see Gotlib & Joorman, 2010). This deficit has been implicated in the inability in depressed individuals to regulate emotion by redirecting attention.

In sum, we are left with a fascinating puzzle. In depression, clear neurophysiological differences have been identified in the network that is implicated in error processing and error adjustments. In addition, clear differences in cognitive function have been documented for depressed populations. Yet surprisingly, behavioural differences in the markers of cognitive control for depressed populations have been lacking and inconsistent (Paulos, 2015; Pizzagalli et al., 2006; Saunders & Jentzsch, 2014). Interestingly though, depression had been associated with deficits in cognitive control when emotional regulation was required (Holmes & Pizzagalli, 2007; Saunders & Jentzsch, 2014). It is also the case that the region implicated in the evaluation of emotional content, the rostral ACC, is implicated in the regulation of

post-error adjustments (Pizzagalli, Peccoralo, Davidson, & Cohen, 2006). These puzzle pieces motivated the investigation documented in Paper 6. This paper is reviewed below, and is then presented in full to conclude this chapter.

Paper 6 Overview

In Section 1 of this thesis I documented our first contribution to the understanding of post-error behaviour. We explored post-error slowing in a novel cognitive game paradigm, *the Buckets game*, which allowed us to examine whether or not post-error slowing generalised beyond rapid choice experiments. Here we use two of the same measurement techniques to explore the post-error adjustments of depressed and non-depressed students, for emotional and non-emotional words, when undertaking a classic emotional Stroop task.

In doing so we built upon the work of Compton and colleagues (2008), who employed a variant of the emotional Stroop task to investigate post-error behaviour in undergraduates assessed for depression symptoms using the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Unfortunately, emotional stimuli and neutral stimuli were not considered independently in Compton's study, making the results difficult to interpret. The experiment also featured performance feedback after each trial, which may give rise to post-error slowing independently of erroneous performance (de Bruin, Mars, & Hulstijn, 2004; Saunders & Jentzsch, 2012). We remedied these issues in two experiments by using an emotional Stroop task with no feedback on performance, and by considering and analysing emotional and non-emotional content separately. Considering emotional and non-emotional words separately allowed us to examine the unique impact of emotional stimuli on reactive control – the specific type of cognitive control

implicated in the post-hoc cognitive adjustments required to adapt behaviour in response to an unplanned or unexpected challenge.

Summary and Transition

In Paper 6 our results were staggering and consistent. This consistency was found between the robust and traditional measurement calculations of post-error adjustments, as well as between Experiment 1 and Experiment 2. We documented a clear and debilitating effect of errors on participants with depression symptoms as compared to controls – and we also documented a clear difference in this effect for emotional and neutral stimuli. Following errors on neutral stimuli, roughly double the amount of slowing was observed for participants with depression symptoms relative to controls, and this slowing was not compensated by an increase in accuracy. Following errors on emotional stimuli, no slowing was observed for participants with depression symptoms but a substantial decrease in accuracy (~9%) was observed.

In our discussion we concluded that our data supported an account of post-error slowing that was commensurate with the cognitive control account – even though we did not find an increase in accuracy associated with an error. We suggested post-error slowing may buffer against a decrease in accuracy that otherwise might result from a processing disturbance associated with an error (Gehring et al., 1993). We suggested that when an error is registered in awareness, the reactive control system is recruited (Nieuwenhuis et al., 2001), and responding on the following trial is slowed to allow for the disturbance in processing associated with the error. However, in the case where an error was made but not registered in awareness we would not expect the reactive control system to be not recruited, and therefore we would expect no post-error slowing and a substantial decrease in accuracy.

Importantly, this account does not necessitate nor exclude the small increase in accuracy sometimes associated (e.g., Botvinick et al., 2001; Dutilh et al., 2012; Laming, 1979) with post-error slowing.

With regards to depression, we documented severe impairments in reactive control. We demonstrated that when exposed to emotional content, participants with depression symptoms did not slow following an error and rather showed a substantial decrease in accuracy. In other words, if emotionally primed, those with depression symptoms showed a complete failure to adjust their behaviour in response to the environment. This finding was commensurate with previous work that indicated reactive cognitive control is impaired for depression when emotional regulation is required (Saunders & Jentzsch, 2014).

In future work, I hope to address my hypotheses that the inconsistency typically found in depression studies of cognitive control may be the result of internally generated emotional priming (automatic thoughts and rumination) that is common in depression. It is possible that the presence of randomly interspersed emotional stimuli in our experiment regulated this otherwise inconsistent deficit (Paulus, 2015). In any event, Paper 6 suggests that when depression symptoms are high, adaptive and goal-driven behaviours might prove extremely difficult to maintain in the face of perceived mistakes or negative feedback. Our data suggest that even mild emotional exposure may lead to severe impairment in executive function and behavioural regulation for depressed individuals. This is a crucial clinical understanding that can only able be confirmed experimentally via sequential effects. Section 2 of this thesis provides a powerful case for further proliferation of sequential effects research. For participants with high levels of depression symptoms, both

emotional content *and* success or failure critically affects performance in the immediate future.

Cognitive dysfunction under emotional exposure: When participants with depression symptoms show no cognitive control.

Journal:	*Psychological Science*
Manuscript ID	Draft
Manuscript Type:	Research article
Date Submitted by the Author:	n/a
Complete List of Authors:	Williams, Paul; University of Newcastle, Psychology Howard, Zachary ; University of Newcastle, Psychology Ross, Rachel; University of Newcastle, Psychology Eidels, Ami; University of Newcastle, School Psychology
Keywords:	Cognitive Control, Depression, Post-error slowing, Emotional Control, Emotional Stroop

Title:

Cognitive dysfunction under emotional exposure: When participants with depression symptoms show no cognitive control.

Authors: Paul Williams[a*b], Zachary Howard[a], Rachel Ross[a], Ami Eidels[a]

Running Head: Cognitive control under emotional exposure

Affiliations:

[a] School of Psychology, University of Newcastle, Callaghan, Australia

[b] The READ Clinic, Psychological Services Centre, Erina, Australia

*Correspondence to Paul Williams, Email: paul.williams.psyc@gmail.com

Abstract:

Adaptive human behavior requires cognitive control - the monitoring of actions and performance, to regulate and coordinate ongoing behaviour. Major depression is associated with neuropsychological differences in cognitive control, however behavioural experiments have failed to consistently reflect this. We resolve this ambiguity and show that in the emotional Stroop task, depression symptoms are linked to severe deficits in cognitive control following errors. For emotional content, major depression symptoms were associated with a failure to instigate behavioural adjustments following errors, leading to reduced performance. For non-emotional content, we found major depression symptoms were associated with substantial adjustments following errors, mitigating reduced performance. These findings suggest that under emotional priming, major depression is marked by a complete failure to adapt behaviour in response to relevant environmental feedback. This work has implications for interpreting prior and future scientific findings, and may also inform clinical applications for depression treatment.

Keywords:

Depression, Cognitive Control, Post Error Slowing, Emotional Stroop, Reactive Cognitive Control, Post Error Adjustments

Introduction

Major Depressive Disorder (depression) accounted for 4.3% of disability adjusted life years worldwide in 2004, and is expected to be the leading cause of disability adjusted life years by 2030 (Mathers and Ma Fat, 2008). Cognitive theories of depression have spawned well-supported psychotherapeutic treatments (e.g., Beck, Rush, Shaw, & Emery, 1979) that target the core symptoms of the disorder – emotional dysregulation, cognitive or information processing deficits, and behavioural deficits. To improve treatment outcomes it is imperative to better understand the relationship between these core symptoms (Gotlib & Joorman, 2010). Here we clarify these relationships by documenting the impact of emotional priming on *cognitive control* - the ability to regulate information processing and maintain goal-directed behaviour under varying environmental demands.

Cognitive control has been conveniently dichotomised as proactive or reactive (Braver, 2012). *Proactive control* describes the preparatory cognitive adjustments required to adapt behaviour successfully for a known environment. For example, it is common to minimise emotional behaviour (e.g., fear, crying) in professional settings as compared to other settings. *Reactive control* describes the post-hoc cognitive adjustments required to adapt behaviour in response to an unplanned or unexpected challenge. For example, additional and specific adjustments are sometimes required to minimise emotional behaviour when an unplanned event occurs (e.g., conflict, a demotion). While several authors have suggested deficits may exist in both proactive and reactive control for participants with depression symptoms (Holmes & Pizzagalli, 2008; West, Choi, & Travers, 2010), recent experimental evidence hints that depression might be associated with a specific deficit in reactive control when

emotional regulation is required (Holmes & Pizzagalli, 2007; Saunders & Jentzsch, 2014). This evidence concurs with anecdotal reports that the increase in emotional dysregulation (e.g., irritability, agitation, crying) typically associated with depression can be masked in some environments or for short periods.

Two long-standing behavioural markers of reactive control are post-error adjustments of both response time and accuracy. Errors result from interaction with the environment, are typically unplanned, and can be used to guide corrective behaviour (Ridderinkhof, van den Wildenberg, Segalowitz, & Carter, 2004). In cognitive tasks participants typically slow following errors (Laming, 1979; Rabbitt; 1966), an adjustment commonly argued to aide performance, in terms of accuracy, on subsequent attempts (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Laming, 1979; Ridderinkhof et al., 2004). In high accuracy tasks this *post-error slowing* is considered ubiquitous and is considered a benchmark of reactive cognitive control (Botvinick et al., 2011)..

Depression has been associated with deficits in reactive control when emotional regulation is required (Holmes & Pizzagalli, 2007; Saunders & Jentzsch, 2014). This is commensurate with the well-established association between depression symptoms and the neurophysiological markers of errors and error awareness; the error-related negativity (ERN) (Olvet & Hajcak, 2008; Ridderinkhof et al., 2004; West et al., 2010) and error-positivity (P_E) (Holmes & Pizzagalli, 2010). Pizzagalli, Peccoralo, Davidson, and Cohen (2006) note neurophysiological studies have also documented structural (Ballmaier et al., 2004), neurochemical (Auer et al., 2000), and functional (Beauregard et al., 1998; George et al., 1997; Kumari et al., 2003) differences for depressed populations in the anterior cingulate cortex, a region central to the network implicated in error monitoring and correction. Therefore, errors

in cognitive tasks seemingly provide an ideal platform to study reactive cognitive control. Surprisingly though, differences in post-error adjustments have been described as inconsistent or non-specific for depressed populations (Paulus, 2015; Pizzagalli et al., 2006; Saunders & Jentzsch, 2014).

Previously, Compton and colleagues (2008) explored the effect of emotional stimuli on post-error slowing in a population with depression symptoms, measured using the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Participants were asked to indicate the number of words presented on a monitor. Up to four emotionally charged words, or emotionally neutral words, were presented on each trial. As depression symptoms increased, post-error slowing increased and post-error accuracy decreased. Unfortunately, these findings are difficult to meaningfully interpret because post-error adjustments were averaged across emotional and neutral trials. Any effect of emotional stimuli on post-error slowing could have been contaminated by neutral stimuli. The experiment also featured performance feedback after each trial, which may give rise to post-error slowing independently of erroneous performance (de Bruin, Mars, & Hulstijn, 2004; Saunders & Jentzsch, 2012).

We remedied these issues in two experiments by using an emotional Stroop task (Williams, Mathews, & MacLeod, 1996) with no feedback on performance, and by considering emotional and non-emotional content separately. Participants were assessed for depression symptoms using the Beck Depression Inventory®–II (Beck, Steer, & Brown, 1996). We suspected that our results would highlight a specific reactive cognitive control deficit following emotionally-valenced stimuli, only for those with depression symptoms. Our findings demonstrate, for the first time, that

differences in the behavioural markers of reactive cognitive control are consistently associated with depression symptoms.

Experiment 1

Method

Participants

Thirty-three undergraduate psychology students (24 Females) volunteered through an online experimental management system. Subject ages ranged from 18-43 (M =24.6, SD = 8.0). Participants had normal or corrected to normal vision, English as first language, and were compensated with course credits.

Stimuli and Apparatus

Stimuli were presented using Python™ which also recorded response times to the 1ms. Button responses were made on a Cedrus response pad and button locations were counterbalanced across participants.

Table 1 shows the complete stimulus set of 24 words, six positive emotional words (e.g., BETTER, GLAD), six negative emotional words (BITTER, SAD) and 12 uncategorised non-emotional words (BUTTER, PAD). Non-emotional and emotional words were matched on frequency, length, and were orthographic neighbours where possible (BITTER, BUTTER). Items were piloted at an earlier stage to ensure accurate categorisation into emotionally positive, emotionally negative and non-emotional conditions. Post-experiment, participants completed a word identification questionnaire to further ensure word conditions were correctly categorised, revealing 98% correctly identified, with no participants excluded for poor accuracy. Words were printed in red or green color (RGB values of 220, 0, 0, and 0, 170, 0, for red and green, respectively). Words appeared in uppercase Arial font, bold and size 30.

Table 1

List of stimulus items for Experiments 1 and 2

Non-emotional	Positive emotional	Negative emotional
ADDRESSED	IMPRESSED	DEPRESSED
BUTTER	BETTER	BITTER
FERRY	MERRY	MISERY
LATELY	LOVELY	LONELY
PAD	GLAD	SAD
TENURE	PLEASURE	FAILURE

Procedure

The experiment was conducted in a quiet, dimly lit room with air-conditioning. Participants sat 60cm from the 17" monitor so stimuli occupied a visual angle of up to 4.77 degrees. Participants were given task instructions followed by 24 demonstration trials in which correct responses were automatically displayed on the screen. This was followed by 24 practice trials with participant responses and feedback, and then 24 practice trials without feedback. Data were collected in the subsequent eight experimental blocks with a forced 1 min break between blocks. On each trial, a single word was presented in the centre of a white background. To avoid adaptation and minimise reliance on local cues we introduced a trial-to-trial spatial uncertainty of up to 40 pixels around the target location. Participants were asked to identify the color of the word while refraining from reading the word, by pressing either the red- or green-marked button. On each trial, presentation of a fixation cross for 500ms was followed by a blank screen for 500ms, then followed by the stimulus for a maximum of 4000ms. The presentations of stimuli were response terminated.

The words used are listed in Table 1 and each appeared four times within a block (i.e., 96 trials per block), with order of presentation randomised. At the end of the experiment participants indicated how they responded to each word during the experiment: non-emotional, emotionally positive or emotionally negative. Participants overwhelmingly concurred with our emotional categorisation of items (98% agreement). Participants then completed the Beck Depression Inventory®–II (Beck et al., 1996) which has demonstrated strong validity and reliability in nonclinical and clinical populations (Beck, Steer, Ball, Ranieri, 1996; Sharp & Lipsky, 2002).

Results

Depression (BDI-II) scores.

Scores on the BDI-II ranged from 0 to 44 (M = 10.09, SD = 9.71). Consistent with Compton (Compton et al., 2008), participants scoring ≥ 20 on the BDI-II (n = 6) were categorized as displaying depression symptoms, and participants scoring ≤ 12 (n = 22) were categorized as controls.

Response time and Post Error Slowing

All analyses were carried out on individual participant's data before mean scores for emotional and non-emotional words were collated and analysed across participants for the variables of interest. We note accuracy, response time (RT), and post-error slowing showed no difference for positive and negative stimuli - hence these data were collapsed to form the stimuli category *Emotional Stimuli.*

For Experiment 1 mean RT across all participants ranged from 353 to 677 ms (M = 478 ms, SD = 67 ms) and showed no relationship to BDI-II (rho = .11, p = .537). Mean accuracy ranged from 78.9 - 99.5% (M = 94.8%, SD = 5.1%) and also showed no relationship to BDI-II (rho = .14, p = .438). Paired sample t-tests revealed no differences across participants for neutral (M = 468 ms; = M = 95.1%) versus

emotional (*M* = 465 ms; M=94.5%) words for either accuracy, $t(32) = .76$, p = .451, or RT $t(32) = 1.63$, $p = .114$. These RT results confirmed there was no *emotional Stroop effect* (Williams et al., 1996) observed in our data, which is defined as slower responding for emotional words relative to neutral words on the current trial. The lack of an emotional Stroop effect is a common finding for intermixed presentation designs (Phaf & Kan, 2007).

Post-error adjustments were calculated using what we term the *traditional* method. The traditional method involves subtracting the mean RT of a participant's post-error trials from the mean RT of their post-correct trials. Similarly for accuracy, it subtracts the conditional probability of a hit preceded by a hit from the conditional probability of a hit preceded by a miss. We calculated post-error adjustments separately for neutral and emotional stimuli. That is, the mean RT of trials following errors on emotional stimuli was subtracted from mean RT of trials following correct responses on emotional stimuli, and likewise for neutral stimuli. These results are presented in Figure 1A, separately for depressed and control groups. Of paramount interest was whether participants with depression symptoms ($n = 6$) showed the same pattern of post-error slowing as controls ($n = 21$; one control subject made no errors on neutral stimuli). Figure 1A highlights participants with depression symptoms showed sizeable post-error slowing following neutral words (*M* = 129.9ms) but very little post-error slowing following emotional words (*M* = 24.8ms). In contrast, controls showed consistent post-error slowing for both neutral (*M* = 54.0ms) and emotional (*M* = 78.3ms) words. A mixed model analysis of variance (ANOVA) confirmed the interaction between depression category and word valence was significant, $F(1,25) = 4.61$, $p = .042$. There was no main effect of depression category or word valence [$F(1,25) = 1.80$, $p = .192$; $F(1,25) = .067$, $p = .738$].

Figure 1B depicts the marginally reliable negative relationship between the difference in post-error slowing for emotional and neutral words and BDI-II scores (*rho* = -.34, *p* = .065).

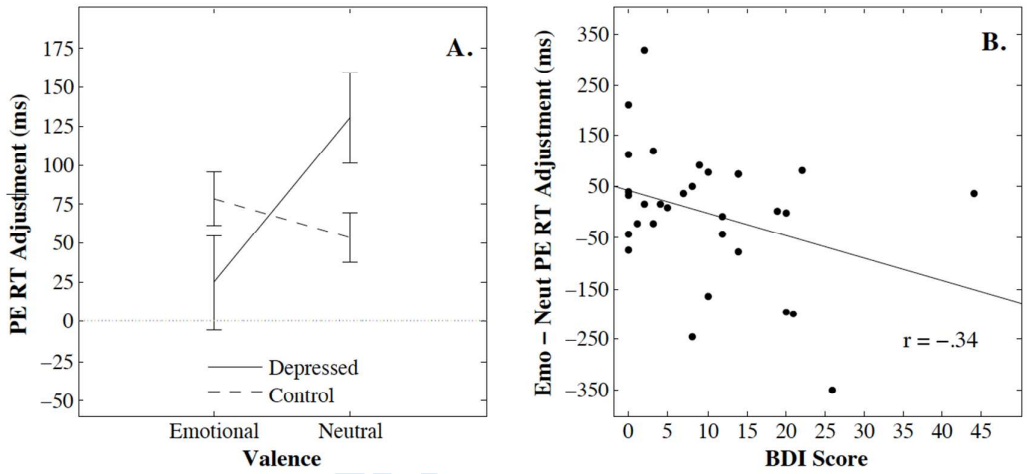


Figure 1. Panel A shows post-error adjustment in response time (PE RT) for emotional and neutral words for depressed and control participants. Positive values indicate post-error slowing. Error bars indicate the within subject corrected standard error of the mean (Morey, 2008). Panel B shows emotional minus neutral post-error RT adjustments by BDI score. The small number of depressed participants is reflected in Panel A by the greater variability of the solid line.

Interim Discussion

The stimuli type of emotional or neutral substantially impacted cognitive control in participants with depression symptoms, but not controls. For participants with depression symptoms, post-error slowing - the behavioural signature of reactive cognitive control - was absent for errors made on emotional words. Furthermore, the difference between post-error slowing on neutral words and emotional words was associated with the degree of depression symptoms across the full range of BDI-II scores, suggesting the degree of cognitive control impairment was associated with the degree of depression symptoms.

The results of Experiment 1 were thus encouraging, but rested on only a few participants with depression symptoms and a marginally reliable correlation. There was also an uneven spread of errors across participants and categories of word valence. For neutral stimuli, two participants made no errors, and a further nine made fewer than five errors. For emotional stimuli thirteen participants made fewer than five errors. Because of this lack of errors, we were unable to meaningfully calculate post-error accuracy adjustments and compare various methods of measuring post-error slowing (see Williams, Heathcote, Nesbitt, & Eidels, 2016). Notably, the traditional method for calculating post-error slowing in Experiment 1 cannot differentiate post-error slowing from long-term effects like fatigue, distraction, or boredom (Dutilh et al., 2012a), which may be more prevalent in populations with depression (Paelecke-Habermann, Pohl, & Leplow, 2005; Veiel, 1997). One solution is to pair post-error trials with immediately preceding pre-error counterparts that are also post-correct trials. Pairwise differences are then calculated (i.e., post-error RT minus [pre-error + post-correct] RT), with the mean of the differences providing a *robust* measure of post-error RT adjustments (Dutilh et al., 2012a). The same type of pairing may be employed to calculate post-error accuracy adjustments (Dutilh, Forstmann, Vandekerckhove, & Wagenmakers, 2013).

We designed a second experiment (Experiment 2) in which we expected participants to make more errors. This allowed us to calculate both the traditional and robust measures for post-error slowing.

Experiment 2

Method

Participants

Thirty-seven undergraduate psychology students (29 Females) volunteered through an online experimental management system. Subject ages ranged from 18-48 (M =23.3, SD = 6.1). In an effort to recruit a greater number of participants with depression symptoms, our recruitment poster indicated a preference for very happy or very sad volunteers. Exclusion criteria were as per Experiment 1.

Stimuli and apparatus

The word set was the same as Experiment 1 (see Table 1 again for the complete list). In Experiment 2, each word could be presented with and without italic letters. When italicised, a single letter in the string -- either at the beginning, middle, or end of each word stimuli -- would be italic. Participants were instructed to respond differently depending on whether the word contained an italicised letter. Thus, the task required scanning of the letters but did *not* require processing of the emotional content. Each word was presented with and without italics an equal number of times for each participant.

Procedures

Participants were given task instructions followed by 12 example trials with automated responses. This was followed by 20 practice trials with participant participation and feedback and then 20 practice trials without feedback. Data were collected in the subsequent eight experimental blocks. The words listed in Table 1 were presented three times per block (i.e., 72 trials per block). Participants were asked to identify the color of the word and whether or not it contained an italic letter. The four response options were “italic red”, “italic green”, “non-italic red” and “non-italic green”. The procedure was otherwise the same as Experiment 1.

Results

Accuracy, response time (RT), and post-error slowing again showed no difference for positive and negative stimuli, thus we again combined these data to form an *Emotional Content* condition. Scores on the BDI-II ranged from 0 to 40 (M = 12.95, SD = 10.47). Participants with depression symptoms (n = 9) and controls (n = 24) were classified as per Experiment 1.

Mean RT ranged from 684ms to 1,587ms (M = 998ms, SD = 211ms) and showed no relationship to BDI-II (*rho* = .17, p = .315). Accuracy ranged from 85.2 - 98.9% (M = 94.0%, SD = 3.2%) and showed a marginal tendency to increase as BDI-II scores increased (*rho* = .31, p = .064). The spread of errors across participants and categories of word valence was much more favourable. For neutral stimuli, just one subject made fewer than five errors. For emotional stimuli, just two participants made fewer than five errors. This allowed us to calculate post-error accuracy adjustments and employ the robust method. A paired sample t-test revealed no difference in accuracy for neutral (M=93.6%) versus emotional (M = 94.3%) words, $t(36)$ = .98, p = .334. However, a similar t-test revealed a positive emotional Stroop effect where RTs were slower for emotional (M = 979ms) than neutral (M = 959ms) words, $t(36)$ = 3.57, p = .001. This emotional Stroop effect showed no relationship to BDI-II (*rho* = -.24, p = .15).

Figure 2, Panels A and B indicate Experiment 2 produced a near perfect replication of the pattern of post-error RT adjustments found in Experiment 1. We use the subscripts T and R to refer to the Traditional and Robust methods, respectively. Participants with depression symptoms showed substantial post-error slowing (M_T = 231.2ms; M_R = 221.7ms) for neutral words and no post-error slowing for emotional words (M_T= 26.9ms; M_R = 7.7ms), whereas controls showed consistent post-error slowing for both neutral (M_T= 80.0ms; M_R = 96.1ms) and emotional (M_T= 65.8ms;

M_R = 99.5ms) words. A mixed-model ANOVA confirmed a significant interaction between depression group and word valence on post-error slowing [$F_T(1,30) = 10.88$, p=.003; $F_R(1,30) = 6.45$, p=.017], and a significant main effect of valence [$F_T(1,30) = 14.38$, p=.001; $F_R(1,30) = 6.05$, p=.020], but a non-significant main effect of depression group [$F_T(1,30) = 2.28$, p = .141; $F_R(1,30) = .13$, p = .721]. Figure 2, Panels C and D show the post-error accuracy adjustments, and confirm the inability of depression participants to implement cognitive control following errors on emotional stimuli. The depression group members were not less accurate following errors on neutral stimuli when they slowed (M_T = 0.2%; M_R = 4.0%), but were far less accurate when they did not slow following errors on emotional stimuli (M_T = -8.9%; M_R = -8.7%). Controls showed no significant decrease or increase in accuracy following errors on neutral (M_T = -1.2%; M_R = 2.9%) or emotional (M_T = -1.3%; M_R = -1.9%) stimuli. Mixed model ANOVAs showed marginally significant interactions between depression group and valence on accuracy [$F_T(1,30) = 3.84$, p=.059; $F_R(1,30) = 2.67$, p=.112], and a marginal and significant main effect of valence [$F_T(1,30) = 4.05$, p=.053; $F_R(1,30) = 13.29$, p=.001]. The main effects of depression group were non-significant and marginal [$F_T(1,30) = 1.75$, p = .195; $F_R(1,30) = 3.99$, p = .055], respectively.

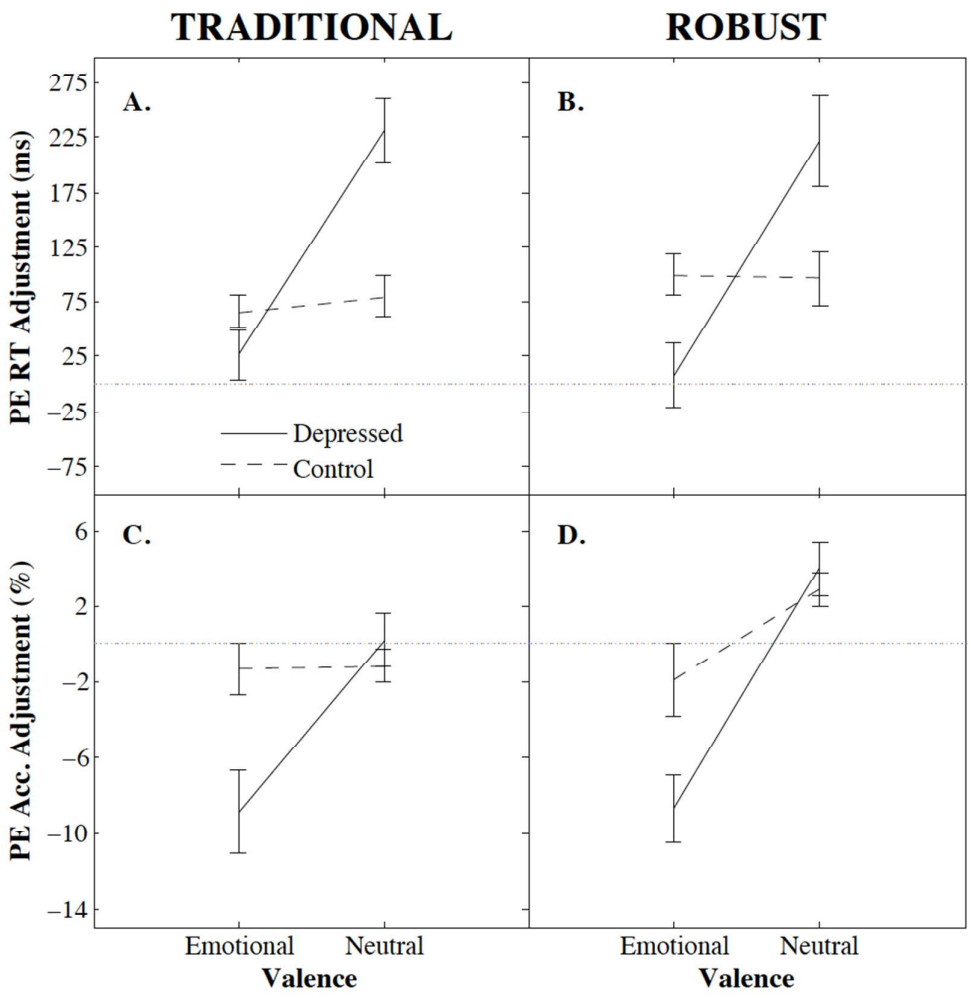


Figure 2. Panel A shows post-error adjustment in RT (PE RT) for emotional and neutral words for depressed and control participants. Positive values indicate post-error slowing. Panel B shows the RT adjustment calculated via the robust method. Panel C shows post-error accuracy (PE Acc.) adjustment for emotional and neutral words for depressed and control participants. Panel D shows the accuracy adjustment calculated via the robust method. Error bars indicate the within subject corrected standard error of the mean (Morey, 2008).

Simple effects tests confirmed the depression symptom group showed lower post-error accuracy following errors on emotional words when compared to the control group, [$T_T(30) = 1.94$, $p = .062$; $T_R(30) = 2.25$, $p = .032$]. Similar tests confirmed that, for neutral stimuli, post-error slowing was greater for the depression symptom group

than controls, [$T_T(30) = 2.90$, $p = .007$; $T_R(30) = 1.74$, $p = .093$], but there was no difference in post-error accuracy for the depression symptom group and controls, [$T_T(30) = .564$, $p = .577$; $T_R(30) = .415$, $p = .681$].

Figure 3, Panels A and B depicts the negative relationship between the difference in post-error slowing for emotional and neutral words and BDI-II scores for all participants for both the traditional ($r = -.51$, $p = .001$) and robust ($r = -.36$, $p = .028$) calculation methods. This completes the replication of Experiment 1 results and indicates that for the traditional measure the difference in post-error slowing for emotional words relative to neutral words accounts for 25% of the variation in depression symptoms.

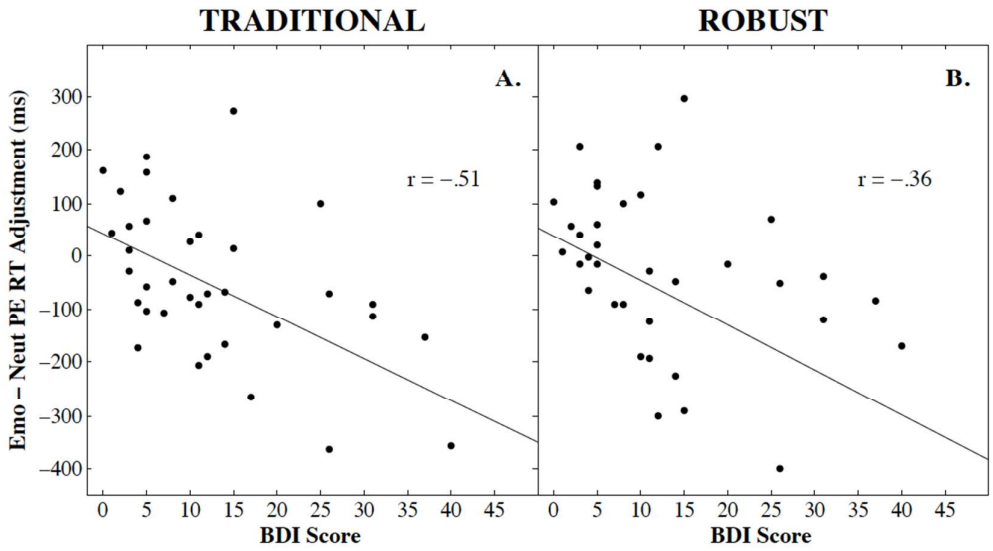


Figure 3. The relationship between the difference in post-error RT adjustments for emotional and neutral stimuli and BDI-II Score.

General Discussion

Post-error slowing is a benchmark effect of reactive cognitive control. We aimed to explore this effect in depression and controls with emotional and non-

emotional priming. Our results indicated that for all participants, errors to non-emotional stimuli impaired subsequent trial performance by slowing responding without a compensatory increase to accuracy. This effect held for both depressed and control participants, although the magnitude of post-error slowing was greater for participants with depression symptoms. Strikingly, we also documented a specific deficit in cognitive control for participants with depression symptoms when exposed to task-irrelevant emotional content. Participants with depression symptoms failed to slow down after errors to emotionally-valenced stimuli, resulting in a substantial reduction in post-error accuracy (~9%) compared to controls (who exhibited the expected post-error slowing).

The results suggest that depression symptoms might be linked with specific deficits in reactive cognitive control that are both qualitatively and quantitatively distinct for emotional and non-emotional stimuli. Our results, therefore, are able to reconcile previously ambiguous evidence for behavioral deficits in depressed participants. Participants with depression symptoms differentially exhibit both substantial increases and decreases in post-error slowing corresponding to specific experimental conditions. Averaging across emotional and non-emotional stimuli (e.g., Compton et al., 2008) would combine these two opposing effects and provide inconsistent results.

The failure for participants with depression symptoms to engage cognitive control following errors on emotional stimuli is commensurate with previous work that indicates cognitive control impairment in depression when emotional regulation is required (Saunders & Jentzsch, 2014). It is also in line with previous work linking error detection and correction to the rostral anterior cingulate cortex (ACC), the brain region associated with evaluating the emotional significance of events (Pizzagalli,

Peccoralo, Davidson, & Cohen, 2006). Studies in perception (Rinck & Becker, 2005; Surguladze et al., 2004), memory (Mathews and MacLeod, 2005), and attention (Eizenman et al., 2003; Joorman, 2004) have also shown that depression is associated with a specific impairment in the ability to disengage from emotional content (Gotlib & Joorman, 2010). Notably, activation of the emotional region of ACC via task irrelevant emotional content has also been implicated in the impairment of proactive cognitive control (Wyble, Sharma, & Bowman, 2008). Within the context of this prior work, our results suggest that for participants with depression symptoms, task irrelevant emotional content may have led to disruption of normal function for the ACC. Combined with an inability to disengage from the emotional content, participants with depression symptoms may have been left them unaware of some or all of their errors, and so reactive cognitive control failed to engage (this account is unpacked further below, by contrast with regular cognitive control function). As a result of this failure, we suspect that the disturbance in processing associated with errors was not corrected. Lower accuracy was therefore observed on subsequent trials. We note the degree of this impairment was strongly associated with the degree of depression symptoms across the full range of BDI-II scores.

The clarity and consistency of our results, across experiments and also methods of measurement, helpfully informs our understanding of both the regular and irregular function of reactive cognitive control. Consider that when post-error slowing was observed – for controls on all stimuli, and for participants with depression symptoms on neutral stimuli - we found no increase or decrease in accuracy following errors. Yet for participants with depression symptoms on emotional stimuli, we found no post-error slowing and a large decrease in accuracy. This pattern of results might best be accommodated by a perspective that assumes post-error slowing buffers

against a decrease in accuracy that might otherwise result from the processing disturbance associated with an error.

Under this framework, errors are associated with a disturbance in processing (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993), and awareness of these errors is required to fully recruit the cognitive control system (Endrass, Reuter, & Kathmann, 2007; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001; Wessel, Danielmeier, Ullsperger, 2011). An examples of such a disturbance in processing might include a lapse in attention, resulting in decreased sensory sensitivity (Purcell & Kiani, 2016). When an error is registered in awareness – as is typical in regular function - the reactive control system is recruited, and response speed is slowed on the following trial (Nieuwenhuis et al., 2001), possibly due to an increase in caution (Dutihl et al., 2012b; Purcell & Kiani, 2016). Given our results, this caution is likely associated with the diversion of resources to the reactive cognitive control system, and the re-orientation of the subject to the task. Such a resolution is broadly consistent with recent approaches that view post-error slowing as a heterogeneous effect (e.g., Ullsperger & Danielmeier, 2016). We stress that this account does not necessitate nor exclude the small increases (e.g., Botvinick et al., 2001; Dutilh et al., 2012b; Laming, 1979) or decreases (e.g., Notebaert et al., 2009) in accuracy sometimes associated with post-error slowing.

When an error is not registered in awareness – as we suspect for participants with symptoms of depression for emotional stimuli - the lapse associated with an error may carry over to the next trial. Critically, this account predicts no slowing and a substantial decrease in accuracy in cases where the cognitive control system is not recruited following errors. This was precisely the result we observed for participants with depression symptoms on emotional stimuli. As we noted above, in depression,

we suspect task irrelevant emotional content may have led to disruption of normal function for the ACC, and combined with an inability to disengage from the emotional content, these participants may have been left unaware of some or all of their errors.

In conclusion, Experiments 1 and 2 provided two major insights. Firstly, for participants with depression symptoms, consistent and severe impairments in reactive control were identified. When these participants were primed with task irrelevant emotional content, there was a total failure in reactive cognitive control. These data suggest that depression symptoms are associated with a severe impairment in behavioural regulation in the face of even mild emotional exposure. In the case of a major depressive episode - where task-irrelevant emotional thoughts are frequent, automatic, and ruminated upon – an inability to monitor environmental feedback may severely disrupt adaptive and goal-driven behaviour. This substantially informs our understanding of the relationship between emotional dysregulation, cognitive or information processing deficits, and behavioural deficits, in depression. Exploring the boundary conditions of this effect may also help determine when and how depressed individuals are able to adapt their behaviour appropriately in response to the environment. Future studies may also wish to determine the boundaries of this effect with regards to depression in the absence of anxiety and other symptoms, given we used non-controlled samples typical of those presenting for therapy. Notably, our results also helpfully constrain current models of corrective behaviour, and provide the impetus for testable theoretical speculation that will drive further experimentation.

References

Auer, D. P., Pütz, B., Kraft, E., Lipinski, B., Schill, J., & Holsboer, F. (2000). Reduced glutamate in the anterior cingulate cortex in depression: an in vivo proton magnetic resonance spectroscopy study. Biological psychiatry, 47(4), 305-313.

Ballmaier, M., Toga, A. W., Blanton, R. E., Sowell, E. R., Lavretsky, H., Peterson, J., . . . Kumar, A. (2004). Anterior cingulate, gyrus rectus, and orbitofrontal abnormalities in elderly depressed patients: an MRI-based parcellation of the prefrontal cortex. American Journal of Psychiatry, 161(1), 99-108.

Beauregard, M., Leroux, J.-M., Bergman, S., Arzoumanian, Y., Beaudoin, G., Bourgouin, P., & Stip, E. (1998). The functional neuroanatomy of major depression: an fMRI study using an emotional activation paradigm. Neuroreport, 9(14), 3253-3258.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). Cognitive Therapy of Depression. New York: Guildford.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. Arch Gen Psychiatry, 4, 561-571.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychol Rev, 108(3), 624-652.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn Sci, 16(2), 106-113. doi: 10.1016/j.tics.2011.12.010

Compton, R. J., Lin, M., Vargas, G., Carp, J., Fineman, S. L., & Quandt, L. C. (2008). Error detection and posterror behavior in depressed undergraduates. Emotion, 8(1), 58-67. doi: 10.1037/1528-3542.8.1.58

de Bruin, E., Mars, B., & Hulstijn, W. (2004). It wasn't me... or was it? How false feedback affects performance. In M. Ullsperger & M. Falkenstein (Eds.), Errors, conflicts, and the brain. Current opinions on performance monitoring (pp. 118-124). Leipzig: MPI of Cognitive Neuroscience.

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E. J. (2013). A diffusion model account of age differences in posterror slowing. Psychol Aging, 28(1), 64-76. doi: 10.1037/a0029875

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012a). How to measure post-error slowing: A confound and a simple solution. Journal of Mathematical Psychology, 56(3), 208-216. doi: 10.1016/j.jmp.2012.04.001

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012b). Testing theories of post-error slowing. Atten Percept Psychophys, 74(2), 454-465. doi: 10.3758/s13414-011-0243-2

Eizenman, M., Yu, L. H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., & Levitan, R. D. (2003). A naturalistic visual scanning approach to assess selective attention in major depressive disorder. Psychiatry Res, 118(2), 117-128.

Endrass, T., Reuter, B., & Kathmann, N. (2007). ERP correlates of conscious error recognition: aware and unaware errors in an antisaccade task. European Journal of Neuroscience, 26(6), 1714-1720. doi: 10.1111/j.1460-9568.2007.05785.x

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. Psychological Science, 4(6), 385-390. doi: 10.1111/j.1467-9280.1993.tb00586.x

George, M. S., Wassermann, E. M., Kimbrell, T. A., Little, J. T., Williams, W. E., Danielson, A. L., . . . Post, R. M. (1997). Mood improvement following daily left prefrontal repetitive transcranial magnetic stimulation in patients with depression: a placebo-controlled crossover trial. American Journal of Psychiatry.

Gotlib, I. H., & Joormann, J. (2010). Cognition and Depression: Current Status and Future Directions. Annual review of clinical psychology, 6, 285-312. doi: 10.1146/annurev.clinpsy.121208.131305

Holmes, A. J., & Pizzagalli, D. A. (2007). Task feedback effects on conflict monitoring and executive control: relationship to subclinical measures of depression. Emotion, 7(1), 68-76. doi: 10.1037/1528-3542.7.1.68

Holmes, A. J., & Pizzagalli, D. A. (2008). Spatiotemporal dynamics of error processing dysfunctions in major depressive disorder. Archives of General Psychiatry, 65(2), 179-188. doi: 10.1001/archgenpsychiatry.2007.19

Holmes, A. J., & Pizzagalli, D. A. (2010). Effects of task-relevant incentives on the electrophysiological correlates of error processing in major depressive disorder. Cognitive, Affective, & Behavioral Neuroscience, 10(1), 119-128. doi: 10.3758/CABN.10.1.119

Joormann, J. (2004). Attentional bias in dysphoria: The role of inhibitory processes. Cognition and Emotion, 18(1), 125-147. doi: 10.1080/02699930244000480

Kumari, V., Mitterschiffthaler, M. T., Teasdale, J. D., Malhi, G. S., Brown, R. G., Giampietro, V., . . . Williams, S. C. (2003). Neural abnormalities during

cognitive generation of affect in treatment-resistant depression. Biological psychiatry, 54(8), 777-791.

Laming, D. (1979). Choice reaction performance following an error. Acta Psychologica, 43, 199-224.

Mathers, Ma Fat,. The Global Burden of Disease 2004 update. (2008): World Health Organisation.

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. Annu Rev Clin Psychol, 1, 167-195. doi: 10.1146/annurev.clinpsy.1.102803.143916

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). reason, 4(2), 61-64.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology, 38(5), 752-760.

Olvet, D. M., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. Clinical Psychology Review, 28(8), 1343-1354. doi: http://dx.doi.org/10.1016/j.cpr.2008.07.003

Paelecke-Habermann, Y., Pohl, J., & Leplow, B. (2005). Attention and executive functions in remitted major depression patients. Journal of Affective Disorders, 89(1–3), 125-135. doi: http://dx.doi.org/10.1016/j.jad.2005.09.006

Paulus, M. P. (2015). Cognitive control in depression and anxiety: out of control? Current Opinion in Behavioral Sciences, 1, 113-120. doi: http://dx.doi.org/10.1016/j.cobeha.2014.12.003

Pizzagalli, D. A., Peccoralo, L. A., Davidson, R. J., & Cohen, J. D. (2006). Resting anterior cingulate activity and abnormal responses to errors in subjects with

elevated depressive symptoms: a 128-channel EEG study. Hum Brain Mapp, 27(3), 185-201. doi: 10.1002/hbm.20172

Purcell, B. A., & Kiani, R. (2016). Neural mechanisms of post-error adjustments of decision policy in parietal cortex. Neuron, 89(3), 658-671.

Rabbitt, P. (1966). Error and error correction in choice-response tasks. Journal of Experimental Psychology, 71(2), 264-272.

Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. Brain Cogn, 56(2), 129-140. doi: 10.1016/j.bandc.2004.09.016

Rinck, M., & Becker, E. S. (2005). A comparison of attentional biases and memory biases in women with social phobia and major depression. J Abnorm Psychol, 114(1), 62-74. doi: 10.1037/0021-843x.114.1.62

Saunders, B., & Jentzsch, I. (2012). False external feedback modulates posterror slowing and the f-P300: implications for theories of posterror adjustment. Psychon Bull Rev, 19(6), 1210-1216. doi: 10.3758/s13423-012-0314-y

Saunders, B., & Jentzsch, I. (2014). Reactive and proactive control adjustments under increased depressive symptoms: insights from the classic and emotional-face Stroop task. Q J Exp Psychol (Hove), 67(5), 884-898. doi: 10.1080/17470218.2013.836235

Surguladze, S. A., Young, A. W., Senior, C., Brebion, G., Travis, M. J., & Phillips, M. L. (2004). Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. Neuropsychology, 18(2), 212-218. doi: 10.1037/0894-4105.18.2.212

Ullsperger, M., & Danielmeier, C. (2016). Reducing Speed and Sight: How Adaptive Is Post-Error Slowing? Neuron, 89(3), 430-432.

Veiel, H. O. (1997). A preliminary profile of neuropsychological deficits associated with major depression. J Clin Exp Neuropsychol, 19(4), 587-603. doi: 10.1080/01688639708403745

Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. Journal of Cognitive Neuroscience, 23(10), 3021-3036.

West, R., Choi, P., & Travers, S. (2010). The influence of negative affect on the neural correlates of cognitive control. Int J Psychophysiol, 76(2), 107-117. doi: 10.1016/j.ijpsycho.2010.03.002

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. Psychol Bull, 120(1), 3-24.

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. Judgement and Decision Making, 11(2), 174-184.

Wyble, B., Sharma, D., & Bowman, H. (2008). Strategic regulation of cognitive control by emotional salience: A neural network model. Cognition and Emotion, 22(6), 1019-1051.

General Discussion

The six chapters of this thesis have explored the impact of the recent past on human performance. The contributions have focused on the impact of two broad categories of sequential effects: (1) whether or not the previous attempt was successful or unsuccessful, and (2) whether a previous stimulus carried emotional content or no emotional content. The six chapters form a coherent whole, however, Section 1 and Section 2 had clearly delineated goals. In Section 1, I documented the development and application of a cognitive game designed to explore normal cognition and extend our understanding of The Hot Hand and Post-error slowing. In Section 2, I documented the use of a well-established paradigm, the emotional Stroop task, to extend our understanding of the impact of emotional content on cognitive processing, for normal and depressed populations. The impact of emotional content was considered on its own in Chapter 5, and in interaction with errors in Chapter 6.

The use of the same measurement techniques throughout this thesis provided a unique opportunity to compare post-error slowing across two differing paradigms. On the one hand, the buckets game was temporally extended (maximum trial length of 8s) and offered a rich risk and reward game play platform through which the conscious trade-off between speed and accuracy was encouraged. . On the other hand, the emotional Stroop offered a classic fast-paced and forced-choice cognitive experiment. The data presented in Section 1 and Section 2 therefore provide a unique opportunity to compare post-error adjustments goal oriented and temporally extended tasks versus rapid choice tasks. Our data in the Buckets game (Chapter 3), a goal oriented and temporally extended task, suggested global contributors accounted for almost half of the post-error adjustments documented by the traditional method. In contrast, our data from the emotional Stroop task (Chapter 6), a rapid choice task, highlighted little or no influence of global contributors such as fatigue or boredom. In

this case the post-error adjustments as measured by the robust and traditional methods were extremely consistent. This thesis then, when taken as a whole, is the first to document this empirical point of difference. Our data suggest global effects such as boredom may influence temporally extended tasks more than fast-paced tasks. This lends further weight to speculations the post-error behaviour in rapid tasks may not be representative of those found in more complex and temporally extended tasks.

The chapters of this thesis also represented modular contributions. In Chapter I motivated my exploration of sequential effects, and provided a novel example of paradigm development in which both computer game design and experimental design principles were utilised. We required our paradigm to be heavily informed by gaming principles to meet the demands of hot hand research, and also to meet the strict experimental specifications necessary for rigorous and scientific study. Ultimately, we combined the principles of game design and experimental design to develop a top-down alien shooter where players were asked to shoot down as many alien spaceships as possible within a fixed amount of time.

Chapter 2 further highlighted the value of our rigorous piloting process. We exposed and explored the failure of the top-down alien shooter to meet a critical design benchmark. We then documented the development of a second game, the Bucket's game, which met our benchmark requirements. In the Buckets game, we also strategically altered the design so that the difficulty measure was precise and near continuous (10^{th} of a second). This alteration brought the appraisal of the hot hand into line with other contemporary decision-making research, and therefore allowed the parallels between the hot hand and post-error slowing to be illuminated.

Chapter 3 documented the application of our cognitive game paradigm. We simultaneously explored the hot hand and post-error slowing using the Buckets game.

The work made several novel contributions, including outlining the theoretical and empirical links that supported simultaneous exploration of the hot hand belief and post-error slowing. Our results were illuminating. In regards to the hot hand belief, we documented a hot hand effect larger than any previous research we are aware of. This finding hinted the hot hand effect might be prevalent in contexts when motivation is lower, which would explain the resilience of the hot hand belief in the face on contradictory evidence at the professional level. In regards to post-error slowing, we reinforced the importance of consideration motivation, which had been heavily considered on other areas of cognitive control, but not post-error adjustments. We also provided the first empirical support for speculation that there may be substantial differences between post-error behaviour in rapid-choice experimental tasks, as compared to goal driven and temporally extended tasks.

Chapter 4 began the shift toward clinically oriented research, and away from an exploration of normal cognition. We chose a well-established paradigm – the emotional Stroop task - to explore clinically oriented sequential effects. To motivate this work and develop the link to sequential effects as explored in Section 1, I demonstrated that (1) errors and negative feedback may influence the cognitive performance of those with depression, (2) the brain region implicated in the evaluation of emotional content is also implicated in the regulation of post-error adjustments, and (3) sequential effects resulting from emotional stimuli, particularly those that might result from an interaction with errors, were largely unexplored. In line with these goals, Chapter 4 provided a methodological review and best practice guideline for the implementation of the emotional Stroop task.

In Chapter 5 I documented our first exploration of sequential effects in the emotional Stroop task with a clinically oriented sample. This work was part of a

larger study in which we tested whether the processing of emotional stimuli was obligatory, non-obligatory, or task dependent. With regards to sequential effects, we documented a reliable, fully randomised experimental and statistical methodology for partitioning fast and slow emotional Stroop (ESE) effects. We also showed that the slow ESE did not generalize to a non-standard emotional Stroop design, and therefore might not be as robust as had been assumed.

In Chapter 6, I documented an exploration of the relationship between sequential effects caused by the emotional content of stimuli, and those caused by errors. After motivating this work, I outlined two experiments that employed a classic emotional Stroop task for participants measured for symptoms of depression. The impact of errors for emotional and non-emotional content was considered separately. We documented a clear and debilitating effect of errors on participants with depression symptoms, and a clear difference in this effect for emotional and non-emotional stimuli. When exposed to emotional content, participants with depression symptoms did not slow following an error and rather showed a substantial decrease in accuracy. When exposed to non-emotional content, participants with depression symptoms slowed roughly twice as much as controls. In other words, if emotionally primed, those with depression symptoms showed a complete failure to adjust their behaviour in response to the environment. We used these findings to draw conclusions about the nature of depression, and also about the nature of the cognitive control system.

The above work, as a whole, provided several avenues for further study. In Section 1 our findings suggested future work might consider whether low motivation may help explain other findings of post-error speeding, particularly in studies where the overall success rate is low. Our findings also suggested that future work might

explore the prevalence of the hot hand in amateur contexts, where motivation is possibly lower than in the professional sporting contexts that provide the typical bedrock for hot-hand research. . In Section 2, I provided a powerful case for further proliferation of sequential effects research. For participants with high levels of depression symptoms, both emotional content *and* success or failure critically affected future performance. This result has great potential in shaping future theorising and experimentation.

References

Adams, E. (2010). Fundamentals of Game Design (2nd ed.). Berkely, CA, USA: New Riders.

Adams, R. M. (1995). Momentum in the performance of professional tournament pocket billiards players. International Journal of Sport Psychology, 26(4), 580-587.

Albright, S. C. (1993). A Statistical Analysis of Hitting Streaks in Baseball. Journal of the American Statistical Association, 88(424), 1175-1183.

Auer, D. P., Pütz, B., Kraft, E., Lipinski, B., Schill, J., & Holsboer, F. (2000). Reduced glutamate in the anterior cingulate cortex in depression: an in vivo proton magnetic resonance spectroscopy study. Biological psychiatry, 47(4), 305-313.

Avugos, S., Köppen, J., Czienskowski, U., Raab, M., & Bar-Eli, M. (2013). The "hot hand" reconsidered: A meta-analytic approach. Psychology of Sport and Exercise, 14(1), 21-27.

Ballmaier, M., Toga, A. W., Blanton, R. E., Sowell, E. R., Lavretsky, H., Peterson, J., . . . Kumar, A. (2004). Anterior cingulate, gyrus rectus, and orbitofrontal abnormalities in elderly depressed patients: an MRI-based parcellation of the prefrontal cortex. American Journal of Psychiatry, 161(1), 99-108.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. Psychology of Sport and Exercise, 7(6), 525-553.

Beats, B., Sahakian, B. J., & Levy, R. (1996). Cognitive performance in tests sensitive to frontal lobe dysfunction in the elderly depressed. Psychological medicine, 26(03), 591-603.

Beauregard, M., Leroux, J.-M., Bergman, S., Arzoumanian, Y., Beaudoin, G., Bourgouin, P., & Stip, E. (1998). The functional neuroanatomy of major depression: an fMRI study using an emotional activation paradigm. Neuroreport, 9(14), 3253-3258.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. Arch Gen Psychiatry, 4, 561-571.

Bocskosky, A., Ezekowitz, J., & Stein, C. (2014). The hot hand: A new approach to an old "fallacy". Paper presented at the MIT Sloan Sports Analytics Conference, Boston.

Bogte, H., Flamma, B., van der Meere, J., & van Engeland, H. (2007). Post-error adaptation in adults with high functioning autism. Neuropsychologia, 45(8), 1707-1714.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. Annu Rev Psychol, 66, 83-113.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychol Rev, 108(3), 624-652.

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

Brown, S. D., Marley, A., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. Psychological review, 115(2), 396.

Camerer, C. F. (1989). Does the Basketball Market Believe in the `Hot Hand,'? The American Economic Review, 79(5), 1257-1261.

Cane, J. E., Sharma, D., & Albery, I. P. (2009). The addiction Stroop task: examining the fast and slow effects of smoking and marijuana-related cues. Journal of Psychopharmacology, 23(5), 510-519.

Carter, C. S., MacDonald III, A. W., Ross, L. L., & Stenger, V. A. (2001). Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. American Journal of Psychiatry, 158(9), 1423-1428.

Clark, R. D., III. (2003a). Streakiness among professional golfers: fact or fiction? International Journal of Sport Psychology, 34(1), 63-79.

Clark, R. D., III. (2003b). An analysis of streaky performance on the LPGA Tour. Percept Mot Skills, 97(2), 365-370.

Clark, R. D., III. (2005). Examination of hole-to-hole streakiness on the PGA Tour. Percept Mot Skills, 100(3 Pt 1), 806-814.

Compton, R. J., Lin, M., Vargas, G., Carp, J., Fineman, S. L., & Quandt, L. C. (2008). Error detection and posterror behavior in depressed undergraduates. Emotion, 8(1), 58-67.

Croson, R., & Sundali, J. (2005). The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos. Journal of Risk and Uncertainty, 30(3), 195-209.

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. Front Psychol, 2(233), 1-9.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Putnam, K. (2002). Depression: perspectives from affective neuroscience. Annu Rev Psychol, 53(1), 545-574.

de Bruijn, E. R., Grootens, K. P., Verkes, R. J., Buchholz, V., Hummelen, J. W., & Hulstijn, W. (2006). Neural correlates of impulsive responding in borderline personality disorder: ERP evidence for reduced action monitoring. Journal of psychiatric research, 40(5), 428-437.

De Bruijn, E. R., Mars, R. B., & Hulstijn, W. (2004). It wasn't me... or was it? How false feedback affects performance. Errors, conflicts, and the brain. Current opinions on performance monitoring, 118-124.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. Psychological Science, 5(5), 303-305.

Dorsey-Palmateer, R., & Smith, G. (2004). Bowlers' Hot Hands. The American Statistician, 58(1), 38-45.

Dudschig, C., & Jentzsch, I. (2009). Speeding before and slowing after errors: is it all just strategy? Brain Res, 1296, 56-62. doi: 10.1016/j.brainres.2009.08.009

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E. J. (2013). A diffusion model account of age differences in posterror slowing. Psychol Aging, 28(1), 64-76.

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012a). How to measure post-error slowing: A confound and a simple solution. Journal of Mathematical Psychology, 56(3), 208-216.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012b). Testing theories of post-error slowing. Atten Percept Psychophys, 74(2), 454-465.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. Psychon Bull Rev, 17(6), 763-771.

Elliott, R., Sahakian, B., Herrod, J., Robbins, T., & Paykel, E. (1997). Abnormal response to negative feedback in unipolar depression: evidence for a diagnosis specific impairment. Journal of Neurology, Neurosurgery & Psychiatry, 63(1), 74-82.

Elliott, R., Sahakian, B., McKay, A., Herrod, J., Robbins, T., & Paykel, E. (1996). Neuropsychological impairments in unipolar depression: the influence of perceived failure on subsequent performance. Psychological medicine, 26(05), 975-989.

Erčulj, F., & Štrumbelj, E. (2015). Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PLoS ONE, 10(6), e0128885.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47(6), 381-391.

Frings, C., Englert, J., Wentura, D., & Bermeitinger, C. (2010). Decomposing the emotional Stroop effect. The quarterly journal of experimental psychology, 63(1), 42-49.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. Psychological Science, 4(6), 385-390.

George, M. S., Wassermann, E. M., Kimbrell, T. A., Little, J. T., Williams, W. E., Danielson, A. L., . . . Post, R. M. (1997). Mood improvement following daily left prefrontal repetitive transcranial magnetic stimulation in patients with depression: a placebo-controlled crossover trial. American Journal of Psychiatry.

Gilden, D. L., & Wilson, S. G. (1995). Streaks in skilled performance. Psychon Bull Rev, 2(2), 260-265.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17(3), 295-314.

Gotlib, I. H., & Joormann, J. (2010). Cognition and Depression: Current Status and Future Directions. Annual review of clinical psychology, 6, 285-312.

Hajcak, G., McDonald, N., & Simons, R. F. (2003). Anxiety and error-related brain activity. Biol Psychol, 64(1), 77-90.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: evidence for generalized cognitive search processes. Psychol Sci, 19(8), 802-808.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24(1), 1-55.

Holle, C., Neely, J. H., & Heimberg, R. G. (1997). The effects of blocked versus random presentation and semantic relatedness of stimulus words on response to a modified Stroop task among social phobics. Cognitive Therapy and Research, 21(6), 681-697.

Holmes, A. J., & Pizzagalli, D. A. (2007). Task feedback effects on conflict monitoring and executive control: relationship to subclinical measures of depression. Emotion, 7(1), 68-76.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev, 109(4), 679-709.

Huber, J., Kirchler, M., & Stöckl, T. (2010). The hot hand belief and the gambler's fallacy in investment decisions under risk. Theory and Decision, 68(4), 445-462.

Kerns, J. G., Cohen, J. D., MacDonald III, A. W., Johnson, M. K., Stenger, V. A., Aizenstein, H., & Carter, C. S. (2005). Decreased conflict-and error-related activity in the anterior cingulate cortex in subjects with schizophrenia. American Journal of Psychiatry, 162(10), 1833-1839.

Klassen, F. J. G. M., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. . Journal of the American Statistical Association, 96, 500-509.

Koehler, J. J., & Conley, C. A. (2003). The "hot hand" myth in professional basketball. Journal of Sport & Exercise Psychology, 25(2), 253-259.

Kumari, V., Mitterschiffthaler, M. T., Teasdale, J. D., Malhi, G. S., Brown, R. G., Giampietro, V., . . . Williams, S. C. (2003). Neural abnormalities during cognitive generation of affect in treatment-resistant depression. Biological psychiatry, 54(8), 777-791.

Laming, D. (1968). Information theory of choice-reaction times. New York: Academic Press.

Laming, D. (1979a). Choice reaction performance following an error. Acta Psychologica, 43, 199-224.

Laming, D. (1979b). Autocorrelation of choice-reaction times. Acta Psychologica, 43, 381-412.

Larkey, P. D., Smith, R. A., & Kadane, J. D. (1989). It's okay to believe in the "hot hand". Chance, 2, 22-30.

Luce, R. D. (1986). Response Times: Their Role in Inferring Elementary Mental Organization. . 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(5), 397-408.

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. Annu Rev Clin Psychol, 1, 167-195.

McKenna, F. P., & Sharma, D. (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: the role of fast and slow components. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(2), 382.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual review of neuroscience, 24(1), 167-202.

Murphy, F., Michael, A., Robbins, T., & Sahakian, B. (2003). Neuropsychological impairment in patients with major depressive disorder: the effects of feedback on task performance. Psychological medicine, 33(03), 455-467.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology, 38(5), 752-760.

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. Cognition, 111(2), 275-279.

Offerman, T., & Sonnemans, J. (2004). What's Causing Overreaction? An Experimental Investigation of Recency and the Hot-Hand Effect. The Scandinavian Journal of Economics, 106(3), 533-553.

Paulus, M. P. (2015). Cognitive control in depression and anxiety: out of control? Current Opinion in Behavioral Sciences, 1, 113-120.

Phaf, R. H., & Kan, K.-J. (2007). The automaticity of emotional Stroop: A meta-analysis. Journal of Behavior Therapy and Experimental Psychiatry, 38(2), 184-199.

Pizzagalli, D. A., Peccoralo, L. A., Davidson, R. J., & Cohen, J. D. (2006). Resting anterior cingulate activity and abnormal responses to errors in subjects with elevated depressive symptoms: A 128- channel EEG study. Hum Brain Mapp, 27(3), 185-201.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bourma & D. Bouwhuis (Eds.), Attention and performance X (pp. 531-556). Hillsdale, NJ: Erlbaum.

Proudfit, G. H., Inzlicht, M., & Mennin, D. S. (2013). Anxiety and error monitoring: the importance of motivation and emotion.

Rabbitt, P. (1966). Error and error correction in choice-response tasks. Journal of Experimental Psychology, 71(2), 264-272.

Rabbitt, P. (1969). Psychological refractory delay and response-stimulus interval duration in serial, choice-response tasks. Acta Psychologica, 30(0), 195-219.

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. British Journal of Psychology, 70(2), 305-311.

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? an analysis of response programming. Quarterly Journal of Experimental Psychology, 29(4), 727-743.

Rabbitt, P., & Vyas, S. (1981). Processing a display even after you make a response to it. How perceptual errors can be corrected. The Quarterly Journal of Experimental Psychology Section A, 33(3), 223-239.

Rabbitt, P. M. A., & Vyas, S. M. (1970). An elementary preliminary taxonomy for some errors in laboratory choice RT tasks. Acta Psychologica, 33(0), 56-76.

Rabin, M. (2002). Inference by Believers in the Law of Small Numbers. The Quarterly Journal of Economics, 117(3), 775-816.

Rao, J. M. (2009). Experts' perceptions of autocorrelation: The hot hand fallacy among professional basketball players. http://www.justinmrao.com/playersbeliefs.pdf. UC San Diego.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological review, 85(2), 59.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. Psychological Science, 9(5), 347-356.

Rosenberg, D. R., Mirza, Y., Russell, A., Tang, J., Smith, J. M., Banerjee, S. P., . . . Boyd, C. (2004). Reduced anterior cingulate glutamatergic concentrations in childhood OCD and major depression versus healthy controls. Journal of the American Academy of Child & Adolescent Psychiatry, 43(9), 1146-1153.

Saunders, B., & Jentzsch, I. (2012). False external feedback modulates posterror slowing and the f-P300: implications for theories of posterror adjustment. Psychon Bull Rev, 19(6), 1210-1216.

Saunders, B., & Jentzsch, I. (2014). Reactive and proactive control adjustments under increased depressive symptoms: insights from the classic and emotional-face Stroop task. Q J Exp Psychol (Hove), 67(5), 884-898.

Siwoff, S., Hirdt, S., & Hirdt, P. (1987). The 1987 Elias baseball analyst. New York: Collier Books.

Smith, G. (2003). Horseshoe pitchers' hot hands. Psychon Bull Rev, 10(3), 753-758.

Stewart, N., Brown, G. D., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(1), 3.

Thakkar, K. N., Polli, F. E., Joseph, R. M., Tuch, D. S., Hadjikhani, N., Barton, J. J., & Manoach, D. S. (2008). Response monitoring, repetitive behaviour and anterior cingulate abnormalities in autism spectrum disorders (ASD). Brain, 131(9), 2464-2478.

Tversky, A., & Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. Chance, 2(4), 31-34.

van Meel, C. S., Heslenfeld, D. J., Oosterlaan, J., & Sergeant, J. A. (2007). Adaptive control deficits in attention-deficit/hyperactivity disorder (ADHD): the role of error processing. Psychiatry Res, 151(3), 211-220.

Walsh, P. D. (1996). Area-restricted Search and the Scale Dependence of Path Quality Discrimination. Journal of Theoretical Biology, 183(4), 351-361.

Waters, A. J., Sayette, M. A., Franken, I. H., & Schwartz, J. E. (2005). Generalizability of carry-over effects in the emotional Stroop task. Behaviour research and therapy, 43(6), 715-732.

Wilde, G. J. S., Gerzke, D., & Paulozza, L. (1998). Risk optimization training and transfer. Transportation Research Part F: Traffic Psychology and Behaviour, 1(1), 77-93.

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. Psychol Bull, 120(1), 3-24.

Wyble, B., Sharma, D., & Bowman, H. (2005). Modelling the slow emotional Stroop effect: Suppression of cognitive control. Progress In Neural Processing, 16, 291.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev, 111(4), 931-959.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. Phil. Trans. R. Soc. B, 367(1594), 1310-1321.