Characterising goal neglect by investigating the effects of complexity and task structure

Francesca Biondo

This dissertation is submitted for the degree of

Doctor of Philosophy

Lucy Cavendish College

University of Cambridge

October 2017

ABSTRACT

A fundamental question of human existence is how much control we have on our behaviour. This dissertation aims to add to our understanding of cognitive control by characterising how a particular failure of performance, Goal Neglect (GN), is affected by different forms of complexity manipulations. In Chapter 2, I develop a new task to test GN and unlike previous studies, I manipulate complexity qualitatively by altering the instructional cues - the cues instructing the participant to shift to a different rule set. GN was sensitive to this kind of complexity manipulation and this is linked to a failure in recognizing the significance of the instructional cues. In Chapter 3, I propose a new entropy-like measure to quantify the temporal clustering of GN and use this to test the differential temporal patterns that are predicted by two theoretical models of GN. The results suggest that both models are likely to be operant, but with their relative dominance being different across time: GN early on in the task appears to be mostly driven by failures which are "task model" like, whilst GN which manifests later on is better aligned with the "monitoring" account. Chapter 2 also revealed that GN can be sensitive to manipulations of complexity *during* task performance, which motivated the question of whether previously published studies suggesting the contrary, were perhaps due to insufficient complexity. Hence, in Chapter 4, using the new GN task, I investigate this further. Overall, the results were mixed and indicated that complexity does not appear to affect GN unless the complexity manipulation is more closely associated to the critical event. Throughout this dissertation, I refer to models and empirical evidence from the Prospective Memory (PM) literature given the apparent similarity between PM and GN experimental paradigms. In Chapter 5, I take this further and investigate how PM failures and GN are different, if at all, with the broader aim to integrate what are otherwise isolated domains. I found a mixture of null findings which suggest that it is not entirely clear if GN and PMf reflect different capacities. Nonetheless, while investigating the differences between GN and PMf, a much more interesting question emerged with respect to what structural features of a task predict different signatures of GN-like and PMf-like errors. The key finding to this theory-neutral approach was a general rule about task structure: a combination of extended practice and low frequency of critical events predict both a larger amount of errors and with more of these occurring late in the task. Overall, this research has shed further light on task conditions that may result in different error signatures and that may reflect different cognitive resources.

ACKNOWLEDGEMENTS

To all my yoga teachers, for the energy and focus that the practice has given me. And all of the friends at the Hughes-Lucy Cavendish badminton club.

To Charlotte Eades, for her courage and the charity setup in her memory - Charlotte's Bag (www.charlottesbag.com) which aims to raise funds for research into glioblastoma multiforme.

Finally and most importantly, thanks to my wonderful parents and my siblings, Claudia and Andrea, for taking good care of me, always. This PhD would not have been possible without their loving support.

TABLE OF CONTENTS

LIST OF FIGURES

10

LIST OF TABLES

11

ABBREVIATIONS

ANOVA	Analysis Of Variance
ATC	Actual Task Complexity
ATCv1	Actual Task Complexity version 1
ATCv2	Actual Task Complexity version 2
BOLD	Blood-Oxygen-Level Dependent
EF	Executive Functions
fMRI	functional Magnetic Resonance Imaging
GN	Goal Neglect
IC	Instructional Complexity
IET	Inter-Event Time
LMT	Letter Monitoring Task
LMTv2	Letter Monitoring Task version 2
MD	Multiple Demand
MSE	Mean Side-Error
PFC	Prefrontal Cortex
PM	Prospective Memory
PMf	Prospective Memory failures
RT	Reaction Time
SGNT	Semantic Goal Neglect Task
WCT	Word Categorization Task

Chapter 1. INTRODUCTION

A fundamental question of human existence is how much control we have over our behaviour. Not surprisingly, a central topic in cognitive neuroscience is "cognitive control" - the ability to flexibly adapt behaviour to reach goals. One approach to understand the mechanisms underlying cognitive control is to examine how goal-directed behaviour breaks down when facing a complex task. However, complexity is an ill-defined concept sometimes leading to counterintuitive effects. For example, additional complexity may be followed by unaltered or even improved performance (Maylor, 1993, 1996; Bhandari and Duncan, 2014). The objective of this dissertation is to advance mechanistic accounts of cognitive control and to do so I focus on the specific error of Goal Neglect (GN) as a function of different forms of complexity manipulations, including changes in task structure.

This introductory chapter provides a framework of the basic issues and concepts that motivate the experiments presented in this dissertation. The literature review in this introduction is not exhaustive and some topics are dealt more fully in later chapters. This first chapter is divided into five main sections. The first discusses the problem of cognitive control. The second focuses on how studying errors using goal-directed tasks provides crucial information on the way cognitive control works. I then introduce the specific error of GN. In the fourth section, I discuss various aspects of complexity and capacity limitations. Finally, I provide a preface to the chapters which follow.

Chapter 1 | Introduction

1.1 COGNITIVE CONTROL – WHY STUDY IT?

In nature, one way that organisms survive and reproduce is to go through a series of programmed behaviours that are activated based on specific stimuli in the environment. At first, this may seem to imply that such behaviours are simplistic, however, on the contrary these innate releasing mechanisms (IRMs, Lorenz, 1970 as cited in Duncan, 2010) may be quite complex as is illustrated in the mating behaviour of sticklebacks described by Duncan (2010a):

At the start of the mating season, the male-stickle-back turns red, stakes out a territory, and builds a nest consisting of a hole covered with weeds. The stage is set for a complex mating sequence, driven by a concatenated series of male and female IRMs. The first male IRM is triggered by the sight of a female stickleback, with swollen belly and a specific, posturing movement, entering the territory. The male approaches and begins a characteristic zigzag dance. Now the first IRM of the female comes into play; seeing the zigzag, she approaches the male. Her approach drives the next male IRM – he turns and swims rapidly toward the nest; seeing him turn, the female is enticed to follow. As the female is seen to approach the nest, the male responds by pointing his head to the opening; the female responds by entering. At the sight of the female in the nest, the male beings to stimulate spawning; he repeatedly thrusts his head at her rump, and in response, the eggs are laid. Finally, the male detects fresh eggs in the next and, in response, releases his sperm…. Each step in this sequence is somewhat separate from the others; it is the approach of the female to the nest that releases the male's head point; it is the head point that releases the female's entry. The separate IRMs form the elements of the fishes' behaviour; in combination; they create a complex whole. (p. 4)

Despite a certain degree of complexity, such sequences of behaviour are inherently limited in as much as they are contingent to specific events in the environment. For example, if the male zigzag dance fails to occur, this could halt the mating sequence and lead to a failure to reproduce. A first level of improvement over this limitation is receptivity to a larger array of inputs or less stringent thresholds, as long as these events are sufficiently good indicators of the required conditions – for example, movements approximating a zigzag dance could be sufficient if these reflected equally good reproductive fitness. However, a far more radical solution to maximize achieving a

desired goal is the ability to *create* the optimal conditions required for the target behaviour out of an environment which does not readily offer them. In other words, learning novel programmes (behaviours, thoughts) to obtain goals (reproduce) overrides the necessity to go through a series of fixed and limited programmes and allows the organism to deal flexibly with an intrinsically chaotic environment.

In fact, we could imagine this process to cycle iteratively through two main stages. First, new programmes would be learnt. Second, these learnt problem-solving sequences could be stored and re-used when needed by using feed-forward control (if x then y) constrained by goals. Situations may arise in which none of the stored programmes are adequate to solve the problem at hand, which would then require returning to the first stage. The second stage is assumed to be relatively faster, automated and less laborious than the first stage which instead is slower and largely error-driven. Learning a new programme requires formulating the problem space, setting local goals, attempting actions to fulfil these goals, monitoring the outcome and updating the problem space or model of the problem (Reason, 2013). In the cognitive sciences, these two stages are closely related to the distinction between "controlled" and "automatic" behaviour (Schneider and Shiffrin, 1977). In Schneider's and Shiffrin's paper, automated behaviour is one in which the sequence of behaviour is always activated in response to the input (if x then y) and does not require "active control or attention". For example, the sight of a comb may trigger a "combing" action without much effort. Nonetheless, automated behaviour may often be accompanied by controlled behaviour since other inputs may trigger competing behaviours (if x and z, then $\sim y$) - for example, a comb (x) on a doctor's desk (z) should result in suppressing combing one's own or, the doctor's hair ($\sim y$)! An inability to modulate automated behaviours is typically observed in patients with lesions to their prefrontal cortex (PFC), a brain region thought to be important for cognitive control (Bianchi, 1922; Luria, 1966; Desimone and Duncan, 1995; Miller and Cohen, 2001). In summary, we not only need cognitive control to create new programmes but we also need it to modulate the deployment of stored, automatic programmes.

In the cognitive neuroscience literature, the concept of "cognitive control" is broadly interchangeable with the term "executive functions" (EF). A slight difference does however exist, with cognitive control referring to a more abstract phenomenon and more general processes such as, reconfiguration, biasing and on-line maintenance of contextual

information (Desimone and Duncan, 1995; Botvinick et al., 2001). Instead, EF is relatively less generic and, although it still allows for unified views, its focus is modular, by referring to explicit separable models of control or sub-functions such as, "inhibition", "updating", "shifting/switching" and others (Miyake et al., 2000; Friedman and Miyake, 2017). Overall, both terms are not defined in a consistent manner (Astle and Scerif, 2009), however the core feature is that they refer to complex processes which support and optimize goal-directed behaviour in spite of prepotent alternatives (Baddeley, 1986). To avoid confusion, throughout this dissertation I will use one term, cognitive control, with the understanding that it grossly overlaps with the concept of EF.

The concept of cognitive control raises one important issue – if cognitive control is organized in a modular way then, this might imply the absurd existence of some "director" who orchestrates the running of these different modules. This is also referred to as the "homunculus" (little man) problem. So, the key question is, what are the mechanisms which allow cognitive control to deal with the chaos of the external (and internal) environment? One widely used approach to attempt answering this question is studying errors, which I describe in more detail in the next section.

1.2   ERRORS

What is an error? This can be defined as a case when a planned sequence of mental or physical activities fails to obtain the intended goal without the intrusion of some chance event (p.10, Reason, 2013). An error has at least two modifiers. The first is the goal. Often one may have multiple conflicting goals such that the same behaviour classifies as correct depending on one goal (e.g. goal 1: reduce commute time to work in the city centre; behaviour: rent expensive flat in city centre) and an error according to another co-existing goal (e.g. goal 2: save money for a holiday). The second modifier is the observer: the same action can classify as correct from the point of view of one agent but an error according to another agent/system (e.g. establishing traffic precedence for two vehicles at an unsigned crossroad).

The first basic step to study how a system works is to systematically observe it. Based on these observations, one can subsequently propose an initial explanation of the functions operating between the input and the output in the system. This explanation, or model

(some representation of the world), can then be tested by making predictions (hypotheses) of the output based on a given input. Any deviation from the expected output in the test results, provides information that the model is either incorrect and needs to be adjusted or, the deviation might simply be due to chance/test error. If we think of an error as a deviation from an expected result, then in a similar way, errors inform us about the cognitive mechanisms which intervene between setting up a goal and achieving it. For example, the task of preparing a cup of tea requires the control of a sequence of sub-actions: find cup, add teabag, add boiling water, add sugar, add milk, etc... Chaining models suggest that each sub-action is dependent on the previous one. However, we often observe errors that involve a failure in one step (e.g. forget to add sugar) without this failure affecting the following step (add milk) (as cited in Henson, 1996; lecture by Badre, 2015). Hence, this evidence does not support chaining models and alternate theories are favoured (e.g. hierarchical models). This example also illustrates how our intuitions (initial models) of how the mind works are not necessarily the way the mind is actually functioning.

If a complex system (e.g. the mind) is made up of different mechanisms, then each one of these mechanisms is likely (although not necessarily) to give rise to different kinds of errors or pattern of errors. In practice, the process of studying human errors allows us to make inferences about mechanisms in precisely the opposite direction − qualitatively different errors are assumed to indicate the existence of different sub-functions in the system. The way in which errors can be fractionated into categories is manifold, debatable and easily confusing. A conceptual distinction is offered by James Reason who proposes two classes of errors "slips/lapses" versus "mistakes" (2013). Slips are when the plan for action is appropriate, but the actions do not go as planned, whilst mistakes are when the plan to achieve the goal are inadequate hence, even if actions go as planned it is likely to result in failure (Reason, 2013). Although it is easy to imagine the distinction of these errors from one's own personal experiences e.g. forgetting to lock the car (slip) vs. failing to find a destination because of an outdated map (mistake), it is not entirely clear if this distinction actually exists in the mind.

Another perspective on the categorization of errors can be drawn from the theory of control proposed by Norman and Shallice (1986; as cited in Ward, 2015). These authors suggested that sets of tasks referred to as "schemas" are activated or suppressed according

to a biasing mechanism depending on the current goal. This biasing mechanism arises from a balance between i) bottom-up processes driven by cues in the environment and automated procedural behaviours (e.g. habits) and, ii) top-down processes (e.g. setting up plans and long-term goals, task instructions, decision-making). Hence, failure of these two distinct processes would lead to different errors. Although not identical, this differentiation shares some similarities to Reason's (2013) slips-vs.-mistakes categorization: slips are similar to failures of bottom-up processes given that the latter tend to include an appropriate action plan/intention; mistakes are similar to impairments in top-down processes because the focus here is getting the plan right. Evidence to support the bottom-up vs. top-down distinction should come in the form of a double dissociation so that interfering with process (i) would lead to increased number of bottom-up errors and leave process (ii) unaltered (no change in the number of top-down errors) whilst interfering with process (ii) would lead to the reverse pattern, an increase in top-down errors but no change in the number of bottom-up errors. To the best of my knowledge this precise type of double dissociation has not been reported however, there is evidence for a single dissociation. For example, neuropsychological studies have shown that damage to the prefrontal cortex (thought to be important for biasing mechanisms) leads to a marked increase of top-down errors like socially inappropriate behaviour (e.g. undress in public). This kind of error may reflect an impaired top-down representation of the social context (e.g. public space) which fails to override an intact automated schema of undressing behaviour (e.g. triggered by warm weather). This kind of inappropriate behaviour is not infrequent in patients with damage to the prefrontal cortex (Bianchi, 1922; Luria, 1966; Shallice and Burgess, 1991).

Another distinction of errors is made in terms of failures of either "task-setting" or "monitoring". Task setting refers to the ability to generate a solution to a relatively open-ended problem (Ward, 2015). Monitoring refers to the ability to observe and check one's own behaviour and assess if it is aligned to the intended one. Neuropsychological studies have indicated a hemispheric specialization of the prefrontal cortex of these two sub-functions. For example, one study by Stuss *et al.* (2000) used a modified version of the Wisconsin Card Sorting Test (WCST) and showed that patients with left prefrontal cortex lesions were worst on one aspect of the task related to task setting, whilst right prefrontal cortex patients were more impaired on another component related to monitoring. In the

WCST a series of cards must be sorted into categories following a certain rule, for example yellow cards must be stacked on yellow cards, red on red and so on. On each trial, participants are told if their response is correct. After a few trials, the participant is told they are incorrect which should prompt a change in sorting technique even though this is not an explicit instruction. Generally, participants with damage to the frontal lobe show impairments with this task by persevering with the incorrect rule despite the negative feedback (Demakis, 2003). In the modified version of the WCST, patients were informed of the different sorting rules at the start and explicitly told what the first sorting rule was and when this would change. Stuss and colleagues (2000) found that patients with left- compared to patients with right-prefrontal lesions were more impaired on the original version of the task, which heavily depended on the ability to generate new sorting rules (task setting). Instead, the flip pattern was observed for the modified WCST, right-prefrontal patients showed more errors suggesting a heightened inability to monitor their behaviour compared to left-prefrontal patients.

1.3 A FRACTIONATED ACCOUNT OF COGNITIVE CONTROL AND THE PROBLEM OF TERMINOLOGY

Besides "monitoring" and "task setting", many other functions of cognitive control have been proposed giving rise to the "fractionated" perspective (Miyake et al., 2000). In practice, it is not entirely clear how these functions (and their failures) differ precisely from each other and this confusion only becomes evident as one attempts to map the following three levels of representation together: 1) behavioural models, 2) language and 3) brain models. The paradoxes that emerge from this mapping exercise suggest the potential dangers of relying exclusively on any one level (e.g. overfitting) and ignore other layers of complexity that may instead provide useful constraints to reflect the real phenomena. I will use examples to illustrate this point. The Stuss *et al* study (Stuss et al., 2000) mentioned in the previous paragraph suggests that "task setting" is both strategic and separable from "monitoring". However, it is quite easy to conceive of "monitoring" as a form of strategy per se: a linguistic consideration would indicate that "monitoring" is somehow dependent on the capacity of "task setting". However, from a brain model perspective the neuropsychological evidence (Stuss et al., 2000) suggests that these functions have hemispheric specialization and are dissociable. Hence the paradox.

Chapter 1 | Introduction

A mental concept riddled with such linguistic vagaries is "response inhibition" which refers to the ability to suppress a prepotent response. First, as Ward rightly highlights, it is not entirely clear how "response inhibition" is dissociable from "monitoring", as impairments in monitoring would lead to impairments in response inhibition (Aron et al., 2004; as cited in Ward, 2015). Second, the existence of "response inhibition" as a separable function is further complicated by its very own fractionation into subtypes. For example, a neuropsychological study which I was involved in (Cipolotti et al., 2016) indicates a dissociation between a Stroop-like-response-inhibition, compared to a Hayling-test-like-response-inhibition with left and right prefrontal cortex patients having comparatively worse performance on each type, respectively. Third, additional subtypes of response inhibition come in the shape of "perseveration" and "capture" errors. To add to the confusion both of these errors are thought to be task-setting-related despite response inhibition being considered to be closely related to monitoring, whilst monitoring is meant to be different to task setting. To illustrate this third point more clearly, let's recall the perseveration errors in the WCST which were mostly observed in patients with lesions in the left PFC and interpreted as an impairment of rule induction, or task setting (Stuss et al., 2000). A similar finding was observed using a different task called the Brixton test (Reverberi et al., 2005). In each trial of this task, a blue coloured circle moves to one of ten possible positions in space and the participant is required to predict where the blue circle will move next. The prediction is made by inducing the current pattern (rule) based on the blue circle's moves, for example, an alternating pattern. Like in the WSCT, the patterns change without explicit notice. Hence, continued predictions based on a previous, but now irrelevant, rule is classified as an error of perseveration. This perseveration occurs despite indirect feedback that the response is incorrect (blue circle does not move to predicted position). An additional component to this task involves red circles, which replace blue circles for short periods of time. For red circles, instead of predicting the next move, the instruction is to simply select the position where the red circle is. The red circle moves around the board following a pattern. When the blue circles reappear, some participants persevere with the pattern established by the red circle - a so called "capture error" given that the participant is captured by the red rule. Patients with right PFC damage show more capture errors compared to left PFC patients (Reverberi et al., 2005).

In summary, this neuropsychological evidence suggests a separable localization for what at a descriptive and behavioural level appear like very similar errors (perseverations and capture errors). Task-setting and monitoring are considered separate and yet, perseveration/capture errors are thought to be both impairments of task-setting but also fundamentally reflecting an inability to inhibit a prepotent response i.e. a monitoring impairment (!) These linguistic paradoxes may be compounded by the phenomena of *neural degeneracy* whereby different brain areas are thought to perform similar functions or, *neural reuse* in which brain areas are reused for a different function without losing their original function.

1.4   A UNIFIED ACCOUNT OF COGNITIVE CONTROL AND THE SPECIAL ERROR OF GOAL NEGLECT

A different view to a fractionated account of cognitive control is one which suggests a more unified mechanism of control (Miyake et al., 2000; Duncan, 2013). This view advocates that despite the studies on hemispheric differences mentioned above, replications of these findings have largely been lacking (Duncan and Owen, 2000; Robinson et al., 2012; Duncan, 2013). Duncan's unified account proposes that cognitive control is driven by "attentional episodes", where each episode involves focused processing of a part of the goal at hand (Duncan, 2013). This proposal is based on the key idea that solving a problem requires breaking it down into sub-problems and processing each sub-problem one after the other. Hence, in contrast to the fractionated account which refers to separate "planning" "monitoring" "strategizing" "inhibition" etc. functions, this attentional episode view is unified in as much as each attentional episode is only considering whatever contents and control processes are required by the current sub-problem and can rapidly adapt to deal with a different set of requirements in the next sub-problem.

Goal neglect (GN) is closely linked to the attentional episode account; in fact its contribution to understanding cognitive control should be understood in the context of various other key ideas that are linked to this account. What follows is a description of these concepts: i) GN, ii) fluid intelligence, iii) GN within the context of a fractionated account of cognitive control, iv) GN and fluctuations of attention v) the multiple demand (MD) network and, vi) adaptive neural coding.

Chapter 1 | Introduction

1.4.1 ATTENTIONAL EPISODES AND GN

Attentional episodes are associated with a particular type of error called GN which is central to this dissertation. During GN, a person shows an understanding of task requirements but fails to execute part or all of these requirements, as if these are not controlling behaviour (Duncan et al., 1996). Hence, this form of failure can be thought to reflect a dissociation between knowledge and action. Although the term GN was coined by Duncan (1996), this phenomenon, which is not uncommon in patients with major frontal brain injuries, was first portrayed in accounts by the neuropsychologists Alexander Luria and Brenda Milner (Milner, 1963; Luria, 1966). Luria describes GN-like failures during a simple stimulus-response task with a frontal patient. In this task, the patient is simply required to press a button when a light is switched on. However, despite the patient's intact ability to recall instructions, the patient fails to press the button when the light is on (Luria, 1966). Luria described these patients as often showing deficits in making any plans when faced with a novel task and instead tended to attempt to solve the task immediately, frequently in a fragmented, disorganized way and typically leading to errors which they would not recover from. From these observations, Luria inferred that the role of the frontal lobes was with planning, controlling and verifying of behaviour (Luria, 1966). In daily life, GN is manifest as a difficulty in managing tasks that are made up of various sub-tasks, such as shopping for food or preparing a meal. The impact of this kind of impairment can be considerable despite other cognitive functions being intact (Shallice and Burgess, 1991).

GN is not only observed in brain lesioned patients but also in the normal adult and children populations (Duncan et al., 1996; Zelazo et al., 1996; De Jong et al., 1999; West and Alain, 2000; Kane and Engle, 2003; Zelazo, 2004; Towse et al., 2007; Duncan et al., 2008; Altamirano et al., 2010; Marcovitch et al., 2010; Bowling et al., 2012; Morey et al., 2012; Bhandari and Duncan, 2014; Roberts and Anderson, 2014; Unsworth and McMillan, 2014). In a child-adapted version of the WCST, children of age +3 could successfully sort cards according to colour and shape. However, when asked to switch from one sorting rule to another (colour to shape or shape to colour) they failed to do so and persevered with the first rule, despite knowing and being able to report the new rule (Zelazo et al., 1996; Zelazo, 2004). In the normal adult population, GN can be observed when tasks are complex, such that one or more rules of the task are completely ignored

(Duncan et al., 1996; De Jong et al., 1999; Kane and Engle, 2003; Duncan et al., 2008; Marcovitch et al., 2010; Bowling et al., 2012; Morey et al., 2012; Bhandari and Duncan, 2014). At the end of the task, when memory for task instructions is assessed, the neglecting participant is often observed to display surprise and embarrassment as they successfully recollect the neglected rule (Duncan et al., 1996).

The first experimental task which resulted with the coining of the term "Goal Neglect" is the Letter Monitoring Task (Duncan et al., 1995; Duncan et al., 1996). In this task, participants saw a stream of letters and numbers appear one pair at a time in the centre of a computer screen (Figure 1.1). They were instructed to attend one side, left or right, and ignore the other side of the character stream. The task was to read the letters or numbers aloud from the attended side.
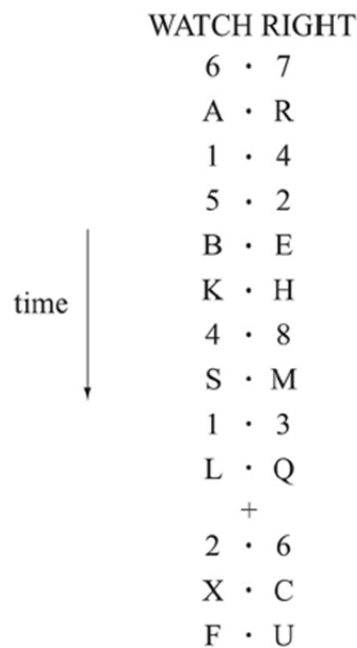
Figure 1.1 The Letter Monitoring Task

The picture shows a trial from the Letter Monitoring Task, (illustration reproduced with no permissions required from Duncan et al., 2008). In this trial, the first side-cue states "WATCH RIGHT" hence the correct response up until the second side-cue is "7,R,4,2,E,H,8,M,3,Q". The second side-cue is a "+" symbol which indicates attending the right side so, the correct response from that point until the end of the trial is "6, C, U". If the second side-cue had been a "-" symbol then the correct response for the last part of this trial would have been "2,X,F".

Each trial contained two side-cues, each signalling which side needed to be attended for the letters and numbers which followed. The first side-cue appeared at the start and contained the verbal instruction "WATCH LEFT" or, "WATCH RIGHT". The second side-cue was displayed towards the end of the trial and was either the symbol "+" or "-" which indicated right and left, respectively. The combination of side-cues was such so that for half of the trials the second side-cue required the participant to switch from one side of the stimulus stream to the other. Patients with frontal lobe injuries showed similar performance to healthy controls on the first part of each trial, before the second side-cue was shown (Duncan et al., 1995). However, the patients showed significantly more errors on the final part of the trial. These errors were mostly side-errors, where instead of following the instruction indicated by the second side-cue, the patients persevered with the side indicated by the first side-cue. Despite the neglect of the second side-cue, patients demonstrated intact memory of the task instructions, before and at the end of the task. In addition, patients showed no hesitations or attempt to correct their behaviour (Duncan et al., 1995). Overall, this suggested that this dissociation between knowledge and action was driven by an impaired or incomplete control structure which Duncan (2008) refers to as the "task model".

It is worth noting that the lack of hesitations or attempts to correct the neglect has elsewhere been found to support the idea that this type of error is driven by different mechanisms than say, corrected errors. In an eye-tracking study, Bowling *et al.,* (2012) used an antisaccade task where participants were instructed to gaze away from the target on screen and instead look at its mirror position. Failure to inhibit the reflexive saccade to the target was referred to as a prosaccade error. This study revealed two variants of this prosaccade error: uncorrected saccades and, corrected saccades (i.e. the gaze shifts from the incorrect target position to the correct mirror position). Interestingly, only the uncorrected saccades correlated with measures of spatial memory and inhibition, suggesting that these two errors were driven by distinct processes, with uncorrected saccades considered a form of GN (Bowling et al., 2012).

One important finding about GN is that the extent of neglect was shown to be sensitive to the complexity of the task at the time that the instructions were given. This is referred to as the "Instructional Complexity" (IC) effect. Although I discuss the topic of complexity and related findings in more detail in a separate section of this chapter (see p.13), at this

point it suffices to explain the basic experimental manipulation to test for the IC effect and illustrate this with an example. The experimental design requires 2 conditions; in one condition participants are given basic + extra instructions whilst in the other condition, participants are only given the relevant basic set of instructions. Just before the start of the main task, the group with the extra instructions is told that the extra instructions are no longer relevant. Both groups perform the exact same task which only requires the basic instructions. The IC effect is when the participants in the extra instruction condition show significantly more GN, despite performing the identical task to the other group. The IC effect was observed in manipulations of the Letter Monitoring Task (Duncan et al., 2008) and the Feature Match Task in both adults and children (Duncan et al., 2008; Roberts and Anderson, 2014).

In the Feature Match Task (Figure 1.2), a pair of numbers appears on a display, one pair at a time. The response to each display is determined by the combination of frames and frame properties that surrounded the numbers. The numbers can be either frameless or framed. In the latter case, the frame can vary in colour and shape. For frameless trials, the rule is to sum the pair of numbers and say the result aloud. Responses to the framed trials varies: if the frames are different in both shape and colour then, the correct response is to withdraw response; if the frames match on only one dimension (either shape or colour) then the correct response is to press a button on the side of the larger number (either the *V* or *B* keys on a computer keyboard); if the frames match both dimensions then the correct response is no response, just like when the frames are different on both dimensions. In the extra instruction condition participants are given all of the rules mentioned above, whilst in the basic instructions condition participants are never told of the rule involving a verbal response. In reality, the task for both groups never includes frameless trials requiring verbal responses and before starting the main task, the extra instruction group is told that this rule is no longer relevant and frameless numbers would never appear. Participants in the extra instruction condition showed significantly more neglect, demonstrating the IC effect (Duncan et al., 2008; Roberts and Anderson, 2014). The main interpretation to this finding was that the "task model" had functional limits that were significantly determined by IC.
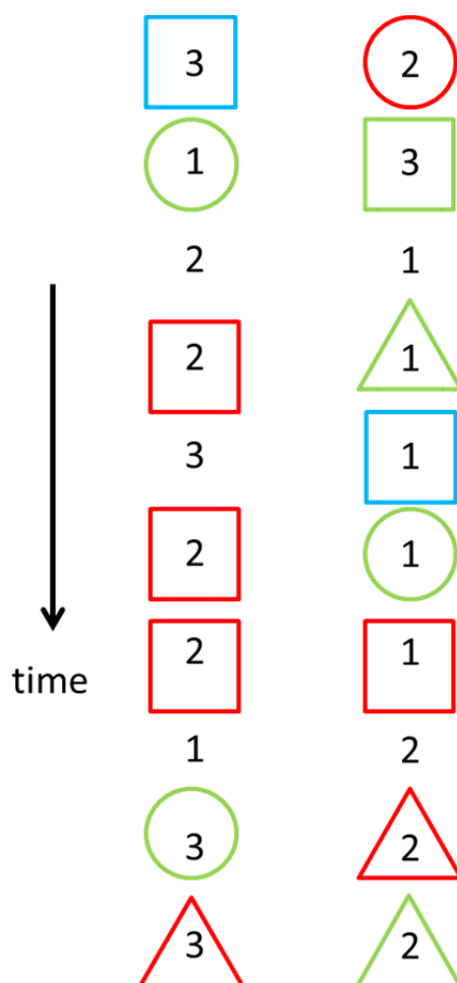
Figure 1.2 The Feature Match Task

This figure illustrates the stimuli stream from the Feature Match Task as described in Experiment 4 of Duncan *et al.*,(2008). One pair of numbers appeared on the display at a time. The combination of number, shape and colour of the frames surrounding the pair of numbers corresponded to different responses. Frameless trials like the 3rd and 8th row in this picture were never actually presented in the main task. Roberts and Anderson (2014) adapted the Feature Match Task for pre-adolescent children and replicated the Instructional Complexity effect in this age group.

Another key finding linked to GN is that this error is significantly more prevalent in adults with poor scores of fluid intelligence (Duncan et al., 1996; Duncan et al., 2008; Duncan et al., 2012; Bhandari and Duncan, 2014). At first, this may seem unsurprising given that poor performance across many cognitive tasks is associated with low scores on fluid intelligence; in fact, in psychometrics, the concept of fluid intelligence is an emergent property of this pattern of performance (further discussion on this topic is provided in the next section on p.27). Nevertheless, the negative correlation between fluid intelligence and GN has been reported to be particularly strong, at least stronger than various forms of working memory (Duncan et al., 2012). One interpretation to this result

is that the cognitive limits underlying GN overlap closely with what is measured by tests of fluid intelligence (Duncan et al., 2008; Duncan, 2013) .

One of these cognitive limits may be related to a form of dissociation between knowledge and action. A dissociation between knowledge and action is not a new phenomenon and has been previously noted in the disorder of hemispatial neglect. This disorder is typical of patients with lesions to the right parietal region and is characterized by deficits of awareness of stimuli in the contralateral hemifield (usually the left visual hemifield). Hemispatial neglect is considered primarily a disorder of attention rather than perception, because of a series of features: the severity of the symptom varies with attentional demands; there is evidence for implicit processing of the neglected stimuli suggesting that these stimuli are in fact perceived, and hemispatial neglect affects other senses not just vision (Ward, 2015). In some sense, hemispatial neglect is to hemianopia (blindness over half of the visual field) as GN is to motor inability. In hemianopia, deficits in awareness are restricted to vision and this deficit cannot be overcome by directing attention to the blind region; in hemispatial neglect, the deficit is not restricted to vision and it can be overcome by directing attention to the neglected region. Analogously, in motor inability, impairments are within the motor system and this deficiency cannot be resolved by paying attention to the paralysed limb; in GN, the deficit is not restricted to motor acts and the failed performance on a neglected rule can easily be restored by directing attention to the rule (Duncan et al., 1996). Interestingly, the similarity to GN is even closer for "spatial imagery neglect" whereby the representation of an imagined (and real) scene is partially reported depending on which perspective the patient is asked to take, thus demonstrating that the spatial knowledge of the entire scene is intact, but only parts of it are available for report (Bisiach, 1996). So, GN, like hemispatial neglect is characterised by fragmented attention of some type of mental representation. GN is also associated to low fluid intelligence. Is there a link between attention and fluid intelligence?

1.4.2 ATTENTIONAL EPISODE AND FLUID INTELLIGENCE

This brings me to the second link, between attentional episodes and fluid intelligence. The psychometric concept of general intelligence originates from the pervasive finding that scores on cognitive tests correlate positively with each other, suggesting that people who perform well on a test are likely to do well on other tests (Spearman, 1904). This

"positive manifold" was interpreted by Spearman as reflecting a general factor or "g", contributing to effective performance across all sorts of abilities both in the lab and in real life – hence the concept of general intelligence, also referred to as "fluid intelligence". However, others like Thorndike and Thomson interpreted the positive correlation as consistent with multiple biological units, rather than a single one (Deary et al., 2010). Going back to Spearman's "g", to measure this factor one could administer a large battery of all kinds of cognitive tests and calculate the mean score. Alternatively, one could capitalize on the positive manifold and administer just one test that consistently correlates most highly with all other tests. In other words, this test strongly captures "g" and hence is a good indicator of general intelligence. Typically tests that load highly across the correlation matrix are the ones that involve novel problem solving and often include matrices and analogies such as Cattell's Culture Fair test (Cattell, 1971; Cattell and Cattell, 1973) and Raven Matrices (Deary et al., 2010). Duncan proposes that attentional episodes underlie core aspects of fluid intelligence and GN and that in the brain, the MD network is a likely neural substrate for this basic aspect of mental programs (Duncan, 2010a, 2013). The link between attentional episodes and fluid intelligence is made via the concept of abstraction. Abstraction can be thought of as the ability to identify a common element across exemplars, for example, closer inspection of disorders of depression, attention-deficit-hyperactivity and substance-abuse may show common impairments in reinforcement learning. Abstraction allows one to deconstruct a complex problem into optimal sub-components (which have common features) which may facilitate efficient cognitive segmentation of the problem via distinct attentional episodes (Duncan et al., 2017). Abstraction is necessary in rational thought required to solve problems – to continue with the previous example, investigating brain networks related to reinforcement learning may be critical to understand disorders of depression, attention-deficit-hyperactivity, substance-abuse etc… Hence, abstraction allows one to transpose a potential solution from one exemplar problem to another after having identified a common feature, as Duncker states in his monograph on problem-solving:

For, one can transpose a solution only when one has grasped its functional value, its general principle, i.e., the invariants from which, by introduction of changed conditions, the corresponding variations of the solution follow each time. (p.5, Duncker, 1945)

1.4.3   GN WITHIN THE CONTEXT OF A FRACTIONATED ACCOUNT OF COGNITIVE CONTROL

In contrast to focusing on the association between GN and fluid intelligence, other studies have instead attributed the phenomenon of GN to momentary lapses of working memory capacity (WMC) (Kane and Engle, 2003; Marcovitch et al., 2010; Morey et al., 2012). Within the unified vs. fractionated accounts of cognitive control, WMC is usually considered within the latter category. Kane and Engle (2003) have shown that WMC significantly predicted GN on a Stroop task in 5 separate experiments in which they varied the relative proportion of congruent and incongruent trials. The basic Stroop task (Stroop, 1935, as cited in Kane and Engle, 2003) involves seeing a colour word on screen (e.g. "red"), one word at a time in a speeded fashion. In congruent trials, the colour word is presented in the same ink colour (e.g. "red" in red ink) whilst in incongruent trials the ink does not match the colour word (e.g. "red" in blue ink). Typically, the task is to read aloud the colour of the ink and ignore reading the word. The incongruent trials are usually found to be particularly challenging which is understood to be driven by competing dimensions of the stimulus (word dimension vs. ink colour dimension). Kane and Engle (2003) showed that WMC predicted GN when the proportion of congruent trials to incongruent trials was high. Instead, when the latter ratio was low, WMC only predicted latency on trials, suggesting that contexts that reinforced the task goal minimized the load on goal maintenance. A similar finding was replicated in a study by Morey *et al.* (2012). Furthermore, Marcovitch *et al.* (2010) extended Kane and Engle's (2003) findings to children using a child-adapted version of the WCST. This study indicated that individual differences in WMC of 4- and 6- year old predicted their ability to maintain novel goals (Marcovitch et al., 2010).

Other studies on GN have focused on other sub-components of cognitive control other than WMC. For example, a study of GN in children (Towse et al., 2007) used a child-adapted version of the Letter Monitoring Task, the Image Naming Task (Figure 1.3).
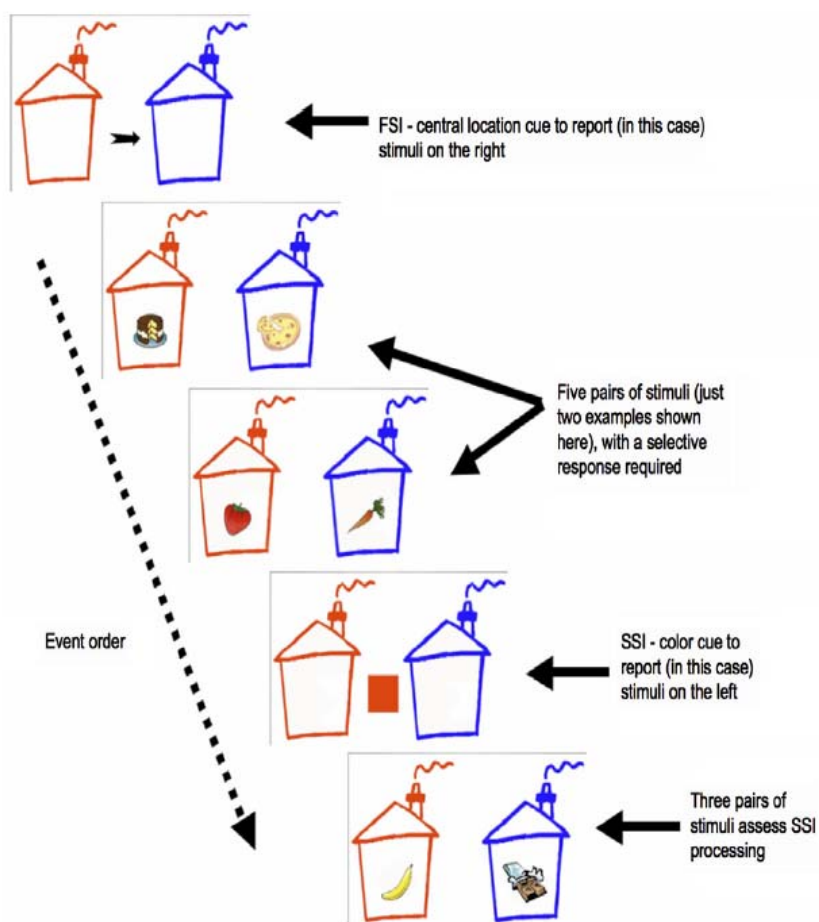
Figure 1.3 The Image Naming Task

The picture illustrates the Image Naming Task (figure reproduced with no permissions required from Towse et al., 2007) which was used to test GN on 4-year old children and is based on the Letter Monitoring Task (see Figure 1.1). FSI stands for first side-cue and SSI, second side-cue. In this study, neglect at different time points was found to load differently on components of cognitive control. Neglect at the start loaded more on response inhibition, whilst neglect of the second side-cue was associated with set shifting (flexibility).

This study revealed that GN at different time points in the task loaded differently on executive functions in 4-year old children. Specifically, neglect of the first side-cue was significantly associated to errors of inhibition as tested by an opposite colour inhibitory task, whilst neglect of the second side-cue was significantly correlated to errors on child-adapted version of the WCST. Overall, this indicated that successful performance on the second side-cue depended more on the ability to shift instructional set than response inhibition, as measured in this study. Interestingly, this fractionation of GN is supported by a study (Altamirano et al., 2010) that found that mental inflexibility as expressed in depressive ruminators was advantageous in tasks that emphasized the GN component of

goal maintenance (via a modified Stroop task) compared to tasks that emphasized the GN component of shifting between goals (via the Letter Monitoring Task).

Overall, the studies presented in this section focused on the association between GN and component parts of cognitive control (e.g. working memory, inhibition, set-shifting), in line with the fractionated view. Although this approach is meaningful, other authors favour the unified view and argue that unitary control factors (e.g. fluid intelligence) which may reflect the coordination of complex behaviour, provide a better fit of successful goal-directed behaviour (Duncan, 2010b; Roberts and Anderson, 2014).

1.4.4 FLUCTUATIONS OF ATTENTION AND GN

Another cluster of GN findings pertains to the idea that fluctuations in attention prior to the event are predictive of GN (De Jong et al., 1999; West and Craik, 1999; West and Alain, 2000; Unsworth and McMillan, 2014). Using a Stroop task, De Jong and colleagues (1999) showed, that GN increases when the response-stimulus interval is large when compared to a short one. In the former, it is thought that the gap in time increases the probability of lapses of attention, whilst in the latter attention is more tightly focused on the task goal. Reaction time analyses further suggested that the largest difference in the Stroop effect was driven by the slowest reaction times. Overall, these findings indicated that GN is at least partly driven by lapses of attention which fluctuate on a trial-by-trial basis (De Jong et al., 1999). In addition, EEG studies have suggested that a slow wave in the frontal brain region prior to stimulus onset is predictive of GN (West and Alain, 2000) which is further supported by studies showing that self-reports of pre-trial attentional states were correlated to GN (Unsworth and McMillan, 2014). This neuroimaging work brings me to the next point – on the Multiple Demand (MD) brain network,

1.4.5 ATTENTIONAL EPISODES AND THE MD NETWORK

A consistent set of fMRI studies have revealed a brain network involving the prefrontal and parietal cortex referred to as the Multiple Demand (MD) system. Activation of this fronto-parietal network has been observed across a wide range of demanding cognitive tasks (Duncan and Owen, 2000; Duncan, 2010b; Fedorenko et al., 2013). Duncan goes further to suggest that the MD network is likely related to processing of attentional

episodes, where configuration of components of the current attentional episode are "bound" together, one episode at a time, as the problem at hand is solved (Duncan 2013). Interestingly, BOLD (Blood Oxygenation Level Dependent) activity in the MD network appears to ramp up at the boundary between perceived events which may reflect the subdivision of cognitive events into sub-components, with maximal BOLD for activities higher in the hierarchy (Farooqui et al., 2012).

The link between attentional episodes and the brain is primarily derived from the association between fluid intelligence and the MD network. First, a strong clue that the MD network is a neural correlate of fluid intelligence is an analogous finding to the positive manifold. Similarly to how the concept of general intelligence is an emergent property of the positive manifold, so is the BOLD activation in the MD network which is observed across many different tasks. Duncan (2013) proposes that MD regions may in fact be controlling all sorts of complex behaviour via the operation of attentional episodes. Second, is support from neuropsychological studies. Frontal lobe lesions have been associated with impairments of cognitive control, as mentioned earlier in this chapter. Especially sensitive to such lesions are open ended problem-solving tasks in which the patient is given an end-point and a starting-point and the participant must generate their own solution, often with no feedback (e.g. tower of London task Shallice, 1982; Morris et al., 1997). Historically, such control deficits were dissociated from intelligence test performance, suggesting that intelligence was independent of the frontal lobes (Hebb, 1940; Luria, 1966; both as cited in Duncan, 2013). However, this is likely to be due to the type of intelligence tests used which did not distinguish between fluid and crystallised intelligence, the latter concerning acquired knowledge and being less sensitive to brain damage (Cattell, 1971). Studies have shown fluid intelligence deficits in patients with frontal lobe lesions (Duncan et al., 1995) and specifically for damage within the MD network (fronto-parietal) but not outside it (Woolgar et al., 2010). In addition, patient-control differences on control tasks such as the WSCT, have been reported to be largely reduced, if not completely removed, when fluid intelligence was partialled out (Roca et al., 2010; however see our paper for persisting differences across frontal-left and -right patients after partialling out fluid intelligence on two 'inhibition' tasks, Cipolotti et al., 2016). Third, is support from fMRI (functional Magnetic Resonance Imaging) studies, where contrasting activity in fluid intelligence tasks compared to sensorimotor tasks

shows an MD pattern (Prabhakaran et al., 1997; Duncan et al., 2000; Bishop et al., 2008). However, there is also evidence which does not support the MD network as key to age-related differences in fluid intelligence (Kievit et al., 2014), and instead indicates that the integrity of frontal white matter tracts and grey matter in BA10 are significantly better candidates.

1.4.6 ATTENTIONAL EPISODES AND ADAPTIVE NEURAL CODING

The MD network has been theorized to work as a single unit in flexibly coding task-relevant information for goal-directed behaviour via an adaptive neural system. Evidence for adaptive neural coding comes primarily from single-cell recording studies which indicate that large parts of the prefrontal cortex code for activity which is relevant to the current episode, for example, relevant stimuli, responses, rules and so on. This activity changes flexibly as the task progresses to the next episode (Duncan, 2001, 2013). In addition, neural activity has been investigated across different stages of a task such as presentation of the cue, delay between cue and choice stimuli and selection of choice stimulus as target or non-target (Stokes et al., 2013). Activity in coalitions of neurons is reorganized adaptively, reflecting the current behavioural relevance of each different stage (Duncan, 2013; Stokes et al., 2013).

1.4.7 INTERIM SUMMARY

In the previous sections I described how errors can provide critical information on the underlying mechanisms of goal-directed behaviour, with evidence being interpreted in two broad categories: a fractionated and a unified view of cognitive control. The error of GN has been presented within the constellation of related findings which are the basis for building a mechanistic account of cognitive control. Next, I turn to the concept of complexity.

1.5 COMPLEXITY, STRUCTURE AND THE LIMITS OF ATTENTION

1.5.1 HOW DO ERRORS RELATE TO COMPLEXITY?

When faced with a complex task, poorer performance is generally more likely than if the task were simple. In fact, measures of performance are often used as an index of task complexity. Considering that human endeavour is always pushing itself towards more complex and novel problems, the question of how complexity affects behaviour is of fundamental relevance. In addition, manipulating complexity has in itself been a method

Chapter 1 | Introduction

to explore cognitive control. For example Atkinson and Shiffrin demonstrated controlled processing by investigating one specific aspect of cognitive control, rehearsal, to which they applied various experimental conditions which varied the rehearsal demands (Atkinson and Shiffrin, 1968; as cited in Schneider and Shiffrin, 1977). Nonetheless, investigating complexity is challenging and merely defining the term is not straightforward. In this section, I endeavour to introduce key ideas surrounding the concept of complexity and how this can inform our understanding of cognitive control. However, before I proceed with this task, I will first clarify the distinction between the terms "complexity" and "demand" as per use in this dissertation.

A term which is closely linked to task complexity is "demand" – poorer performance on a complex task is driven by an increased likelihood of the complex task being more demanding than a simple task. However, it is not entirely clear how "demand" can be dissociated, if at all, from "complexity". Can a complex task be low-demand? Or vice-versa, can a simple task be highly demanding? Vigilance and sustained attention tasks (e.g. see the SART, Sustained Attention to Response Task by Robertson et al., 1997) may, at first glance, fall within the simple-but-demanding category. The instructions to these tasks are relatively straightforward with simple stimulus-response mappings whereby the participant needs to respond with a button press to certain stimuli and withdraw response to other stimuli. A key feature of these tasks is that the stimuli are displayed for extended periods of time (Dillard et al., 2014). Performance at the start can be relatively high, with decline appearing later on which is thought to reflect decreases in "sustained attention". Hence, the critical difficulty in these tasks seems to lie within their prolonged nature. Arguably, even though the instructions appear simple, the demand of such tasks lies within managing resources over long periods of time, and in this sense the task is not spared of being complex. In a neuroimaging study, keeping unchanged the rules of a stimulus-response task whilst increasing the difficulty of perceptual discrimination showed poorer performance and critically, significant higher coding of task information in frontoparietal (MD) regions which suggests increased task demands (Woolgar et al., 2011). In this example it is not really possible to determine if changing perceptual discrimination simply increased demand without affecting complexity or, vice-versa. In summary, it appears that distinguishing "complexity" from "demand" is, yet again, a somewhat linguistic problem and to avoid confusion these terms should be

explicitly defined. In this dissertation, I will consider demand to be an emergent property of complexity in such a way that high demand is a direct consequence of high complexity, and similarly for low demand being a result of low complexity.

1.5.2   CAPACITY LIMITATION

In the previous sections, I noted the idea of "attentional episodes" as a hypothetical mechanism underlying cognitive control. Broadly speaking, attention can be thought of as a mechanism for the selection of information (Ward, 2015). A key feature of attention is that it is capacity limited (Kahneman, 1973). Capacity limitation is typically demonstrated in one of two ways (Duncan, 1980). The first is to show that performance suffers as experimental conditions are manipulated to increase recruitment of assumed internal processes. The second is to show poorer performance as a consequence of experimental conditions which bias allocation between these internal processes such that core processes lose capacity. However, the way capacity limitation is inferred is not always correct and other factors may be at play (Duncan, 1980). For example, based on a diffusion model, observing reduced performance as a consequence of increasing the number of response options could be explained by an increased likelihood of choosing the wrong response by chance, rather than due to a capacity limitation (Duncan, 1980).

1.5.3   TYPES OF COMPLEXITY MANIPULATIONS AND TASK STRUCTURE

Understanding capacity limitation is intimately linked to understanding how complex tasks affect controlled behaviour, given that, to some extent, greater complexity imposes greater demands on a limited resource(s). Typically complexity/demand has been investigated in real-time, that is, during task execution. Two main conceptions of demand (which I will refer to as Actual Task Complexity (ATC)) are in the form of either (i) immediate competition, such as processing multiple stimuli or responses simultaneously (Broadbent, 1958, as cited in Duncan et al., 2008) or (ii) delayed competition, such as preparatory processes for different rules within the same task (Pashler, 1994; Rogers and Monsell, 1995). However, a different kind of complexity emerged with research on GN. This is complexity that is present at the outset of the task, *before* the start of task execution and, refers to the relevant task facts, rules and requirements, collectively termed as Instructional Complexity (IC) (Duncan et al., 2008).

Capacity limitation is characteristic of ATC (Fougnie et al., 2006) and, commonly, this capacity is theorized to be made up of various separate components, with processing

limitations bounded within each one of these components (for example Baddeley's theory of working memory and its sub-component systems: the central executive, the visuo-spatial sketchpad, the phonological loop and the episodic buffer, Baddeley, 1986). A similar conceptualization of capacity limitation stands for IC and I will briefly describe how this understanding developed. During early research on GN (Duncan et al., 2008), GN was observed to be immune to several manipulations of ATC which suggested that this error did not arise from capacities dealing with real-time demand, but by a different capacity altogether. Instead, manipulating the task instructions significantly affected GN hence suggesting that GN was associated with a new kind of capacity - one which was limited in the use of complex instructions.

I described the IC effect using the example of the Feature Match Task in an earlier section of this chapter (see p.26). In summary, the IC effect is when 2 groups go on to perform the exact same version of the task (the one with basic instructions), but the group who learnt extra, irrelevant, instructions at the start shows significantly more GN (Duncan et al., 2008). Later GN experiments that used the Panel Task revealed that this capacity was not globally limited, at least not in the ways it was initially thought of (Bhandari and Duncan, 2014). In the Panel Task (Figure 1.4) participants went through 4 tasks and at the start of each task, instructions to two sub-tasks were administered (Bhandari and Duncan, 2014). The sub-tasks within each task had different stimuli, for example see the top row of Figure 1.4 which shows two different sub-tasks, the vehicles and the books sub-tasks. The sub-tasks were either both complex, both simple or, one was simple and the other complex. The top row of Figure 1.4 displays sub-tasks in the complex form, whilst in the bottom row is displayed the simple form of the vehicles sub-task. For simple forms of the sub-task there were less rules and less stimuli on the display. For example, the rules for the vehicles sub-task were as follows: the participant was first instructed that they had to look at the cue in the centre of the screen, if this was a minus symbol then they had to look at the two panels on the left, if it was a plus symbol then look at the two panels on the right. The second rule regarded the choice of the panel, such that, on the side they were looking at as a result of the first rule, they had to choose the panel which displayed a motorbike. Finally, they had to respond by pressing the lowercase letter underneath the motorbike. If however, there was a dot directly below the lowercase letter, then they had to press the letter T instead of the lowercase letter. The combination in

which the letter T required pressing occurred 25% of the time and these were defined as critical trials. The rest of the trials required responding by pressing the lowercase letter and were regular trials. GN was primarily measured as cases of incorrect responses on critical trials.
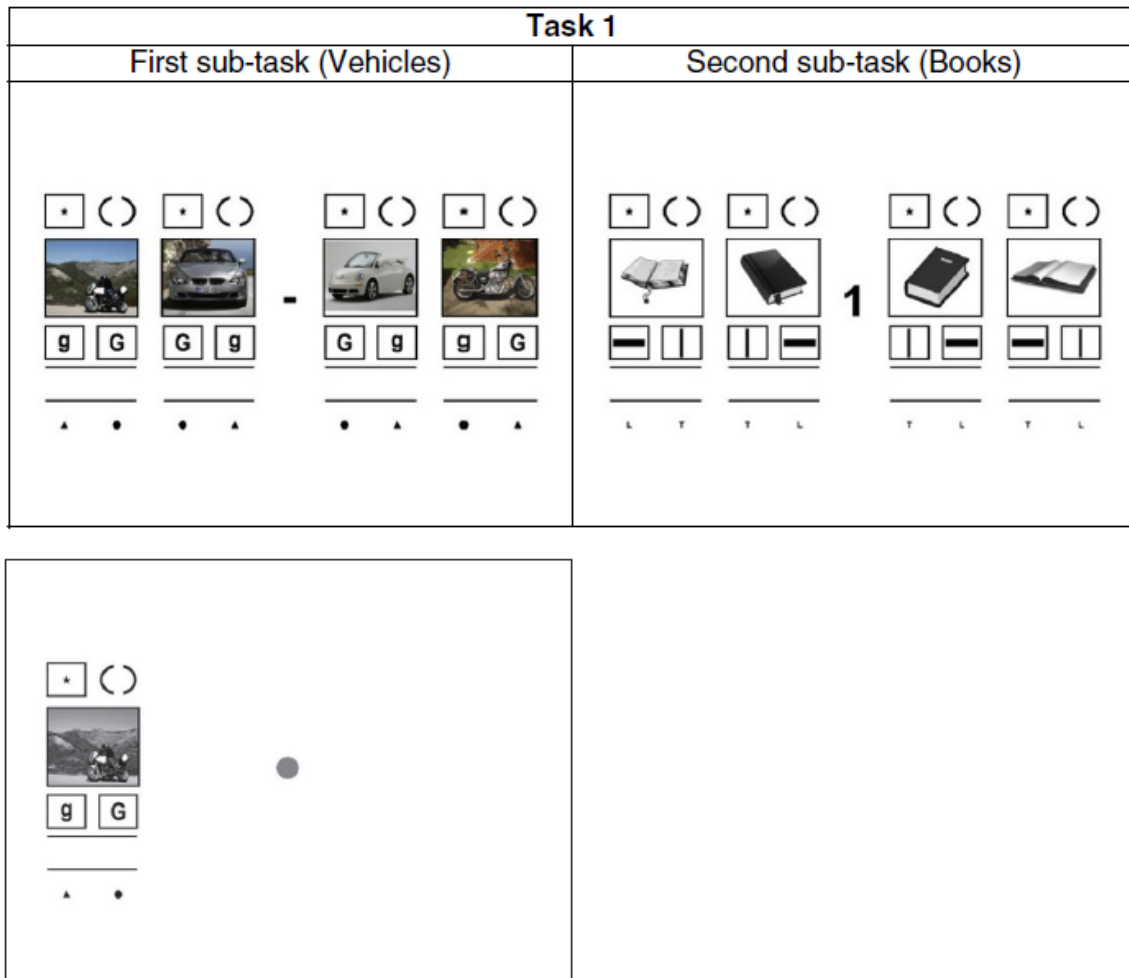
Figure 1.4 The Panel Task

The figure illustrates stimuli from the Panel Task (figure reproduced from Bhandari and Duncan, 2014). The top row shows one of the four tasks administered to participants. Each task was made up of two sub-tasks, in this case the vehicles and books sub-tasks. The sub-tasks came in one of two forms, complex or simple. In task 1 (top row), the sub-tasks are both in the complex form. The bottom row shows the simple form of the vehicles sub-task. In the simple form only one of the four panels appeared, at one of the 4 possible positions, in a random fashion. Hence, the simple form did not require instructions that lead to the choice of the panel hence reducing the instructions to keep in mind.

The key finding of this experiment was that rather than the total amount of IC affecting GN, it was the IC within each sub-task that was important (Bhandari and Duncan, 2014). This suggested that the limits of this capacity were bounded by the task structure, which

is both reminiscent of the concept of chunking and, as mentioned earlier, similar to the way real-time capacity is constrained within the limits of each one of its sub-components (e.g. phonological loop vs. visual sketchpad components within Baddeley's model of working memory).

A good starting point is to consider complexity according to three dimensions via which it can be manipulated externally: 1) quantity, 2) quality and 3) time. Quantity refers to the intensity of a given demand, for example, the level of degradation of stimuli. Quality denotes the type of manipulation, for instance altering the semantic link of a cue to the rule it denotes (transparency) or changing the structure of the task. The latter could include changes in the structural relations between rules or parts of an argument (for e.g. see Badre and D'Esposito, 2009). Time refers to the temporal order of the complexity manipulation, for example, IC occurs early on, before the start of the task whilst ATC occurs during task performance.

1.5.4 COMPLEXITY AS MEANS TO GAUGE COGNITIVE ARCHITECTURE

As previously mentioned, the assumption is that as external complexity is manipulated this either affects the number of internal processes involved or the way that resources are allocated between them to deal with this external demand (Duncan, 1980). Critically, *how* the mind deals with this demand can provide information on the internal architecture of cognition. To illustrate this point I will refer to an example from the literature on cognitive offloading which concerns the use of actions to modify the processing of a task to decrease cognitive demand (Risko and Gilbert, 2016). When participants face a rotated stimulus (for instance a slanted book), they will frequently tilt their head to align to the stimulus' orientation. As the rotation demand becomes stronger (few degrees rotation vs. 180 degrees) or the number of items in the display increases, individuals are more likely to show this behaviour (Risko and Gilbert, 2016). This phenomenon is referred to as external normalization and can be thought of as a way to offload internal normalization - an internal transformation (in this case a mental rotation) that "aligns the representation of a stimulus with a representation stored in memory" (Risko and Gilbert, 2016). Hence, manipulating complexity and any accompanying behavioural output can help us understand internal mechanisms.

1.5.5 GN, PROSPECTIVE MEMORY, MULTITASKING AND TASK STRUCTURE

As internal processes deal with the external complexity, it is hard to tell whether the distinction between these external parameters is at all preserved within the mechanisms of the mind. It is easy to imagine how the intensity (quantity) of one manipulation of complexity may reach a threshold, in a step-like function, and then be dealt with in a qualitatively different way. Similarly difficult to tell, is how complexity applied at the outset may or may not interact with complexity during the task. This point is of particular relevance in research on GN. Considering that manipulations of ATC have not been shown to affect GN (Duncan et al., 2008), this may suggest that there isn't significant competition between the capacity set at the outset and load in real-time. Yet, elsewhere in the experimental literature a paradigm involving a similar task structure to the one used in GN, reports different results. I refer to the studies on Prospective Memory (PM) which is the ability to remember to perform an intended action or thought, at some future time point (Einstein et al., 1997). PM lab-based tasks (such as the one described in Chapter 5, see Figure 1.1) show errors called Prospective Memory failures (PMf) which appear similar to GN. Experiments on PM suggest that forming an intention for a future action before the start of the task interacts with the complexity of the ongoing task (Marsh and Hicks, 1998; Marsh et al., 2002; Harrison et al., 2014). Considering the similarities of the GN and PM task structures, how do we reconcile these findings? How are PMf different from GN, if at all? Alternatively, is GN also similarly sensitive to certain kinds of ATC but these have simply not been tested yet? What do these differences/similarities suggest about the underlying capacities and/or mechanisms? I refer to the PM literature throughout this dissertation with a particular focus on these questions in Chapter 5.

It is worth mentioning here, how a similar ambiguity to the GN-PMF distinction, is found between GN and errors in multitasking. At the descriptive level, errors in multitasking appear to be a variant of GN in as much as multitasking requires switching across multiple goals/sub-tasks. However, data appear to suggest that GN and multitasking errors reflect different cognitive resources. For example, the Computerised Multiple Elements Test, CMET (Hynes et al., 2015; Cullen et al., 2016) involves 4 computerised games, each with different task rules and, importantly, requires the participant to engage in all of 4 games within a limited time. Despite knowledge of the overall goal (engage in at least all 4 games), participants may show impaired performance by persevering with a subset of these games. This dissociation between knowledge and action appears to be a

form of GN and yet, this perseveration was not found to correlate with GN as measured by the Feature Match Task (Hynes et al., 2015). Instead this perseveration correlated with errors on other standard tasks of multitasking/goal management (Cullen et al., 2016) such as the Six Elements Task and the Hotel Task (for more info on these tasks see Shallice and Burgess, 1991; Wilson et al., 1996; Manly et al., 2002). Although the null result between GN and the errors on the CMET should be read with caution, other studies have also shown a dissociation between GN and multitasking. For example, GN was found to correlate to fluid intelligence and to associate to a fronto-parietal network, whilst multitasking was not typically predicted by fluid intelligence and was instead associated to a fronto-polar brain region (BA10) and other default mode network areas which are largely orthogonal to the fronto-parietal network (Shallice and Burgess, 1991; Roca et al., 2010; Roca et al., 2011; Crittenden et al., 2015). One possibility that might explain the difference in effects between GN and multitasking is that in multitasking experiments the participant often needs to spontaneously decide to switch across sub-tasks, instead in GN experiments the switch is often triggered by an external cue. This is related to the difference between time-based and event-based PM experiments which is further discussed in Chapter 5 (p132).

1.6 PREFACE TO THE NEXT CHAPTERS

In the next chapters, I attempt to add to our understanding of the mechanisms underlying goal-directed behaviour by investigating how different forms of complexity affect performance failures with a focus on the error of GN. In Chapter 2, I develop a new task to test GN and, unlike previous studies, I manipulate complexity qualitatively by changing the semantic transparency of the instructional cue. In Chapter 3, I propose a new entropy-like measure to quantify the temporal clustering, or "clumpiness", of GN and use this to test between competing accounts of GN. In Chapter 4, I consider quantitative, qualitative and temporal dimensions of complexity manipulations, with the aim to clarify if the previous lack of ATC effects on GN still stands, or whether these results were due to insufficient manipulations of complexity. In Chapter 5, I systematically manipulate task structure by morphing GN and PM paradigms together, in an attempt to understand how this kind of complexity affects performance failures and whether GN and PM

failures refer to the same phenomenon. Finally, in Chapter 6, I summarise and discuss the implications of the results described in the dissertation.

Chapter 2. THE EFFECT OF TRANSPARENCY OF THE INSTRUCTIONAL CUE

One property of GN is that it is immune to various manipulations of ATC (Actual Task Complexity) but significantly modulated by the complexity of the instructions which are specified early on in the experiment (see Duncan et al., 2008 for a review; Bhandari and Duncan, 2014). This chapter aims to further characterize GN by investigating the effect of complexity, not of the overall task, but specifically of the instructional cues – the cues instructing the participant to shift to a different rule set. One variation of the LMT is the type of symbols that were used as the second side-cue signalling to watch right or left of the stream of characters towards the end of each sequence. In the original LMT, these were plus "+" and minus "-" symbols representing "watch right" and "watch left" respectively. A later variant used greater-than ">" and smaller-than "<" arrows (Duncan et al., 2008) which, intuitively, are more closely related to the rule they represent. The degree of pre-existing association between a cue and the rule it denotes is referred to as the *transparency* of the cue. Unpublished results did not show any differences in the amount of neglect observed across these two versions of the LMT (Duncan et al., 2008). This seemed to suggest that the type of cue used was not critical for GN. In contrast, evidence from a GN study on children (Towse et al., 2007) and in particular, both PM and the task-switching literature, suggest that the transparency of the cue is important for performance.

Several PM studies have investigated the effect of transparency of the PM cue (McDaniel and Einstein, 2000; Marsh et al., 2003; McDaniel et al., 2004; Pereira et al., 2012). All of these studies revealed improved prospective remembering when the PM cue was strongly related to the action to be performed. Some PM theorists propose that correct PM

performance goes through a series of cognitive stages. These stages are: (i) recognizing the PM cue as critical, (ii) verifying that the PM cue meets all criteria for response, (iii) retrieving the PM response and iv) executing/coordinating the PM response (Marsh et al., 2003; for similar stages see Ellis and Freeman, 2008). The first stage is also referred to as the *prospective* component, the second and third phase together as the *retrospective* component. A suggested interpretation for the PM transparency effect is that non-transparent cues hinder the "retrospective" component of a PM intention whilst leaving the "prospective" component unaffected (Marsh et al., 2003, see also Meier & Zimmermann, 2015 in which, not transparency, but the response mapping is manipulated).

Task switching studies (see Jost et al., 2013 for a review) similarly indicate that cues with high transparency produce lower switch costs than ones with low transparency (Mayr and Kliegl, 2000; Arbuthnott and Woodward, 2002; Logan and Bundesen, 2004; Miyake et al., 2004). The focus of these task switching studies is on reaction times across different task transitions, instead of gross performance errors. Instead, my primary scope was to assess the effect of cue transparency on the quantity and quality of the observed errors.

This chapter describes two separate experiments which use two different GN tasks. In the first experiment a newly developed task was used which involved judging photographs of faces. This task had a within-subject design so that for each participant I had two measures of GN based on the transparency of the cues in the critical trials: high transparency and low transparency. From here onwards I refer to these as, Transparent and Nontransparent cues, respectively. To anticipate the results of the first experiment, I found that GN was significantly more likely to occur with Nontransparent cues. To confirm that this result generalised beyond the specific task used, I then ran a second experiment to replicate this transparency effect using a different task: an adapted version of the Letter Monitoring Task (Duncan et al., 2008). This second experiment used a between-subjects design, whereby participants were allocated to either the Transparent or Nontransparent condition.

A second aim of the current study was to investigate how individual differences of fluid intelligence and age mediate any potential transparency effects on GN. I expected to replicate the basic finding that GN is more likely to be observed in participants with low

fluid intelligence (Duncan et al., 1996; Duncan et al., 2008; Bhandari and Duncan, 2014). In addition, I further predicted that fluid intelligence would be more strongly associated to GN on trials with Nontransparent cues. The best measures of fluid intelligence typically require some form of novel problem solving (as cited in Duncan et al., 2008). Hence, given that a cue with low transparency would necessitate learning a novel association I expected this encoding process to rely more heavily on resources common with fluid intelligence. The interest in possible ageing effects stems from studies in PM. Considering the importance of intact PM functioning for independent living (e.g. remembering to take medication), age effects in PM studies have been subject to considerable attention and debate (Zeintl et al., 2007; Kliegel et al., 2016). Hence, I was interested to test how the degree of transparency of the instructional cue might interact with age, in terms of GN. Therefore, in both experiments described in this chapter, the relationships between age, fluid intelligence and GN were tested under different conditions of cue transparency.

2.1   EXPERIMENT 1 – THE SEMANTIC GOAL NEGLECT TASK (SGNT)

In this first experiment I designed a novel GN task in which I manipulated the transparency of the instructional cue within each participant. The task was composed of a dominant sub-task and two less frequent critical sub-tasks, each cued by either a Transparent or a Nontransparent cue. The errors on the critical sub-tasks were used to measure GN.

2.1.1   METHOD

2.1.1.1   *Participants*

Sixty-four individuals, who were not suffering from any psychological condition at the time of testing and had no history of neurological disorders, gave informed consent and completed the experiment in exchange for monetary compensation. They were recruited via a panel of volunteers at the MRC Cognition and Brain Sciences Unit. Selection of participants was such as to ensure a wide distribution of both age (28-70 years) and fluid intelligence scores (IQ 86-145). Age ranged from 28.7 to 70.9 years, Mean=56.9(SD=10.9), Median=60.9. Fluid intelligence ranged from 86 to 145 IQ points, Mean=108.1(SD=14.9), Median=108.1.

2.1.1.2   *General setup and tools*

Experiments were run using a standard desktop computer using Windows Vista. The viewing distance was not strictly controlled but was approximately 57cm such that 1cm subtended about 1 degree of visual angle and the screen size was approximately 30cm by 38cm. The stimulus presentation program was written in Matlab using Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007). R freeware (R Core Team, 2013) and SPSS (IBM Corp., 2013) were used for analyses.

2.1.1.3   *Task*

In summary, the SGNT required participants to make judgments of target photographs by making button presses, whereby the type of judgment to be made was instructed by the preceding cue. A trial consisted of a series of three displays: a cue, the target stimulus consisting of a photograph of a face, and a blank screen, lasting 400ms, 600ms and 1000ms respectively (Figure 2.1). The trials were of four types: regular, neutral and two types of critical trials (Figure 2.2). There were a total of 320 neutral trials, 32 regular trials and 32 critical trials: 16 of each subtype.
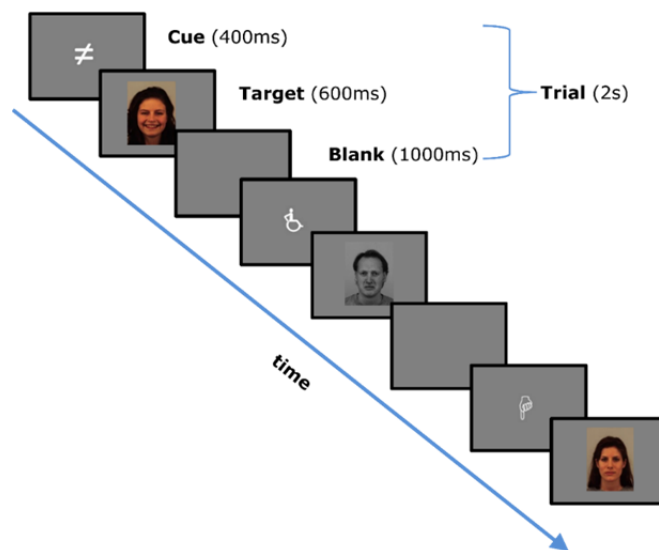
Figure 2.1 A trial in the Semantic Goal Neglect Task (SGNT).
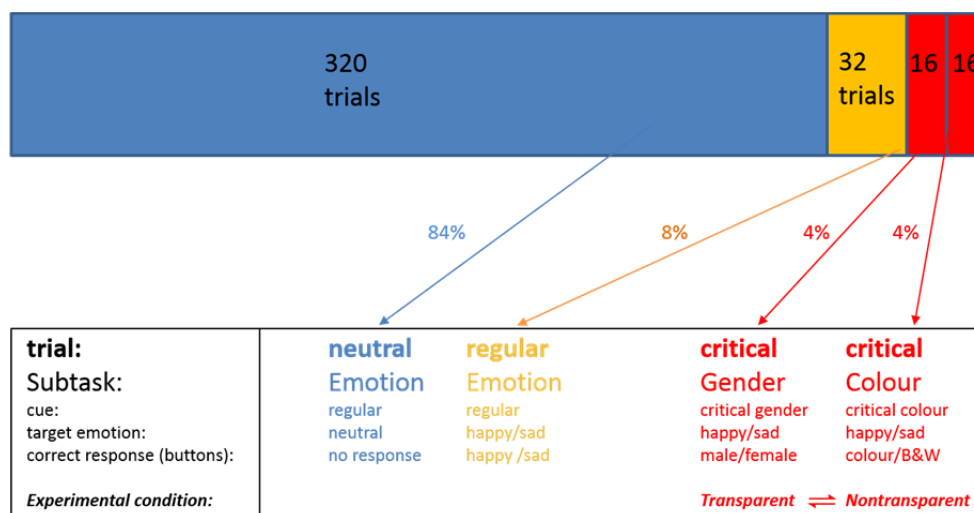
Figure 2.2 The SGNT trial types.

There were a total of 384 trials. The majority of trials (neutral and regular) were preceded by regular cues which instructed the participants to follow the Emotion sub-task. The remaining 8% were critical trials, half of the Gender sub-task type, and the other half, the Colour sub-task type. GN was defined as cases when participants responded according to the Emotion sub-task on a critical trial, hence neglecting the critical cue and treating it as a regular trial. Conditions of transparency (Transparent, Nontransparent) were allocated to each one of the two critical sub-tasks, thus resulting in two types of GN: Transparent and Nontransparent. This allocation was counterbalanced across participants. The effect of transparency of the critical cue on GN was tested by contrasting Transparent GN to Nontransparent GN within each participant.

Each cue instructed the participant which sub-task to apply to the subsequent target. Regular and neutral trials were preceded by regular cues, whilst critical trials were preceded by critical cues. Regular cues instructed participants to do the Emotion sub-task which involved judging the facial expression of the target as happy, sad or neutral by pressing the corresponding happy or sad button or nothing at all for neutral. Instead, critical cues required participants to judge the targets based on other features of the photograph. There were two critical cues – Gender and Colour. If the Gender cue occurred, then the participant had to judge the gender of the target by pressing either the male or female button (Gender sub-task); for the Colour cue, the participant had to judge the colour of the photo by pressing either the colour or black & white button (Colour sub-task).

The experimental manipulation in the task occurred via critical trials only. Conditions of transparency (Transparent, Nontransparent) were allocated to each one of the two critical sub-tasks. This allocation was fixed throughout the task for each participant and counterbalanced across participants so that half the sample did the Gender sub-task in the Transparent condition and the Colour sub-task in the Nontransparent condition. The reverse combination was true for the other half of the sample. GN was defined as cases when participants responded according to the Emotion sub-task on a critical trial, hence neglecting the critical cue and treating it as a regular trial. Therefore, each participant had two different scores of GN, one in the Transparent and one in the Nontransparent condition. To test the research question of whether transparency of the instructional cue affected GN, these two GN scores were compared to each other.

Each trial type had a specific function and accordingly, specific features. As previously mentioned, critical trials were used to gauge GN in the two conditions of transparency. The targets' emotions in critical trials were always either happy or sad, which was an irrelevant target feature if the participants correctly applied the appropriate critical sub-task rules (Gender/Colour). If, however, they responded with a happy or sad button press then this was taken as evidence that they neglected the critical cue and persevered with the Emotion sub-task rule. In other words, neutral targets were intentionally omitted in critical trials because in such cases, neglect (i.e. executing the Emotion sub-task rule during critical trials) would lead to omissions which are generally hard to interpret. Similarly to critical trials, regular trials' target emotions were always either sad or happy.

Thus both critical and regular trials always required a button press and were therefore collectively referred to as response trials (Figure 2.3). Accuracy on regular trials was used to assess baseline task performance. Neutral trials were similar to the regular trials, except that the target emotions were always neutral thus requiring withholding of response. The function of neutral trials was to avoid performance failure on response trials due to rapid task switching by providing gaps of no responses. This padding varied between 3-10 trials. Thus, after a response trial, there were at least 6 seconds before the onset of the next response trial. The variable numbers of neutral trials were selected at random but fixed across participants.
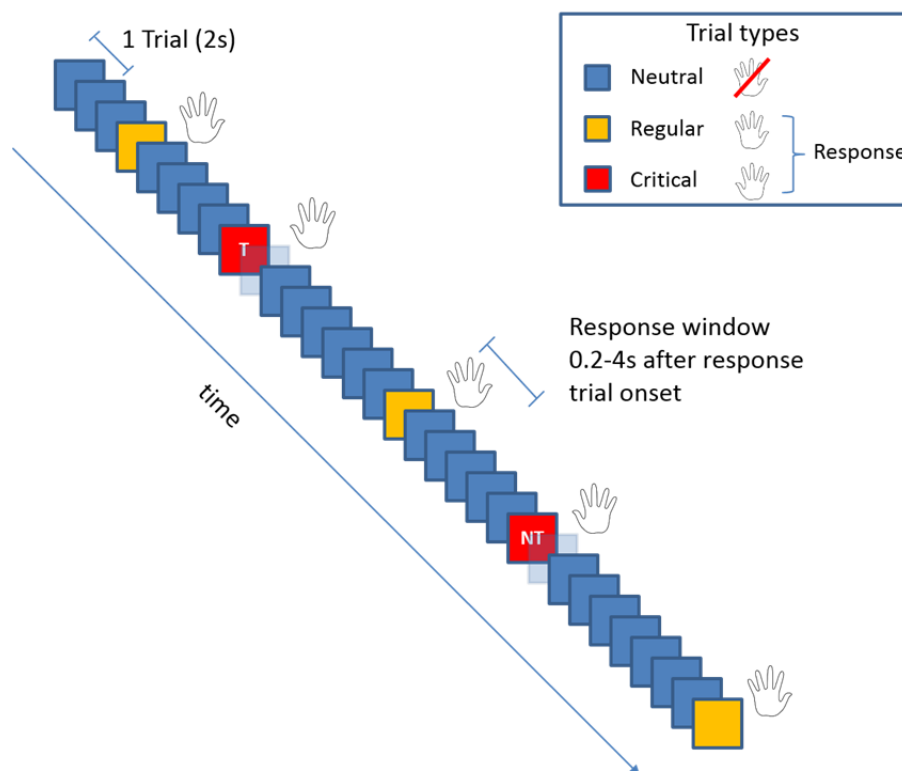
Figure 2.3 SGNT task structure.

Regular and critical trials required a button press response and hence are collectively referred to as response trials. Response trials were separated by sequences of non-response neutral trials with a variable length of 3-10 trials. The sequence of response trials alternated between regular and critical trials and started with a regular trial. The sequence of critical trials alternated between Transparent (T) and Nontransparent (NT) trials. Button presses within a response window of 0.2-4s after each response trial onset were analysed in this experiment.

The order of trial types was balanced. The sequence of response trials always started with a regular trial and alternated between regular and critical response trials (Figure 2.3). The order of critical trials alternated between Transparent and Nontransparent cue conditions, and which of these appeared first was counterbalanced across participants.
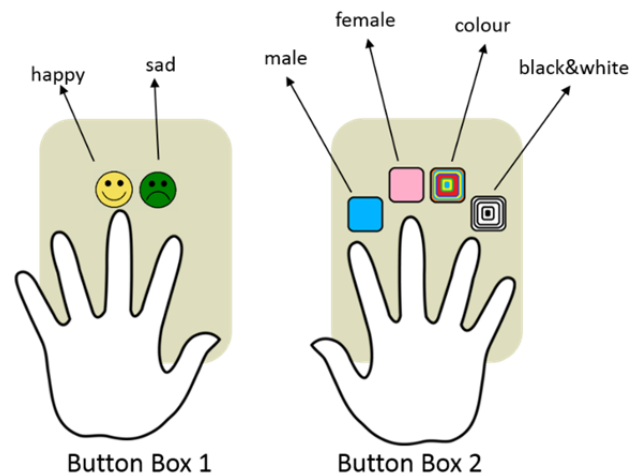
Figure 2.4 Button boxes for the SGNT task.

The non-dominant hand was used with button box 1 which coded for the Emotion sub-task with a happy and sad button. The dominant hand was used with button box 2 which coded for the critical sub-tasks: Gender (male and female buttons) and Colour (colour and B&W buttons). To discourage participants from looking away from the screen during the experiment, none of the buttons had any labels on them. Instead, the button-response mapping was displayed at the bottom of the screen throughout the task. The allocation of the button labels of box 2 was counterbalanced across participants to avoid biasing performance across the critical sub-tasks as follows: if instructions to the Colour sub-task were introduced first then the left pair of buttons on the critical response button box were allocated to the Colour sub-task and the right pair to the Gender sub-task; the inverse response-button mapping was used if the Gender sub-task was explained first.

2.1.1.4   *Apparatus and stimuli*

Responses were collected using two button boxes connected via USB (Figure 2.4). All button presses were recorded, however, only button presses made within a response window of 0.2s to 4s from the onset of a response trial were analysed. Any button presses outside of this window were false positives during neutral trials (Median=28.5 false positives in the whole sample), which are not of primary interest in this study and hence ignored. The first button box had 2 buttons which coded for the Emotion sub-task: "happy" and "sad". The other button box coded the two critical sub-tasks, Gender and Colour. This second button box had 4 buttons: "female", "male", "colour" and "black and white". Participants were asked to use their dominant hand with the second button box. None of the buttons had any labels on them to discourage participants from looking away

from the screen during the experiment. Instead, the button-response mapping was displayed at the bottom of the screen throughout the task.

The target stimuli were made up of 162 photographs of human faces. These were obtained from the Karolinska Directed Emotional Faces (KDEF) database (Lundqvist et al., 1998). Each measured 10 degrees in width and 13.56 degrees in height. The photos were of 54 actors, half of whom were female. For each actor there were three photos each showing one of three possible emotional states: happy, sad and neutral. Each photo was duplicated into colour and greyscale versions.
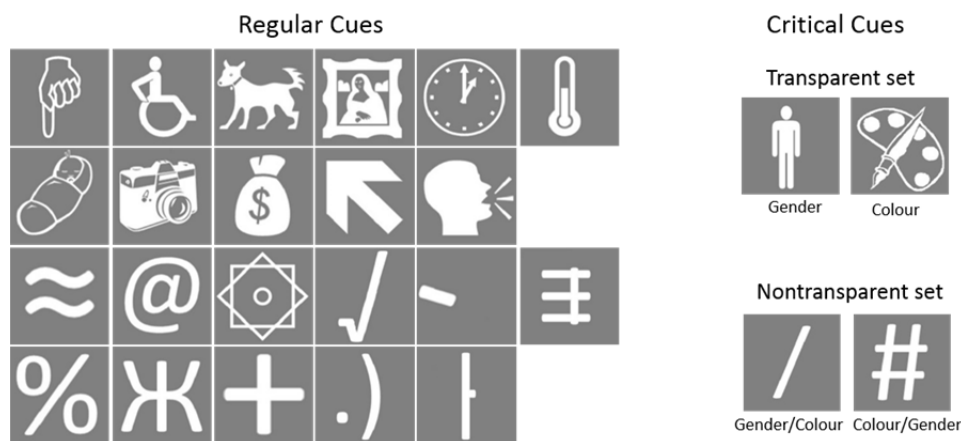
Figure 2.5 SGNT cues.

Regular cues were randomly selected from a pool of 22 symbols, each appearing 16 times in total. Each participant saw only 2 critical cues, one from the Transparent set and another from the Nontransparent set. This was fixed throughout the task and the different combinations of critical cues and conditions of transparency were counterbalanced across participants.

The cue stimuli were made up of a total of 26 white symbols (Figure 2.5). A single cue was displayed alone in the centre of the screen against a grey background. Each symbol measured approximately 2.2 x 2.2 degrees. For regular cues, the symbol was randomly selected from a pool of 22 different symbols (Figure 2.5, left), each appearing 16 times throughout the task. For critical cues, there was a separate pool of 4 symbols (Figure 2.5, right) and each participant saw only a pair of these symbols throughout the task. The pair was made up of one symbol from the Transparent set and another from the Nontransparent set. The Transparent set consisted of a symbol of a man for the Gender

sub-task, and a symbol of a paint palette for the Colour sub-task. The Nontransparent cue set was made up of a hash tag and a forward slash symbol. To avoid the critical cues from appearing salient with respect to the regular cues, half of the regular cues were of the pictorial type to match the Transparent cue, whilst the rest were abstract shapes or logical operators to match the Nontransparent cue. During instructions, single syllable words ("paint", "man", "slash", "hash") were used to refer to the critical cues to avoid biasing the retrieval of the critical rules because of any potential "word length effect" (Baddeley et al., 1975; Campoy, 2008).

2.1.1.5 Procedure

Instructions were introduced in a step-by-step manner. During instructions, participants were invited to give responses at each step, in this way serving as a form of practice. The Emotion sub-task instructions were explained first, followed by the instructions to the Gender and Colour sub-tasks. Which of the two critical sub-tasks was explained first was counterbalanced and this order was congruent with the response-button mapping (for details see Figure 2.4). The order of transparency conditions was independently counterbalanced to the instructional order of the critical sub-tasks. In addition, it was specified that when a critical cue occurred, the corresponding critical rule only applied to the target photo which immediately followed and subsequently they had to switch back to the Emotion sub-task. Participants were instructed to respond as quickly and as accurately as possible. At the end of the instructions, there was an initial recall phase in which a series of questions assessed the participant's comprehension and memory of the task rules including the critical cue-response mapping. The experimenter corrected the participant if necessary, in which case the entire cued recall was repeated until all components were remembered correctly. The instructions were presented once for all participants with the exception of 2 participants for whom the cued recall was performed twice. No further feedback was allowed after this stage. A card displaying the two critical cues was left in full view during the whole running of the task as a memory aid. After task execution, the display card was removed, followed by a second recall phase. All participants remembered the entire set of rules correctly.

The Culture Fair test of fluid intelligence, Scale 2 Form A (Cattell, 1971; Cattell and Cattell, 1973) was administered at the end of each testing session, unless a previous score not older than 5 years was already available.

2.1.2   Results and discussion

The dependent variable in this task is a form of error referred to as GN. The operational definition of GN in this dissertation is the event in which the response to a critical trial was a button press for the Emotion sub-task. In other words, GN is when, during a critical trial instead of making a judgment on the colour of the photo or the gender of the face, a judgment on the emotion of the face is made by pressing the happy or sad button. A Wilcoxon Signed-Rank paired test revealed that GN for critical trials with Nontransparent cues (Median=3.0) was significantly higher than for critical trials with Transparent cues (Median=0.0), W=128.0, z=4.74, p<.0001, r=.46 (Figure 2.6). This is the key result of this experiment and suggests that, in this task, the transparency of the critical cue had a significant and moderately strong effect on the frequency of GN.
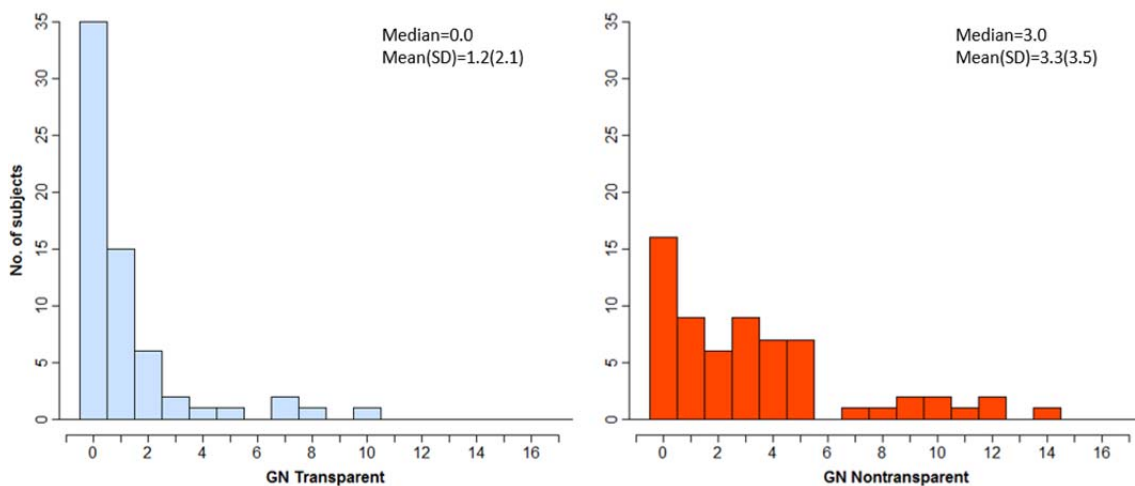
Figure 2.6 Distribution of GN in the critical sub-tasks.

In the 64 participants tested on the SGNT task, significantly more GN was observed in the Nontransparent critical cue condition (red) compared to the Transparent condition (blue). GN is when on critical trials, the participant presses the happy or sad button instead of one of the critical response buttons.

As predicted, a significant association was found between total GN and fluid intelligence ρ=-.24, p=.03, 1-tailed (Figure 2.7, left). However, this association was less robust than previous findings in two ways. Firstly, using Fisher's r-to-z transformations (Preacher, 2002), this relationship was found to be significantly weaker than in previous experiments (p<.01; comparison with Experiment 2 which uses a task that is closest in design to the

SGN task, with n=65, r=-.66, from Duncan et al., 2008). Secondly, this association was removed when age was controlled for, reducing the coefficient from ρ=-.24 to ρ=-.12, which is not significant p=.18(1-tailed).
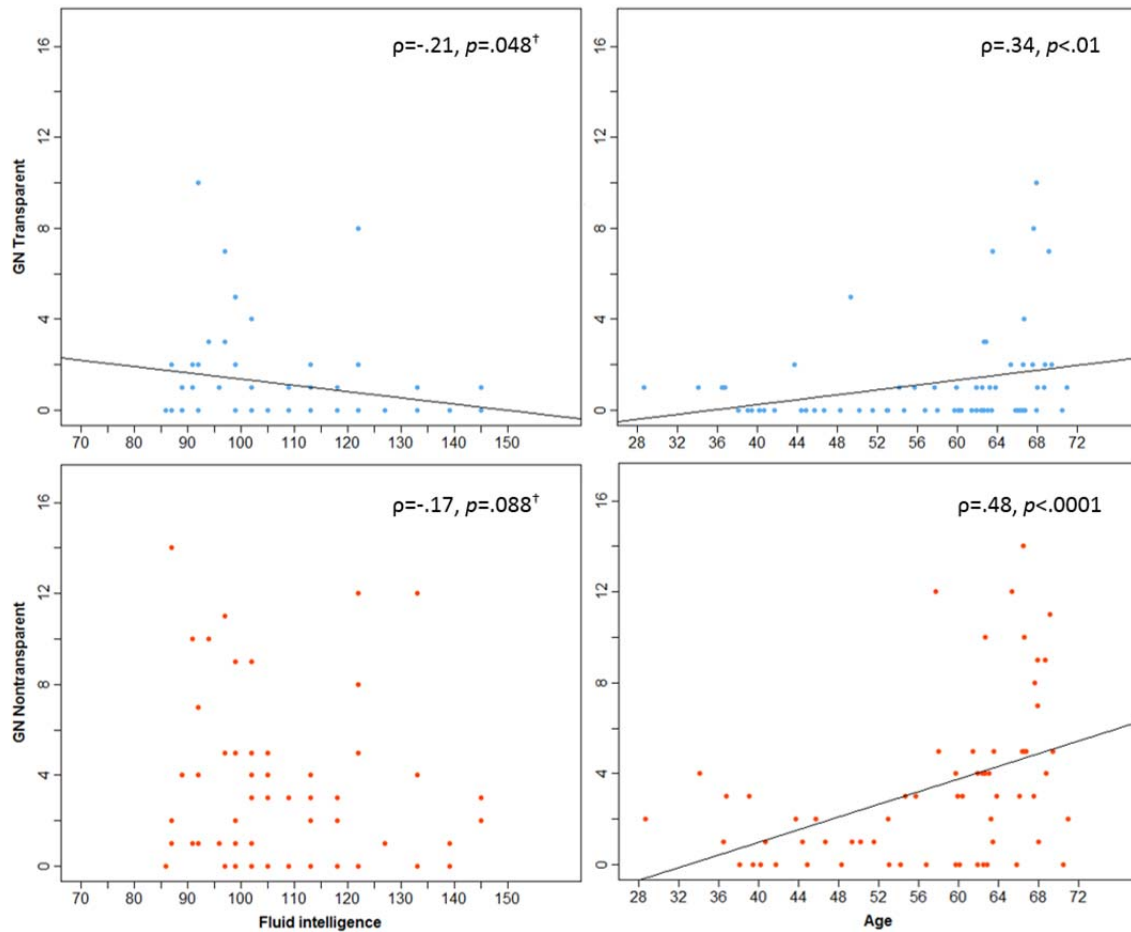
Figure 2.7 Correlations of GN with Fluid intelligence (left column) and Age (right column)

Associations are split according to the condition of the transparency of the critical cue, in blue in the top row is GN in the Transparent condition and in red, bottom row, the Nontransparent condition. [†]Correlations with fluid intelligence are 1-tailed because the hypothesis was directional. Significant correlations (p-value <.05) are marked with a plot line.

A second prediction was that the relationship between fluid intelligence and GN would be modulated by the condition of transparency so that it would be stronger in the Nontransparent cue condition. To test this, I correlated fluid intelligence to a transparency-contrast score. Scores on the Culture Fair test A were converted to IQs using the manual (Cattell and Cattell, 1973) with a mean of 100 and a standard deviation of 16. The IQ scores were normalised on adolescents (Cattell and Cattell, 1973) and are not age-corrected but are absolute values. The transparency-contrast score was calculated as the

difference in GN between Transparent and Nontransparent condition per participant. The correlation between this contrast score and fluid intelligence was not significant $\rho=.01$, $p=.94$ (2-tailed) even after partialling out age $\rho=.06$, $p=.62$ (2-tailed) (Figure 2.8, left). This suggested that the transparency of the cue did not interact with the association between fluid intelligence and GN. Instead, GN was significantly predicted by age $\rho=.52$, $p<.0001$ (2-tailed) and this strong association was preserved even after controlling for fluid intelligence $\rho=.51$, $p<.0001$ (2-tailed). When GN was split according to the condition of transparency, this positive relationship to age was found in both conditions: $\rho=.34$, $p<.01$ (2-tailed) for the Transparent critical cue trials and $\rho=.48$, $p<.0001$ (2-tailed) for the Nontransparent condition (Figure 2.7, right). To test if the relationship between age and GN was modulated by the condition of transparency, a within-subject transparency-contrast score was correlated to age. This correlation was significant $\rho=.25$, $p=.047$ (2-tailed) even after partialling out fluid intelligence $\rho=.26$, $p=.038$ (2-tailed) (Figure 2.8, right). This suggests that the cognitive resources involved in the Nontransparent condition are particularly vulnerable to aging.
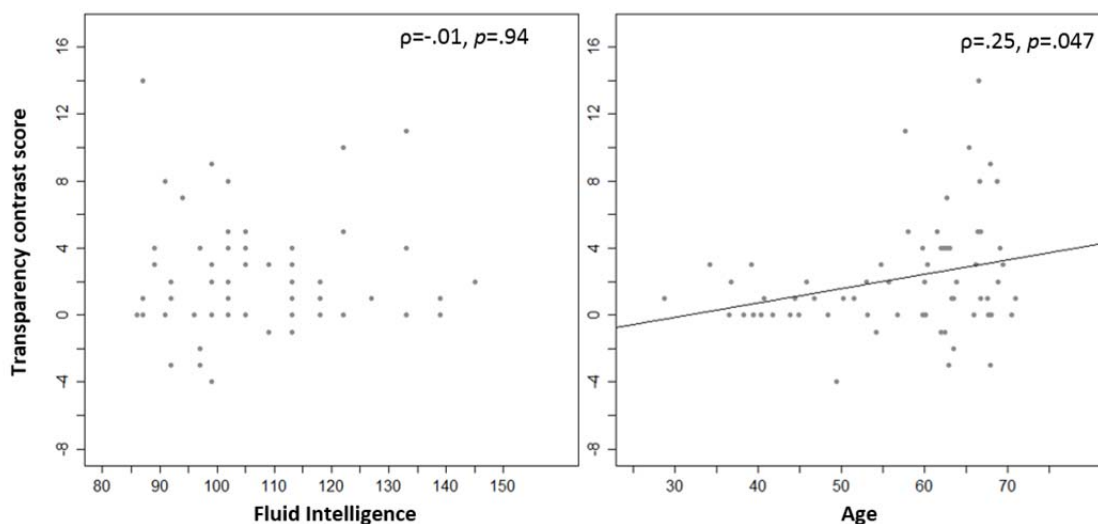
Figure 2.8 Condition of transparency on critical trials is sensitive to age

For each participant, the number of GN cases in the Transparent condition is subtracted from the Nontransparent condition resulting in a Transparency contrast score (y-axis) which is correlated to Fluid Intelligence or Age (x-axis). Only the association with age is significant suggesting that older participants find the Nontransparent condition disproportionally more difficult than the Transparent condition, compared to younger participants.

Why are participants neglecting more Nontransparent than Transparent critical trials? There are a series of possible explanations which I outline below, with the first "failing to notice" theory being the most compelling. This "failing-to-notice" explanation is derived from an analysis aimed at testing if the observed transparency effect was driven by participants intending to execute the Nontransparent sub-task, but were unable to suppress the dominant Emotion sub-task. If recognition of the cue did occur and assuming this was immediate, then I would expect that any aspect of cue processing (e.g. recognition and/or rule retrieval and/or initiation of response) would compete with the execution of the incorrect Emotion sub-task response (neglect) and slow down this response. West and Craik (1999) defined this as cue sensitivity. One way to test cue sensitivity, is to compare reaction times on correct Regular trials to reaction times on neglected Transparent and Nontransparent critical trials. Two Wilcoxon Signed-Rank tests were run and these revealed that in both cases reaction times were not significantly different from correct Regular trials : W=175.0, Z=-0.92, r=-0.12, p=0.36 (Regular vs. Transparent), and : W=628.0, Z=-0.41, r=-0.04, p=0.68 (Regular vs. Nontransparent) see Figure 2.9 and Figure 2.10.
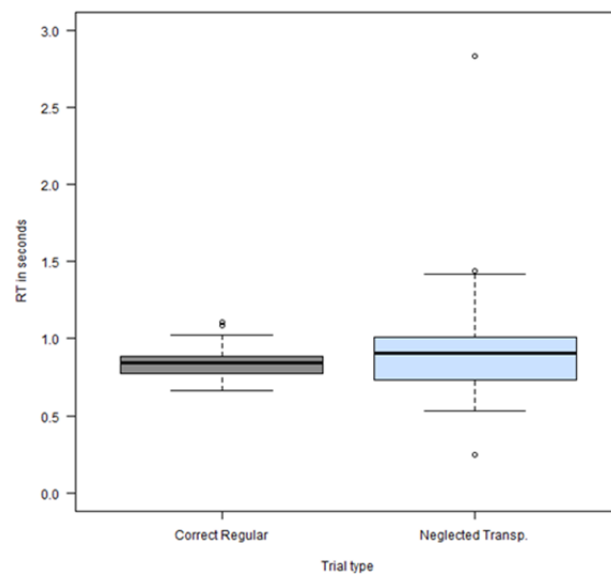
Figure 2.9 Boxplot for reaction time (RT) on correct Regular trials and neglected Transparent critical trials (n=29)

RTs for Correct Regular trials: Median=0.846, Mean(SD)=0.847(0.118).
RTs for Neglected Transparent critical trials: Median=0.904, Mean(SD)=0.961(0.439).
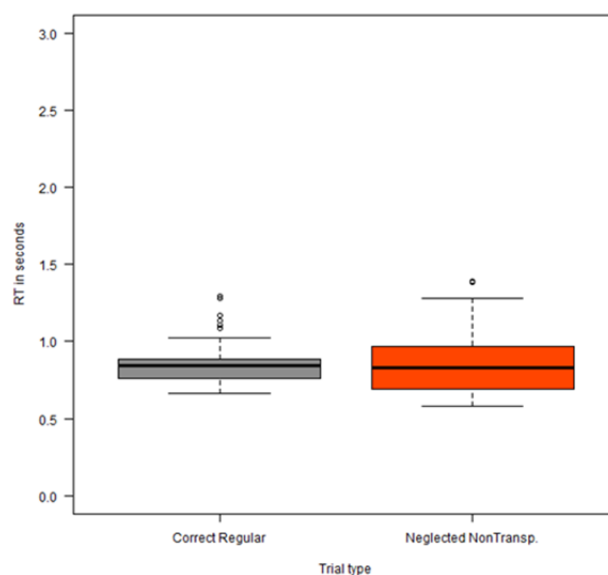
Figure 2.10 Boxplot for reaction time (RT) on correct Regular trials and neglected Nontransparent critical trials (n=48)

RTs for Correct Regular trials: Median=0.843, Mean(SD)=0.858(0.147).
RTs for Neglected Nontransparent critical trials: Median=0.830, Mean(SD)=0.855(0.197).

This cue sensitivity result suggests that in addition to the keypresses being the same, the neglected critical trials were not different from correct Regular trials even in terms of latencies. While the interpretation of null results should be made with caution, this result may indicate that when participants neglected a critical rule they treated the trial as if it were a regular one and were insensitive to the cue. In contrast, preserved cue sensitivity would have indicated that the critical cue was processed enough to elicit the start of a critical response which competes and slows down the regular response. Interestingly, there is a trend for some preserved cue sensitivity in neglected Transparent critical trials (Median=0.904s), but not in the Nontransparent condition (Median=0.830s). Although speculative, this could suggest that Transparent cues are more likely to be processed as significant events than Nontransparent cues.

A potential alternate explanation to the results of the cue sensitivity analysis is delayed awareness of the significance of the critical cue. Rather than immediate, recognition of the cue may have been delayed. In such cases it is possible that processes related to the critical sub-task response may have either not occurred at all or, may have been delayed enough to avoid much temporal overlap, and hence interference, with the incorrect Emotion sub-task response. Delayed processes may have resulted with participants attempting to correct their GN responses. However, it is worth noting that there were very few trials that may have been classified as "corrections" i.e. two successive responses with the first being a incorrect regular response (GN), whilst the second being a correct regular response (correction). In the SGNT task, any case in which both a correct critical response and an incorrect regular response were made within the same response window were classified as "potential neglect" but were not counted as cases of GN. These "potential neglect" cases are rather mixed in type, such that they may involve multiple presses (not just 2), have critical before regular, or regular before critical responses etc... The total frequency of such cases was of 6% of the total number of critical trials across the whole sample. Given that cases of "corrections" are a subtype of the "potential neglect" category, this implies that there were less than 6% cases of "corrections". Overall, given that the frequency is rather low (<6%) I would suggest that this possible explanation is unlikely to explain GN.

Another possibility is that in Nontransparent critical trials, participants correctly realize that they are meant to apply the relevant Nontransparent sub-task rules, but this processing is more demanding than in Transparent critical trials. West and Craik (1999) referred to this as cue accessibility, defined as the amount of processing required to elicit a correct response. One proposed way to test this, is to compare reaction times on correct trials (mean within condition) across the 3 different trial types, Regular, Transparent and Nontransparent. A Friedman rank sum test revealed that these reaction times were significantly different from each other $\chi^2(2)=87.2$, $p<.00001$ (Figure 2.11). Post-hoc tests were used with Bonferroni correction applied. It appeared that Regular vs. Transparent trials (difference=72), Regular vs. Nontransparent trials (difference=102) and Transparent vs. Nontransparent trials (difference=30) were all real differences. In all cases, the critical difference ($\alpha=.05$ corrected for the number of tests) was 26.9. Pairwise comparisons using Wilcoxon signed-rank tests were all significant with Holm correction applied: Regular vs. Transparent $W=20.0$, $z=-6.82$, $r=-.60$, $p<.0001$, Regular vs. Nontransparent $W=1.0$, $z=-6.89$, $r=-.61$, $p<.0001$ and Transparent vs. Nontransparent $W=479.0$, $z=-3.62$, $r=-.32$, $p<.001$. These results suggest that processing Nontransparent trials was more difficult than Transparent trials which in turn was more difficult than Regular trials. A Steiger test (Lee and Preacher, 2013) which compares the size of coefficients, revealed that the difference was larger between Regular and critical trials, than between the two types of critical trials, $p=.05$.
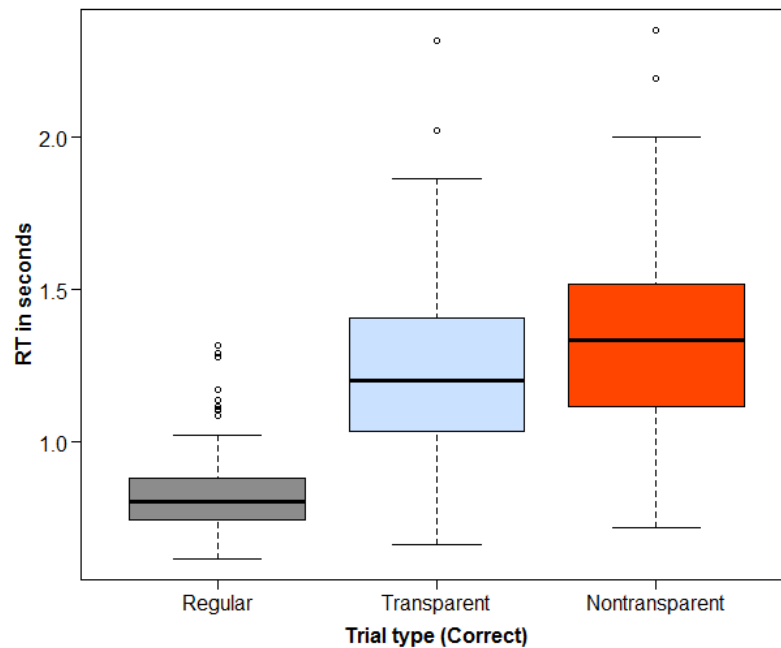
Figure 2.11 Boxplots for reaction time (RT) on correct trials

RTs for Correct Regular trials: n=64,   Median=0.807, Mean(SD)=0.849(0.158).
RTs for Correct Transparent critical trials: n=64,   Median=1.216, Mean(SD)=1.249(0.343).
RTs for Correct Nontransparent critical trials: n=63^, Median=1.336, Mean(SD)=1.356(0.350).

Tests revealed that reaction times on correct Regular trials are faster than ones on correct Transparent critical trials which are faster than reaction times on Nontransparent critical trials. The magnitude of these differences is larger between the Regular compared to both critical trials, than the difference between the two critical trials. ^One participant neglected or omitted all Nontransparent critical trials hence no RT score was available.

Does cue accessibility explain the age effect? To test this, I ran correlations between age and reaction times on correct trials on the 3 different trial types (Figure 2.12) and the magnitude of the coefficients were compared. If the strength of the correlation between age and reaction time is stronger in the Nontransparent condition compared to the Transparent condition then this would suggest that older participants are significantly slower at processing the Nontransparent critical trials. The effect was in the opposite direction than predicted, with the strength of the association in the Transparent condition (p=.48) found to be stronger than in the Nontransparent condition (p=.39). However, three Steiger's tests revealed that none of the coefficients were significantly different from each other: Regular vs. Transparent, Regular vs. Nontransparent and Transparent vs. Nontransparent, p=.32, p=.62, p=.62 (Holm corrected). Hence, there is no evidence to

suggest the interaction between age and transparency could be explained in terms of cue accessibility across trial types.
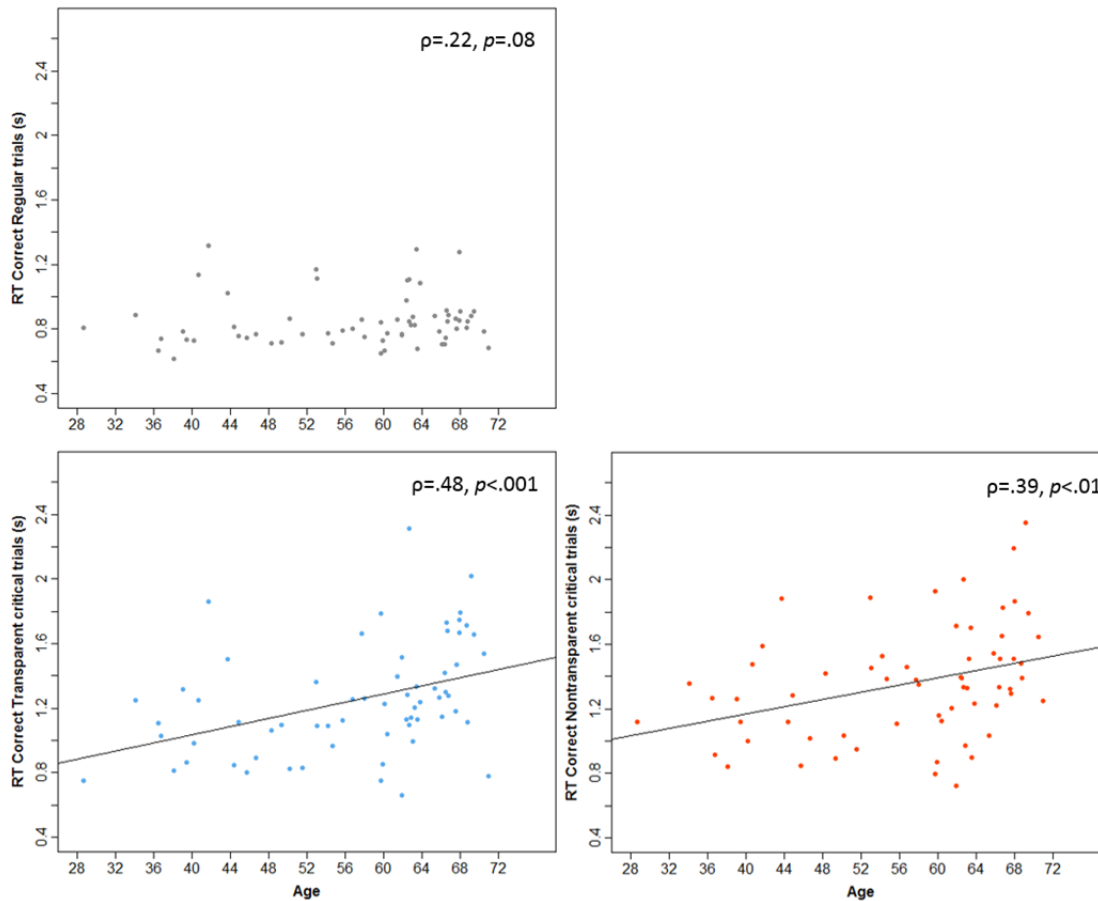
Figure 2.12 Correlations of age and reaction time on all correct trial types

Scatter-plots of RT against age for Regular trials (top row, grey), Transparent (bottom row, blue) and Nontransparent trials (bottom row, red). Plot lines indicate that the spearman correlations are significant at p < 0.05 and survive Holm correction for multiple comparisons.

Finally, an additional potential explanation for participants performing worse on the Nontransparent critical trials is that access to rule content or execution of the Nontransparent rule is sufficiently difficult for participants to give up, leading to an omission. However, the data suggest that this explanation is unlikely. Firstly, the measure of GN in this experiment did not include omissions. Secondly, Steiger tests reveal that transparency does not modulate any association (all *p*>0.05) between age and omissions on critical trials (Figure 2.13).
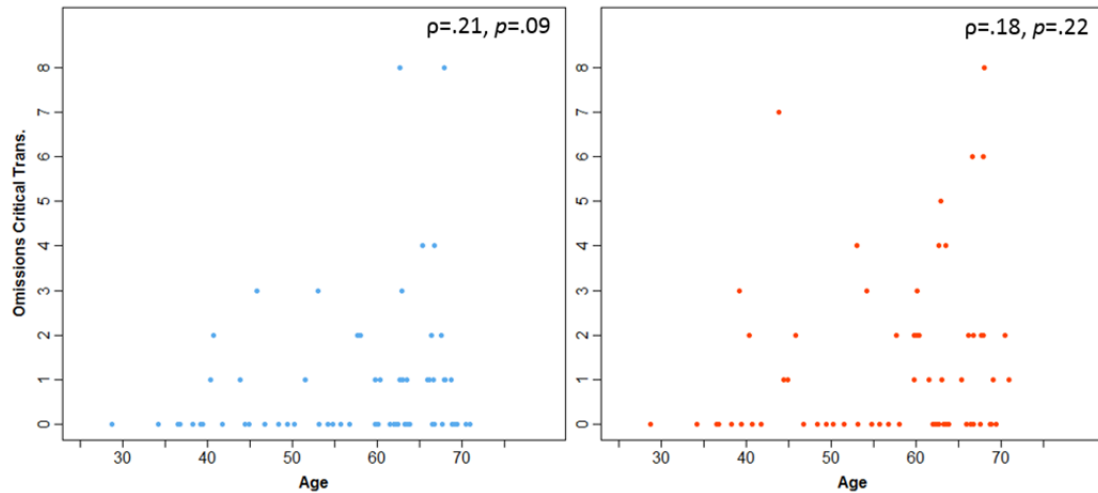
Figure 2.13 Correlation between age and omissions across transparency.

On the left the correlation between age and omissions in the Transparent condition (blue) and on the right In the Nontransparent condition (red). Both associations are not significant (2-tailed) and Steiger tests reveal that the coefficients are not significantly different from each other (p=.86; 2-tailed).

2.1.3 CONCLUSION

The main finding of this experiment is that the transparency of the critical cue significantly modulates GN. When the critical cue is Transparent, GN may occur, however if the critical cue is Nontransparent then GN is reliably more likely to be observed. Reaction time analyses, in terms of cue accessibility and cue sensitivity, shed further light on the observed neglect. The cue sensitivity analysis suggested that when participants neglected they treated the critical trial as if it were a Regular trial – failing-to-notice the importance of the critical cue. The cue accessibility analysis revealed that for correct performance, Regular trials were easier to process than Transparent trials which were in turn easier than Nontransparent trials. To conclude, the data suggest that transparency effect on GN may be driven by a failure to recognize the significance of the Nontransparent critical cues.

In terms of individual differences, age was found to be strongly associated with GN, such that older participants neglected more than younger participants. In addition, older participants were found to be disproportionally worse at the Nontransparent critical trials suggesting that the mechanisms involved in processing these trials were more vulnerable

in the older participants. However, there was no evidence to suggest that older people were disproportionately slower in processing Nontransparent critical trials.

Contrary to our hypothesis, fluid intelligence only weakly predicted GN, and only in the Transparent condition. The latter correlation did not survive when age was accounted for. Given that this relationship was significantly smaller than previously reported using traditional GN tasks such as the Letter Monitoring Task (Duncan et al., 2008), one possibility is that the transparency effect may be specific to the SGNT task and not generalizable to other GN tasks. We therefore went on with a second experiment to test whether the transparency effect would replicate in a standard test of GN.

2.2 EXPERIMENT 2 – THE LETTER MONITORING TASK (LMT)

2.2.1 INTRODUCTION

The main aim of this experiment was to test whether the effect of transparency found in Experiment 1 using the SGNT task would replicate in a different GN task. The task used in this experiment was a modified version of the LMT which was previously found to show no effect of transparency in unpublished experiments (Duncan et al., 2008). In these experiments two versions of the LMT were used, one version involved a plus and minus symbol to represent the rules right and left respectively, whilst the other used relatively more transparent cues for the same rules: right- and left-pointing arrows. GN across these two versions of the LMT was found to be similar, despite the change in degree of transparency. In Experiment 2, I sought to formally test the effect of transparency on GN in the LMT by modifying this task such that the contrast between Transparent and Nontransparent cue conditions was maximized. The manipulation of cue transparency in this experiment was between-subjects, to avoid altering the structure of the task. In addition to an effect of cue transparency, I expected to find a strong negative relationship between GN and fluid intelligence in the LMT, as reported in several other experiments (Duncan et al., 1996; Duncan et al., 2008). Furthermore, given that Experiment 1 revealed a moderately strong relationship between age and GN, I predicted a similar finding in the following experiment using the LMT.

2.2.2 METHOD

2.2.2.1 Participants

Fifty-eight participants were recruited in the same manner as in Experiment 1 (p.45). Age ranged from 50.1 to 73.0 years, Mean=61.0 (SD=5.54), Median=61.75. Fluid intelligence ranged from 73 to 145 IQ points, Mean=104.8, Median=100.5. Half of the participants were allocated to the Transparent cue condition and the other half to the Nontransparent cue condition. The two groups did not differ in terms of age $t(56)=-.88$, $r=.12$, $p=.38$ or fluid intelligence scores $t(56)=-.71$, $r=.09$, $p=.48$ (Figure 2.14).
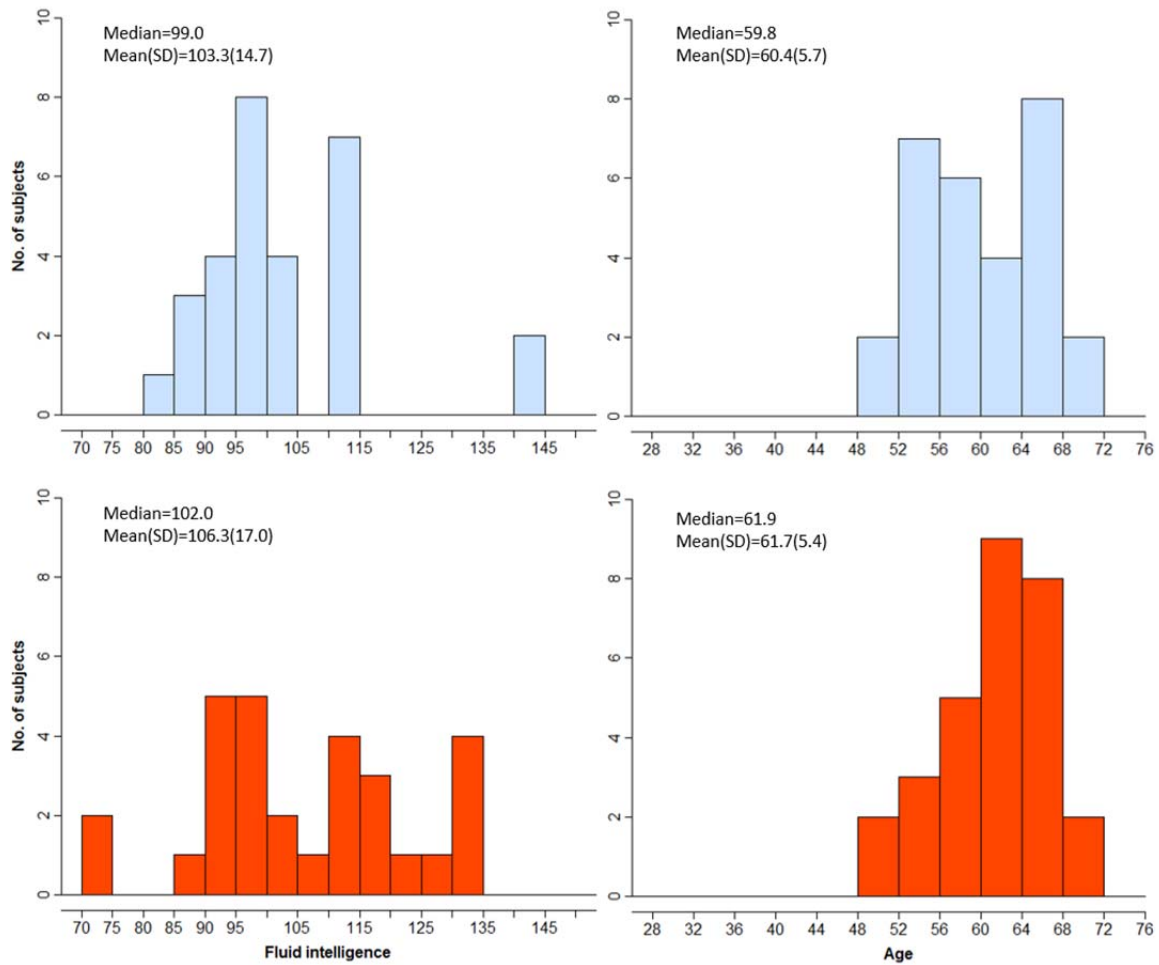
Figure 2.14 Distributions of age and fluid intelligence in the LMT

Age and fluid intelligence (as measured by Cattell's Culture Fair test) were similar across conditions. Transparent condition top row in blue, Nontransparent condition bottom row in red.
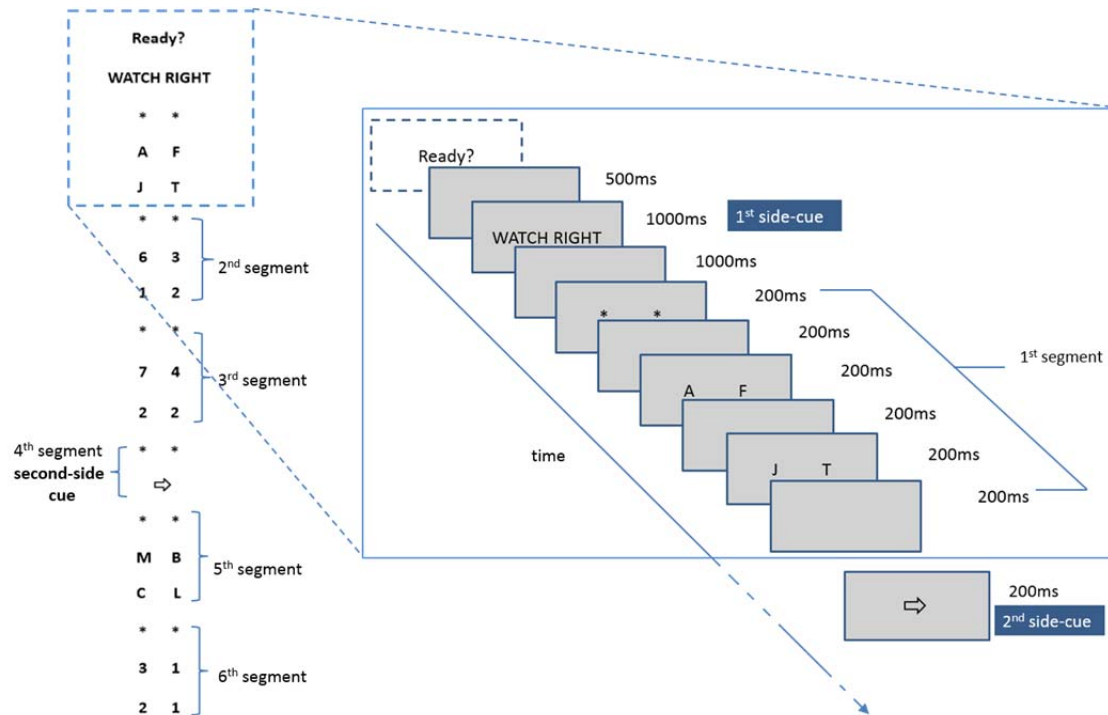
Figure 2.15 A trial in the LMT.

On the left, is a sample of the sequence of stimuli in one trial out of a total of 16, in the LMT. A trial starts with the first side-cue which contains a verbal instruction and is followed by a sequence of frames organized in 6 segments. On the right are details of the durations of the displays. A letter or number segment is made up 6 frames: an asterisk display and two letter (or two number) displays each preceded by a blank interval. Participants had to attend to the side they were instructed to by the side-cues, and on that side read letters aloud or say the sum of numbers within that segment. For example, in this trial the correct response is "F,T, 5, 6, B, L, 2".

2.2.2.2 Task

The LMT was closely modelled to the mixed-trial version used by Duncan and colleagues (2008; see their Figure 3, Panel C). Participants were told they would see a stream of characters appear one pair at a time in the centre of the screen (Figure 2.15). They were instructed to attend one side, left or right, and ignore the other side of the stimulus stream. From the attended side they had to either read letters aloud or say the sum of the numbers within each segment.

There were a total of 16 trials, each lasting 8.1 seconds. The start of each trial was manually triggered by the experimenter. Every trial contained two side-cues, which appeared at fixed positions, and each signalled which side needed to be attended for the subsequent stimuli. The first side-cue appeared at the start and contained the verbal instruction "WATCH LEFT" or, "WATCH RIGHT" with a duration of 1s and was

followed by a 1s blank interval. The remaining stimulus sequence was made up of frames displayed for 200ms. These frames were organized in 6 segments: 5 letter/number segments and 1 second side-cue segment. Letter (or number) segments each contained 6 frames: the first contained a pair of asterisks followed by two frames each displaying a pair of letters (or numbers) and each of these was followed by a blank frame. The second side-cue segment contained 4 frames: a pair of asterisks, a blank frame, the second side-cue symbol positioned in the centre of the display and another blank frame.
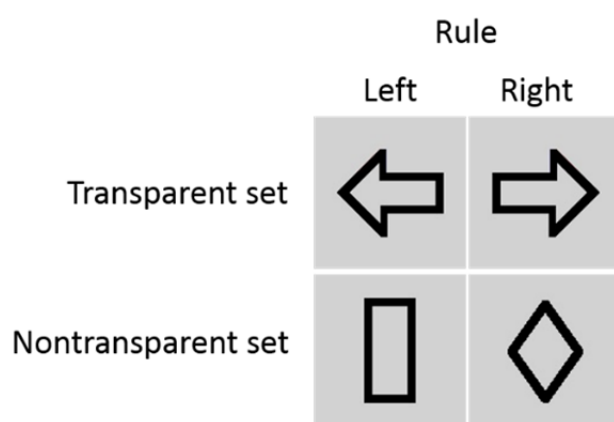
Figure 2.16 Critical cues in the LMT.

The second side-cue was the critical cue in the LMT and was in either the Transparent (top row) or the Nontransparent condition (bottom row).

The second side-cue was the critical cue in the LMT and was in either the Transparent or the Nontransparent condition (Figure 2.16). For the Transparent condition the second side-cue symbols were a left- and right-pointing arrow; for the Nontransparent condition, a diamond (indicating left) and rectangular shape (indicating right). All participants did the exact same task except for the condition of transparency of the second side-cue, with half the sample in the Transparent condition. The first side-cues (verbal instruction) and second side-cues (symbols) gave four possible combinations of instructions indicating which side to attend: non-switch trials, *left-left*, *right-right* and switch trials, *left-right, right-left*. The 16 trials were organized in blocks of 4, each block containing all side-cue combinations in random order.

The task had equal amounts of letter and number segments. For half of the trials the third, fifth and sixth segments followed a *number-letter-number* segment pattern, whilst the remaining half of trials used the reverse combination, *letter-number-letter*. This configuration was used to ensure task probabilities were balanced surrounding the second side-cue. For the remaining first, and second segments, allocation to either *number* or *letter* type was random.

2.2.2.3 Apparatus and stimuli

The computer setup was the same as Experiment 1 except that no button boxes were used. The stimuli in this task included uppercase letters, numbers, asterisks and symbols for the second side-cue. All stimuli were presented in black on a grey background; characters were rendered using Calibri Bold font with a character height of 0.5 degrees, and second side-cues were of comparable size. The distance between pairs of characters was 0.9 degrees measured edge-to-edge. Per trial, letters were randomly selected without replacement from the full alphabet excluding the letters D, I, O, V and W. Per number segment, numbers were randomly selected with replacement from the set 1-8 for the first number display, and the set 1-2 for the second display. The allocation of numbers in each segment was such as to ensure that the sum of the numbers on the left side was different from the right side.

2.2.2.4 Procedure

Participants were initially shown a printed example of a trial (Figure 2.15 left, without the segment labelling). Instructions were followed by a practice phase in which trials were run until the participant succeeded at reporting at least one character, irrespective of its correctness. 33 participants required just one practice trial, 22, two trials and 3, three trials. The practice was followed by cued recall of the instructions; if recall was not perfect, learning cycled through the cued-recall questions until it was perfect (see Appendix for verbatim instructions). Only 2 participants out of the whole sample needed to go through a second loop of the cued recall. The main experiment was then started. A card showing the relevant pair of second side-cue symbols (Figure 2.16, top or bottom row symbols only) was left on the table in full view of the participant throughout task execution, as a memory aid. The experimenter manually recorded responses on a score sheet. At the end, the participants' memory for the instructions was tested via another

cycle of cued-recall of instructions. This revealed intact memory of instructions for the whole sample.

The scoring protocol for the LMT was the same as in previous versions of the task (Duncan *et al.*, 1996 and Duncan *et al.,* 2008). Two scores were calculated per participant: i) performance on the first part of the trial, *before* the second side-cue and, ii) performance on the second part, *after* the second side-cue. Performance on the first part was calculated as the proportion of correct characters reported. This score reflected a basic ability to attend to one side, read letters, and make a simple addition. Performance on the second part was used to gauge GN and reflected the participant's ability to attend and follow the instruction represented by the second side-cue. GN was typically manifested as cases in which participants persevered with the side they had started with, ignoring the second side-cue. During non-switch trials, the behavioural outcome of either attending or neglecting the second side-cue was largely indistinguishable. GN was calculated using the *side-error score*, which is the probability that letters and numbers reported from the final part of the trial came from the incorrect side. The side-error per trial was calculated by giving a score of 1 if most responses were from the incorrect side, a score of 0.5 if the same number were reported from both sides and a score of 0 if more responses were reported from the correct side. Hence, a score of 1 reflects evidence for GN on a single trial. For each participant, the Mean Side-Error (MSE) was calculated by averaging the side-error across all 16 trials. The minimum MSE score was 0, which indicated no evidence of GN across the task. In theory, the maximum MSE score was 1, however, in practice, the maximum score was expected to be 0.5 because GN was unlikely to be gauged other than in the switch-trials which made up half the task. The advantage of using the MSE is that it allows equal weighting of all trials. There were no reaction time data for the LMT.

Finally, Cattell's Culture Fair Test Scale 2, Form A (Cattell, 1971; Cattell and Cattell, 1973) was administered at the end of the testing session, unless an existing score not older than 5 years was already available.

2.2.3 RESULTS AND DISCUSSION

The key finding of this experiment was that GN, measured via the MSE, was significantly larger in the Nontransparent condition (Median=0.16) than in the Transparent condition (Median=0.00), W=149.5, Z=-4.08, r=-0.54, p<0.0001 (

Figure 2.17 and Table 2.1). This indicates that the transparency of the cue modulates GN in the LMT. Participants in both conditions showed good performance in the first part of the LMT, with a median score above 96% correct. A Mann-Whitney-Wilcoxon test for independent samples revealed a significant difference between the two conditions, W=550.5, Z =2.60, r=0.34, p=0.039 (2-tailed). This suggested that condition of transparency also affected the basic performance on the first part of the task.
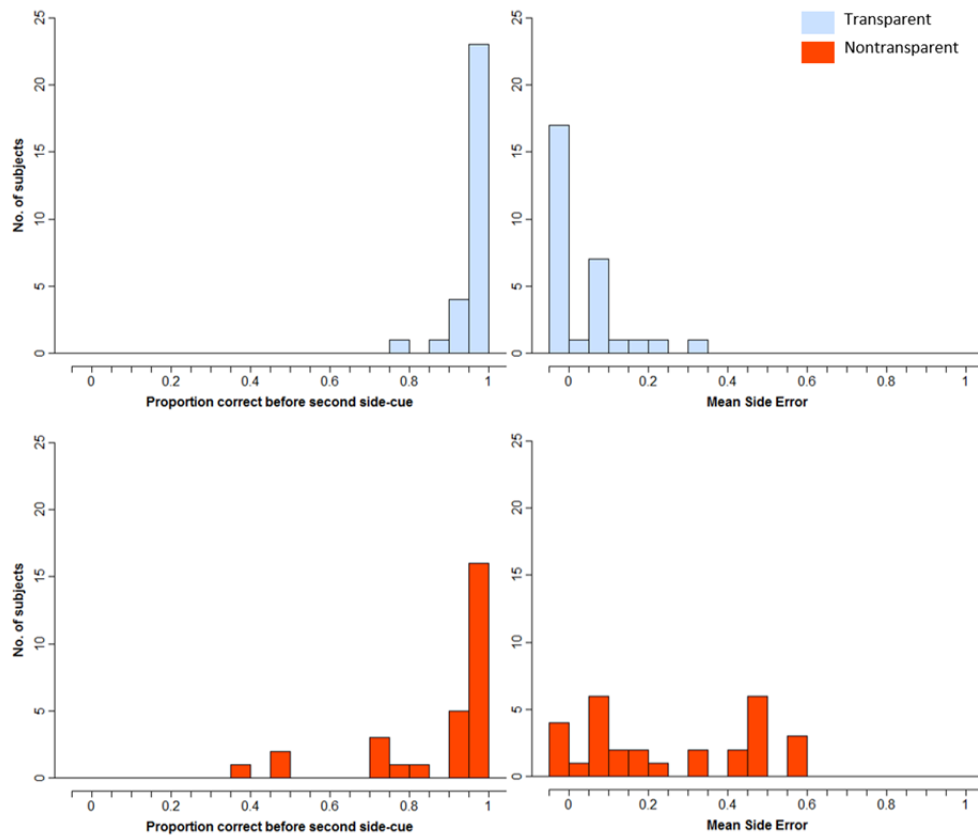
Figure 2.17 Performance on the LMT.

On the left column is the performance on the first part of the LMT as proportion correct. On the right-hand side is GN measured in terms of the MSE.

Table 2.1 Performance on the LMT.

Part	Condition	Mean(SD)	Median
First (correct)	Transparent	.97(.043)	.99
	Nontransparent	.88(.172)	.96
Second (MSE)	Transparent	.05(.080)	.00
	Nontransparent	.26(.217)	.16

Congruent with my hypothesis, low scores of fluid intelligence predicted high GN, in both conditions of transparency, ρ=-.34 (Transparent), ρ=-.42 (Nontransparent), both p<0.05 (1-tailed) (Figure 2.18). These associations remained significant even after partialling out age. Using Fisher's r-to-z transformations (Preacher, 2002), the strengths of these relationships (ρ=-.34 vs. ρ=-.42) were not found to be significantly different from each other (p=.74, 2-tailed) nor were they different (p=.14, 2-tailed) from previous published findings (comparison with Experiment 2 with n=65, r=-.66 as found in Duncan et al., 2008). This suggests that contrary to my prediction, the association between fluid intelligence and GN did not increase significantly with a decrease of transparency.
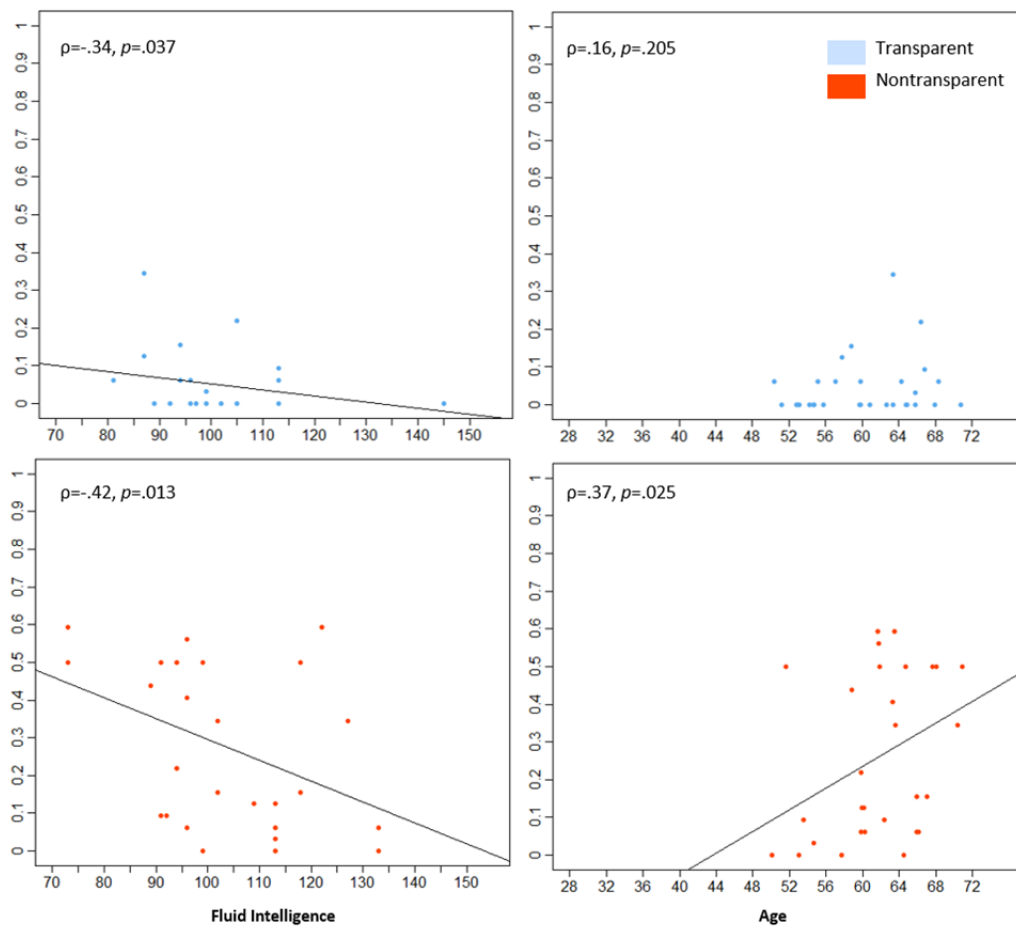
Figure 2.18 Correlations in the LMT

The left-hand column illustrates the associations between GN and fluid intelligence. The right-hand column shows the correlations between GN and age. In red, data for the Nontransparent cue condition (bottom row) and in blue the Transparent cue condition (top row). Plots with a line indicate a significant relationship and *p*-values are all set to 1-tailed because the hypotheses are directional.

Low fluid intelligence scores were associated to poor performance in the first part of the LMT only in the Nontransparent condition, ρ=0.33, but at borderline statistical significance *p*=0.076 (2-tailed).

Congruent with the results from the SGNT in Experiment 1, age was also found to be a significant predictor of GN in the LMT, with older participants performing worse than younger adults. This finding was restricted to the Nontransparent condition, ρ=.37, *p*=.025 (1-tailed) whilst the association in the Transparent condition, ρ=.16, was not significant (Figure 2.18). Partialling out fluid intelligence did not alter these findings. Based on the nominal findings, these associations suggest that like in Experiment 1, older

participants are performing significantly worse than younger participants when the instructional cue is Nontransparent. However, using r-to-z transformations, these coefficients (ρ=.37 vs. ρ=.16) were not found to be statistically significantly different from each other p=.41. This null finding needs to be interpreted with caution, given the small sample size compared to Experiment 1.

In addition, I wanted to test whether the association between individual differences and GN was different across the SGNT and the LMT. To do this I compared the coefficients between tasks within condition of transparency (refer to Figure 2.7 and Figure 2.18 for the correlation plots). Using Fisher's r-to-z transformations (Preacher, 2002), none of the strengths of these relationships were found to be significantly different from each other (Table 2.2). This suggested that individual differences in terms of age and fluid intelligence were similarly associated with GN across both tasks. However, it is worth noting that the fluid intelligence-GN association in the SGNT (Figure 2.7), unlike in the LMT (Figure 2.18), could be explained by age.

Table 2.2 Comparison of effect sizes of individual differences vs. GN, across tasks

Condition (IV)	Task	Coefficient for IV vs. GN	p
Transparent (fluid intelligence)	SGNT	-.21	.55 n.s.
	LMT	-.34	
Nontransparent (fluid intelligence)	SGNT	-.17	.23 n.s.
	LMT	-.42	
Transparent (age)	SGNT	.34	.41 n.s.
	LMT	.16	
Nontransparent (age)	SGNT	.48	.57 n.s.
	LMT	.37	

IV= Independent Variable

Comparisons of effect sizes as measured in Spearman's rho across tasks (SGNT and LMT) within condition of transparency and within IV (fluid intelligence or age). None of the contrasts were significant suggesting that the associations between GN and individual differences was similar across tasks.

2.2.4 CONCLUSION

Overall, the results from the LMT in this Experiment 2 converged with the results from the SGNT in Experiment 1. Experiment 2 presents further evidence to reject the null hypothesis, indicating that manipulating the degree of transparency of the instructional cue significantly affects the observed GN. Fluid intelligence predicted GN in the LMT, however this association was not modulated by the degree of transparency, which is consistent with Experiment 1. Older participants (in the Nontransparent condition only), showed significantly more GN. In addition, there was a trend for the effect of transparency to interact with age, suggesting that older people found the Nontransparent condition disproportionally more difficult; however, perhaps due to lack of power, this was not statistically significant. Both of these age results converge with the ones found in Experiment 1.

2.3 GENERAL DISCUSSION

The main question of this study was whether the transparency of the instructional cue affected failures in performance as measured by GN. Both experiments revealed that when the cue was strongly associated to the rule, neglect may have occurred; however if the association was novel, then GN was reliably more likely to occur. This is congruent with one GN study (Towse et al., 2007), findings in analogous PM studies (McDaniel and Einstein, 2000; Marsh et al., 2003; Pereira et al., 2012; Meier and Zimmermann, 2015) and the task switching literature (Mayr and Kliegl, 2000; Arbuthnott and Woodward, 2002; Logan and Bundesen, 2004; Miyake et al., 2004). An additional novel finding was that age significantly predicted GN in both tasks.

Is the transparency effect in GN the same as in the PM studies? Some PM studies suggest that transparency affects the *retrospective* component of the PM intention but leaves the *prospective* component unaffected (Marsh et al., 2003). The equivalent processes in the GN tasks could be thought of as, (i) recognizing the cue as significant/special (*prospective)* and (ii) accessing rule information (*retrospective*). The cue accessibility RT analysis in the SGNT suggested that the *retrospective* component is affected by transparency, thus converging with the PM findings. What about the *prospective* component? In the SGNT, GN could be due to either one of two cases: i) the participant recognized the significance of the cue (intact *prospective* component) but failed to inhibit the default Emotion sub-task or, ii) the participant failed to recognize the significance of the cue (impaired *prospective* component) and hence naturally proceeded to respond according to the default rule. The cue sensitivity analysis suggested that the first case could be excluded, although this rests on a null result. Therefore, by exclusion, this is indirect evidence that GN is likely to be driven by case (ii) rather than (i). Consequently, the larger incidence of GN within the Nontransparent condition would suggest that transparency is not only affecting the *retrospective* component as reported by PM studies, but also the *prospective* component. Nonetheless, it is worth noting that it is not entirely clear how these proposed components (prospective and retrospective) and their functional abstraction into sub-components (e.g. recognition, verification, retrieval, execution) are distinct from each other in terms of cognitive processes, as well as their temporal order and duration.

This study also investigated how individual differences were related to GN and whether they moderated the transparency effect. As expected, there was a negative relationship between fluid intelligence and GN in both tasks which replicated previous findings (Duncan *et al.,* 1996; 2008). However, in the SGNT this association could be explained by age, perhaps indicating that the cognitive mechanisms underlying GN are not identical across the two tasks. Fluid intelligence did not interact with the transparency effect, hence suggesting that fluid intelligence is not sensitive to this kind of manipulation of complexity unlike with some other kinds (Duncan et al., 2008; Roberts and Anderson, 2014).

For the first time, a relationship between age and GN is reported. GN published studies (Duncan et al., 1996; Duncan et al., 2012) simply report that when age was partialled out from the GN-fluid intelligence correlation the latter remained unchanged; however it is not clear what the GN-age relationship was to start with. In this present study, age was strongly and positively correlated to GN, at least in the Nontransparent condition, even after controlling for covariance with fluid intelligence. This age effect is congruent with PM findings, suggesting that older adults perform worse in prospective remembering when this is tested in the lab (Maylor, 1993, 1996; Kvavilashvili et al., 2009; Schnitzspahn et al., 2011). The interaction of transparency and age was mixed. In the SGNT older participants found the Nontransparent condition disproportionally harder than the Transparent condition when compared to younger participants. This could not be explained by reaction time differences on correct critical trials. In the LMT, the relationship between age and GN was not modulated by transparency. This is congruent with the PM study by Pereira, Ellis and Freeman (Pereira et al., 2012) in which they manipulated the transparency of the semantic association between a cue-action word pair. Overall, this mixed finding supports the hypothesis that the two GN tasks may be recruiting partially different cognitive mechanisms.

2.4   CONCLUSION

Previous GN experiments (Duncan et al., 2008) suggested that load *during* the task (ATC) does not affect GN, whilst load *before* the task (instructional) does. In contrast, this chapter presented two novel experiments which revealed that some forms of load *during* the task do affect GN, specifically, when the complexity is manipulated via the

transparency of the instructional cue. Is it possible, then, that the absence of ATC effects reported in previous published GN studies is simply due to low power or to the type of load that was used? I test this via two experiments in Chapter 4. Also, if participants neglect because they fail to recognize the significance of the cue, why is this occurring? In Chapter 3, I use an entropy-like measure to investigate the temporal pattern of GN in both the SGNT and LMT in an attempt to answer this question. Finally, some differences in the relationships between age/fluid intelligence and GN across the tasks may suggest different underlying cognitive mechanisms at play. I investigate this further in Chapter 5.

Chapter 3. MODELLING THE TEMPORAL DYNAMICS OF GN USING AN ENTROPY-LIKE MEASURE

A functional account for GN was advanced in Chapter 2. This suggested that a likely explanation for GN is that it is driven, at least in part, by a failure to recognize the significance of the critical cue during task performance. Yet, the comprehension and memory for the task rules was intact when tested via cued recall, both at the start and end of the task. How is it possible that neglecting participants knew the meaning of the critical cues, and yet behaved as if they did not know? This chapter has two main aims. First, it reviews two different theories that may explain this phenomenon. Second, it investigates support for these models by analysing the temporal dynamics of GN as observed in the SGNT and LMT.

3.1 COGNITIVE MODELS FOR GN

3.1.1 TASK MODEL

One proposed model for GN is a failure of a control mechanism referred to as the *task model* (Duncan et al., 2008; Duncan, 2013; Bhandari and Duncan, 2014). The first account of the *task model* defined it as a capacity-limited set of representations of the different task components which guides task performance. Given a limited capacity, task components must compete for representation and hence, the more weakly a component is represented the more likely it is for it to be lost from the task model and hence, neglected (Duncan et al., 2008; Duncan, 2013). Later findings (Bhandari and Duncan, 2014)

revealed that GN[1] was not driven by the total complexity of the task components, instead what appeared critical was the way in which this task information was organized - GN in a sub-task was sensitive to the complexity within that chunk, but not much affected by the complexity of the accompanying sub-task (see p.37 for more details on the panel task studies). In addition, these experiments also revealed that for a subset of participants who showed major poor execution, their performance was unstable over the first few trials, suggesting that the *task model* was also affected by the participant's own performance early on in the task. Together these recent findings prompted a revision of the *task model* account, where: (1) capacity is not limited, (2) GN is likely driven by a weakening of the *task model* and/or a failure to activate the right component because of increased competition from multiple task components, and (3) this competition for representation which shapes the *task model* occurs early on in the task, at instructions and early performance, and then stabilizes, setting the behaviour for the remaining part of the task.

Chapter 2 revealed that the transparency of the critical cue affects GN. The *task model* framework can explain this finding by suggesting that a Nontransparent-cued rule, being more complex, is more weakly represented than a Transparent-cued rule, and hence is more likely to be lost from this control structure. Interestingly, in the Nontransparent condition of the LMT, worse performance was also observed in the regular part of the trials where the critical event was absent (see Chapter 2, Experiment 2). This is congruent with the idea that competition from a complex task component may not only affect performance locally - when that task component needs to be applied - but it may be more diffuse in its impact and affect other parts within the same task set (for example, manipulations of complexity led to all sorts of errors aside from the ones to critical responses, in the experiments reported by Bhandari and Duncan, 2014). At first this may sound contradictory to the interpretation of the main finding in the work by Bhandari and Duncan (2014) which suggests that complexity is bound by the sub-components of task, i.e. complexity is bound by the sub-task. However, it is possible that limits of the diffused effects of complexity within a chunk/sub-task are somewhat dependent on the degree of salience of the boundary that creates two sub-tasks in a task. For example, in their panel

---

[1] GN in this study was loosely referred to as "major performance failures" and consisted of all kinds of errors, not just failures on critical events.

task, the sub-tasks (e.g. vehicles vs. books) are relatively easily distinguishable since they involve different task sets altogether (Figure 1.4).

As the *task model* theory developed, so did its predictions of the type of temporal distributions of GN. In earlier GN studies, it was observed that once the participant recovered from neglect this recovery was substantial and often complete (Duncan et al., 1996). The temporal function of GN was sigmoidal, almost step-shaped - a run of poor performance followed by substantial recovery once the *task model* was correctly established. Instead, in later studies (Bhandari and Duncan, 2014), an analysis on the task dynamics within a subset of participants who were showing major performance failures demonstrated unpredictable performance very early on in the task which then stabilized into a continuous display of errors. However, it is important to note that Bhandari and Duncan's analysis is not directly comparable to both the GN results reported in earlier GN studies and, the GN results I present for the following reasons: (1) the analysis focused on a sub-set of the data and collapsed across participants in this sample (2) it included all sorts of errors, not just errors on critical events of the task, which is what GN is otherwise typically defined as, and (3) it included practice trial data. In summary, these studies together suggest that the *task model* is established early on in the task and that once it stabilizes it tends to lead to either a pattern of full recovery or a pattern of sustained performance failures (Duncan et al., 1996; Bhandari and Duncan, 2014). In addition, even the type of error produced tends to stabilize, with one main type of error dominating (Bhandari and Duncan, 2014).

3.1.2 MONITORING AND OTHER ACCOUNTS

Arguably, GN may be driven, at least in part, by failures of cognitive mechanisms which are different, although not necessarily incompatible, from a *task model* account. Fatigue and motivational factors may be obvious examples (Pessoa, 2009). As mentioned in previous chapters, GN has a similar definition to PM failures (PMf). Therefore it seemed valuable to explore the theoretical frameworks offered within the PM literature as possible explanations for GN.

One of the dominant theories of PM is the Multiprocess Theory proposed by McDaniel and Einstein (2000). The Multiprocess theory does not consider anything like a *task model* in which task information may be fragile at the outset or early on in the task, despite intact memory of instructions. In the absence of any such predictions, it therefore

appears to assume that all components of task knowledge is equally intact. Instead, it focuses on investigating the mechanisms via which task knowledge interacts with varying task demands. Precisely, the concern is to understand how much of one's limited attentional resources need to be allocated to detect the PM target (analogous to the critical event, in GN terms), at the expense of the rest of the ongoing task (analogous to the regular sub-task, in GN terms, see Figure 3.1).
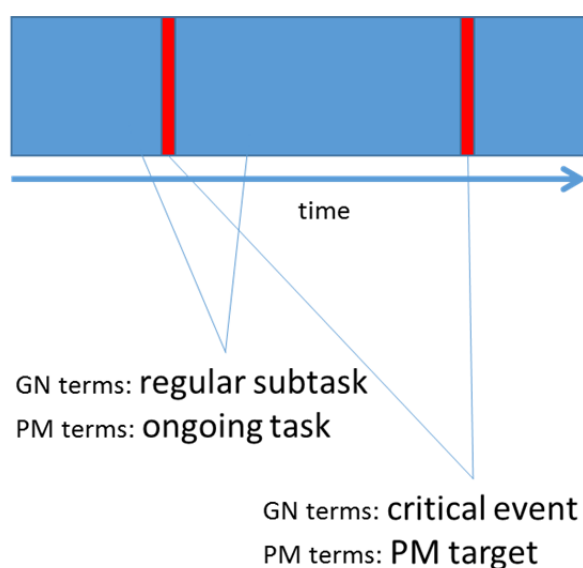
Figure 3.1 Similarity between GN and PM task structure

The task represented in the horizontal bar, illustrates the frequent regular sub-task (blue) which in PM terms is referred to as the ongoing task, and the less frequent critical events (red) which is equivalent to the PM targets in PM terms.

The Multiprocess Theory proposes that there may be two such mechanisms, *monitoring* and *spontaneous retrieval*, and their relative importance depends on the characteristics of the task (McDaniel and Einstein, 2000; Einstein et al., 2005). The *monitoring* account suggests that a PMf (Prospective Memory failure, which is analogous to GN, in GN terms) is driven by a failure of a mechanism which is actively and constantly monitoring for the PM cues (critical cues, in GN terms) and it does so by consuming from the pool of limited attentional resources. Hence, insufficient resources allocated to *monitoring* would result in some reduced readiness to detect these signals in the environment leading to a PMf. The presence of *monitoring* is established by the detection of costs (decreased accuracy and/or increased latencies) to the primary task (regular sub-task, in GN terms)

due to the presence of the secondary task, the PM instruction. This is typically calculated by subtracting the performance on a version of the ongoing task which does not involve the PM instruction, from performance on the ongoing task with the PM instruction (Smith, 2003; Einstein et al., 2005). The presence/absence of such costs has been interpreted as evidence of monitoring/no monitoring, respectively (Einstein et al., 2005; Harrison et al., 2014). Under conditions of no cost, PM retrieval has been hypothesized to occur via "spontaneous" processes (Einstein et al., 2005; Scullin et al., 2010; Harrison et al., 2014). While *spontaneous* processes are thought to show no significant costs when compared to *monitoring* mechanisms, they are neither thought to be fully automatic (Einstein and McDaniel, 2010; Harrison et al., 2014). Specific task characteristics have been identified to predict which type of mechanism is operative, *monitoring* or *spontaneous* (Einstein et al., 2005; Gonen-Yaacovi and Burgess, 2012). An example is the "focality" of the PM cue. A "focal" cue arises when the dimension of the features being processed in the ongoing task (e.g. lexical decision task) highly overlaps with the dimension of the PM cue (e.g. detect the word "actor"). In contrast, in a "non-focal" cue the overlap of dimensions is minimal (to continue with the previous example, the syllable "tor" as the PM cue). The data suggest that "focal" and "non-focal" cues are more likely to elicit *spontaneous* mechanisms and *monitoring* mechanisms, respectively (Einstein et al., 2005).

Instead, other PM theories indicate that *monitoring* for the PM target is always necessary and cannot be cost-free nor "spontaneous" (see PAM, Preparatory Attentional and Memory processes theory, Smith, 2003; Smith et al., 2007). If *monitoring* accounts for the attention allocated to the PM task plus the ongoing task, then *monitoring* costs (e.g. latency increase of 10 s) should predict either performance on the PM component or performance on the ongoing task. However, this has not been consistently the case, indicating that *monitoring* resources may vary over the course of the task (Marsh et al., 2003).

In my GN experiments, I lack task conditions in which the regular sub-task is free from the GN instruction hence, I am unable to directly establish the presence or absence of *monitoring* mechanisms. However, given that (1) the critical cues in both the LMT and SGNT are "non-focal", (2) the instructions did not minimize the importance of the GN rule, and (3) a considerable amount of GN was detected, the *monitoring,* rather than

spontaneous mechanism is highly more likely to be operative according to various PM studies (Einstein et al., 2005; Gonen-Yaacovi and Burgess, 2012; Harrison et al., 2014). Hence, from here onwards I will mostly ignore the *spontaneous* processing account of PM according to the Multiprocess Theory and instead focus on the *monitoring* mechanism.

3.1.3 SUB-TYPES OF GN BASED ON TEMPORAL DYNAMICS

The *task model* and *monitoring* frameworks are not mutually exclusive and hence it is conceivable that both processes, along with other mechanisms, are relevant at explaining GN. My primary interest is to understand whether it is possible to distinguish GN which is dominantly driven by one versus another mechanism. The relative dominance of different mechanisms is likely to depend on all sorts of factors (e.g. task type, task condition, task phase, context, individual differences etc…) and may therefore vary on a case-by-case basis. Considering that the *task model* and *monitoring* theories make different predictions on the temporal dynamics of GN, this offers a potential test bed to distinguish between these mechanisms.

What predictions does the *task model* account make in terms of temporal dynamics? As described earlier, the *task model* refers to some form of task performance blueprint, which is set at the instructions, is further shaped by early performance and then stabilizes. Hence, a *task model* account would predict that, (1) if there's any GN at all and recovery in performance, then GN would tend to occur early on in the task, (2) GN would tend to occur in clumps given that it's driven by an incorrect blueprint. What about a monitoring account? The *monitoring* theory contrasts with the *task model* account mostly in that it requires a constant vigilance of the environment for relevant events, hence it is dependent on the available attentional resources which may wax and wane as the ongoing task demands change over time. This *monitoring* mechanism would predict that (1) GN could occur at any time – whenever there is a lapse in this checking loop, (2) the errors are more likely to be spread out, in a zig-zag like fashion which would correspond to cycles of lapses and recovery of attention.

Importantly, these predictions are unlikely to be exclusive to these 2 specific mechanisms. For example, a sporadic pattern of errors may be a result of boredom. Hence, it is probably unwise to hypothesize that either one or the other process must account for all GN errors. Instead, the separation could be conceptualized as (1) more continuous, *task-*

model-like vs. *monitoring-like,* and (2) open to be explained by other mechanisms. Essentially, my goal is to explore the potential differentiation of GN into sub-types, as indexed by different temporal signatures. The latter are assumed to reflect different underlying cognitive processes, with a current focus on 2 possible variants.

Eyeballing temporal distributions of GN for data from the SGNT and LMT on a participant-by-participant basis did not appear to show an obvious dichotomy of sigmoidal (*task-model-like)* or zig-zag-like (*monitoring-like)* patterns. In fact, if both *task-model-like* and *monitoring-like* mechanisms are operant, then it would be harder to distinguish between the two, given that it is not unlikely for parallel cognitive processes to yield nonlinear (sub-additive) effects on behaviour (Chatham and Badre, 2015). However, if it were possible to use a quantitative measure of the temporal spread of errors, then one could empirically test for the prevalence of one pattern of errors over another. In the next section, I propose the use of a recent advancement in the field of statistical research to quantify such spread, or "clumpiness" of errors in cognitive science.

3.2 CALCULATING THE ENTROPY-LIKE MEASURE

Data clumpiness can be defined as, possibly irregular, cluster(s) of events close together, and has been of interest across various areas of studies including financial markets, seismology and digital consumption (as cited in Zhang et al., 2013). For example, in sports it can be used to assess the so called "hot hand" effect which describes when a sportsman goes through a period of successful games which are significantly better compared to the sportsman's average play. However, most standard measures of clumpiness have been criticized for lacking power and/or for suffering from other limitations (see Zhang et al., 2013). Hence, this motivated the development of a new class of improved clumpiness measures by Zhang and colleagues (2013). One such measure is based on entropy. Entropy, as found in information theory, is defined as follows:

> *Entropy is a measure of the uncertainty associated with a random variable. It quantifies the expected value of the information contained in a message.*

> *(p.10, Zhang et al., 2013)*

When entropy is high, there is: more uncertainty, more information, more surprise and more questions need to be asked to predict the content of the message. Entropy is maximum when all outcomes are equally likely (Khanacadamy, 2016). Shannon's entropy formula is as follows:

$$H_s = -\sum_{i=1}^{n} p_i \, log_2(p_i)$$

or

$$H_s = \sum_{i=1}^{n} p_i \, log_2(1/p_i)$$

Where the entropy *H* for message *s,* made up of a sequence with a total of *n* symbols, is the sum of the probability *p* of each symbol occurring in the sequence, multiplied by the log of base 2 of the inverse of this probability. This is calculated starting from the first symbol (*i*=1). The unit of Shannon entropy is "bits". For example, a fair coin has 0.5 and 0.5 probabilities that it will show heads or tails, respectively, which is equal to [0.5 x $log_2(0.5)$] + [0.5 x $log_2(0.5)$]=1 bit. Whilst a crooked coin with 0.75 and 0.25 probabilities of showing heads or tails, is equal to [0.75 x $log_2(1/0.75)$] + [0.25 x $log_2(1/0.25)$] is approximately equal to 0.81 bits, which is less entropy than the fair coin. Zhang et al., (2013) propose an entropy-like formula in order to measure clumpiness and they primarily do so by replacing the event probability (*p*) with a distance measure. That is, rather than being based on the frequency (probability) of the events, the measure is now based on the distance between the events.

The formula is as follows:

$$H_c = \sum_{i=1}^{n+1} x_i \log(x_i)$$

Where H_c is the entropy-like measure for a set of events over a given period of time. The formula changed in 3 ways:

1. The probability value p is replaced with a new measure, x, which they define as the inter-event time (IET). Further details on the IET are presented below.
2. The minus sign is removed so that when they eventually convert the entropy-like measure (H_c) to their clumpiness index, this value is intuitive such that a high value is equal to high clumpiness.
3. They add 1 to the number of events n. This is necessary because of the nature of their event measure x. Further details are described below.

The inter-event time (IET) is calculated below following a description from (p.10, Zhang et al., 2013). Where i is the th occurrence of the event, t is the position value of an event and N is the number of trials/positions in the sequence. Let's imagine sequence A with 16 trials ($N=16$) and 4 events ($n=4$):

 Sequence A. 0010010010001000

There are 4 events, hence, i loops from the 1st to the 4th final event. When $i=1$, $t=3$; when $i=2$, $t=6$; when $i=3$, $t=9$; when $i=4$, $t=13$.

| Step 1a | $x_1 = t_1$ | For the first event, the IET is the value of the position of the event. For example, for sequence A, $x_1 = 3$ |

| Step 2a | $x_i = t_i - t_{i-1}$ | For all the other events, the IET is the position in the sequence of the event (t_i) minus the position of the previous event. For example, for sequence A, $x_2 = 3$, $x_3 = 3$, $x_4 = 4$. |

Step 3a	$x_{n+1} = N + 1 - t_n$	Step 2a makes it clear that each IET is equivalent to each gap surrounding events in the sequence. Given that the first interval is measured from position zero to the first event (Step 1a), then, an additional final interval is measured from the final event to N+1 (Step 3a[2]). Hence, for sequence A, the last IET is $x_5 = 4$.
Step 4	$\sum_{i=1}^{n+1} x_i = N + 1$	Given the equations above, it follows that the sum of all the IETs will be equal to N+1. For example, for sequence A, $x_1 + x_2 + x_3 + x_4 + x_5 = N + 1 = 17$.

The authors wanted an entropy-like measure which is independent of the length of trials, hence the entropy-like measure is calculated with each of the IETs scaled by N+1 (see Steps1b-3b[3]).

Step 1b	$x_1 = t_1 \div (N + 1)$
Step 2b	$x_i = (t_i - t_{i-1}) \div (N + 1)$
Step 3b	$x_{n+1} = (N + 1 - t_n) \div (N + 1)$

---

[2] Please note that in the original paper there is no +1 added to N, however this definition has been updated and was added post publication (Zhang et al., 2013) via personal communication (email) with the authors Zhang Y, Bradlow ET (2016) Entropy-like measure of clumpiness In: (Biondo F, ed). -personal communication-.
[3] *Ibid.*

Therefore, the IETs (x) are as follows:

$x_1 = 3 \div 17 = .176$

$x_2 = 3 \div 17 = .176$

$x_3 = 3 \div 17 = .176$

$x_4 = 4 \div 17 = .235$

$x_5 = 4 \div 17 = .235$

Now that we have IET values (x) for the whole sequence, we can proceed to calculate the entropy-like measure of clumpiness using the equation $H_c = x*\log(x)$. Please note that instead of the log to the base of 2, the formula uses the natural log. The intermediate products are as follows: .176*log(.176)=-.306; .176*log(.176)=-.306; .176*log(.176)=-.306; .235*log(.235)=-.340; .235*log(.235)= .340. The sum of these intermediate products is approximately equal to: -.306 + -.306 + -.306 + .340 +.340=-1.60. Hence, the entropy-like value for clumpiness, H_c=-1.60. Please note that the authors then go on to further standardize and scale this entropy-like value to a clumpiness index (Zhang and Bradlow, 2016) but this is beyond the scope of the current analysis.

A problem with applying this entropy-like value (H_c) to my data is that it only cares about the gaps between errors (events) and does not care about the gaps between corrects (i.e. the run length of errors). Such that, if events in sequence A were now represented by "0"s instead of "1"s, this formula would give a different measure of H_c=-2.51 instead of H_c=-1.60. This mismatch is not ideal.

Hence, I propose modifying the equation for an entropy-like value of clumpiness as follows:

$$H_r = -\sum_{i=1}^{n} y_i \log(y_i)$$

Chapter 3 | Modelling the temporal dynamics of GN using an entropy-like measure

The changes from the Zhang et al., (2013) formula (H_c) are as follows:

1. The event measure (x) has changed to (y) which is no longer the intervals between events (IET). Instead, the new distance measure is the distance between transitions of events. In other words, it is the run lengths of each event type. (An example is provided in Table 3.1).

2. Given that I am not interested in converting to a clumpiness index that the authors go on to describe elsewhere (Zhang and Bradlow, 2016), there is no longer the need to remove the minus sign from the original Shannon's entropy formula (H_s). Hence, I reintroduce this minus sign.

3. Given that the distance measure is based on run lengths, there is no longer the need to introduce an event (+1) to n at the end of the sequence. Hence, I remove the +1.

Both entropy-like measures of clumpiness (H_c as proposed by Zhang et al., 2013 and H_r as the modified entropy measure that I propose) which use a distance measure as the event in Shannon's entropy (H_s), are sensitive to the edges of the sequence. This means that as an event approaches the centre of the sequence, the entropy-like measure increases, that is, it becomes less clumpy. This is reasonable given that at the mid-point of the sequence we can be most confident that the event (e.g. error) is surrounded by a non-event (e.g. correct) and hence that the gap is maximal. In other words, as an event approaches the edge of the sequence there is an increasing probability that the event might belong to a clump whose other members have not been observed. Other details on this issue can be found in Appendix C of Zhang et al.'s paper (2013).

What follows is an example to illustrate the application of the modified entropy-like measure to my GN data. In my calculations, transition of events are defined as a change from correct critical trials (coded as "0") to GN (coded as "1"), or vice versa. Let's imagine two sequences of critical trial data (sequences B and C, below), with 16 trials each, each displaying 5 cases of GN. The entropy-like measure is calculated as in Table 3.1.

Sequence B.   1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0
Sequence C.   1 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0

Table 3.1 Calculating the entropy-like measure

For both sequences B and C, N=16 (number of trials), n=5 (number of errors). Total entropy (H_t) is the negative of the sum of $y_i \log(y_i)$ where y_i is the run length scaled by the total number of trials.

Sequence B				Sequence C			
Run length	y_i	$log(y_i)$	$-y_i *log(y_i)$	*Run length*	y_i	$log(y_i)$	$-y_i *log(y_i)$
4	4/16	-1.386	.347	1	1/16	-2.773	.173
1	1/16	-2.773	.173	3	3/16	-1.674	.314
1	1/16	-2.773	.173	1	1/16	-2.773	.173
10	10/16	-0.470	.294	1	1/16	-2.773	.173
				1	1/16	-2.773	.173
				4	4/16	-1.386	.347
				1	1/16	-2.773	.173
				1	1/16	-2.773	.173
				1	1/16	-2.773	.173
				2	2/16	-2.079	.260
H sequence B =		0.987		H sequence C =			2.133
More clumpy				*Less clumpy*			

In sequence B, the entropy-like measure is lower than in sequence C. In sequence B, GN is more clumpy than in sequence C. The lower the entropy-like measure, the higher the clumpiness.

A limitation to these entropy-like measures is that they are sensitive to the number of events - the larger the number of events, the higher the entropy-like measure tends to be (see Figure 3.2). To some extent, this is intuitive: minimal clumpiness is achieved when there are equal numbers of each response type, in alternation; if there is only a single instance of one response type then, trials of the other response types must form one or two large clumps. Considering that my intention is to compare the clumpiness of GN independently of the number of GN cases, then this limitation needs to be addressed. To

do so, in my analyses that use the entropy-like measure H_r, I control for variables which model the relationship between errors and the entropy-like measure.
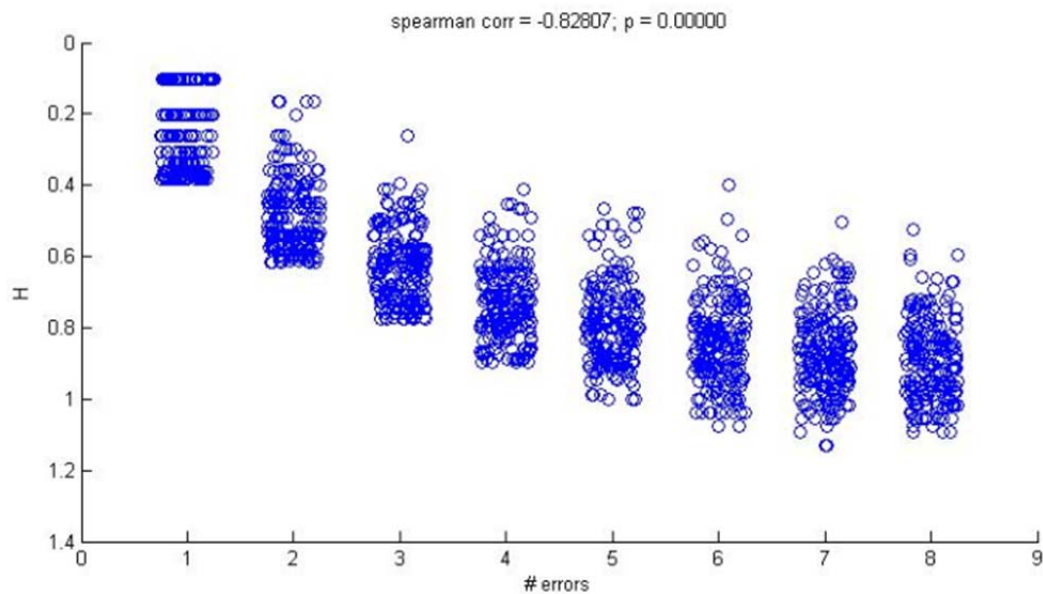
Figure 3.2 Entropy-like measure H_r and number of errors

Using simulated data (errors positioned randomly within trial sequence, for 1-8 errors) to model the relationship between number of errors/events and the entropy-like measure reveals a bias. This bias suggests that the larger the number of events, the larger the entropy (H_l). The function appears to be quadratic.

3.3 USING THE ENTROPY-LIKE MEASURE TO TEST FOR GN SUBTYPES

Now that I have a measure (H_r) to quantify the clumpiness of GN I can proceed with hypothesis testing.

Following the temporal predictions discussed earlier (p.82), the hypothesis is that participants who are neglecting primarily due to a faulty *task model* and show recovery in their performance will (1) display the majority of their GN early on in the task and, (2) this GN will tend to be clumpy. Instead participants who are neglecting primarily because of *monitoring*-like failures will (1) not have any specific bias over which stage in the task GN occurs and, (2) GN would tend to be less clumpy and more zig-zag like in nature.

Considering that the hypotheses suggest that both mechanisms (*task model* and *monitoring*) are present at the start of the task, I can constrain the hypotheses by adding

another premise based on *when* the majority of errors occur. If most of the observed GN happens towards the end of the task, then this is unlikely to be *task model* driven. To measure when the majority of GN occurs I can use the median position of GN. If the median approaches 1, this would indicate that GN tends to happen early on in the sequence of trials, whilst if it approaches 16 then it suggests that GN is predominantly occurring towards the end. Therefore, the revised hypotheses are as follows:

Hypothesis 1.   If most of GN is occurring towards the end of the task (median position of GN is high) then this is less likely to be *task model* driven, and by exclusion, more likely to be a result of a failure of *monitoring-like* mechanisms which should show less clumpy GN (higher H_r) .

Hypothesis 2.   Instead, if most errors are occurring at the start of the task, then it is more likely to be *task model* driven (on top of any additional *monitoring* driven errors) and hence, more likely to be clumpy (lower H_r).

3.3.1   METHOD

Critical trial data from the SGNT and LMT (Chapter 2) were split into four groups according to the condition of transparency: SGNT Transparent, SGNT Nontransparent, LMT Transparent and LMT Nontransparent. The condition of transparency reflected a different critical rule, hence a different sub-component of the hypothetical *task model*. Each group consisted of sequences of 16 trials of critical sub-task accuracy data (regular trials are ignored). There were as many sequences as number of cases (participants). Each trial in each sequence was coded as "correct", "GN", "omission" or "unknown". In the LMT, non-switch trials were (by design) unlikely to show GN, hence instead of sequences with 16 trials, only the 8 switch trials were used. For each case, the median value of all of the positions of GN was calculated. Cases with either no GN-trials or all of the trials showing GN, do not allow a measure of spread of errors and therefore the entropy-like measure cannot be calculated. Hence, such cases were excluded from the analysis. To accommodate the critical trial data to the entropy-like measure calculation, the data were recoded to binary such that, correct=0 and GN=1. In this recoding, "omissions" and "unknown" trials were collapsed to the "correct" category to favour a conservative measure of GN. Finally, the entropy-like measure was calculated for each

case and summary values of the entropy-like measure and position of GN are shown in Table 3.2.

Table 3.2 Summary values for the *Early vs. Late GN* analysis in the SGNT and LMT

	SGNT	LMT
N	50	22
Number of neglectable trials	16	8
Position of GN		
Median	5	3.5
Mean(SD)	5.82(3.5)	3.61(2.42)
Entropy-like measure		
Median	0.97	0.74
Mean(SD)	1.10(0.54)	0.86(0.51)
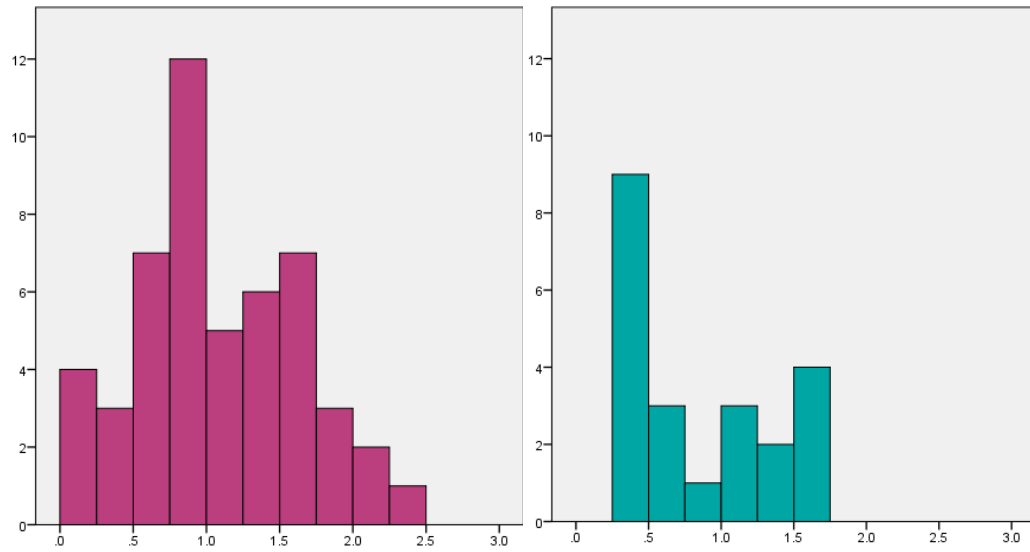
Figure 3.3 Histograms of the *Entropy-like measure* for the SGNT and LMT

On the *x*-axis is the *Entropy-like measure,* for the SGNT (pink, left) and the LMT (green, right). On the y-axis is frequency.
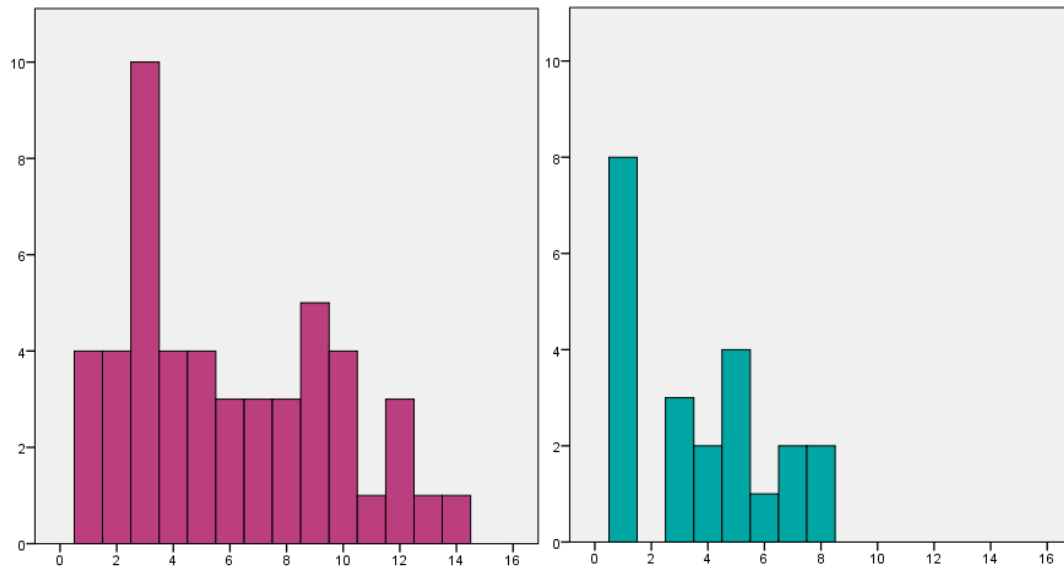
Figure 3.4 Histograms of the *Median position of GN* for the SGNT and LMT

On the *x*-axis is the *Median position of GN,* for the SGNT (pink, left) and the LMT (green, right). On the y-axis is frequency.

To test if lower values of the entropy-like measure (higher clumpiness) are predicted by how early or late GN tended to occur, I ran a linear regression for each task (SGNT and LMT), with the entropy-like measure as the dependent variable and the following independent variables:

- *"Median position of GN"* variable.
- *"Transparency"* factor, which codes for whether the data were in the Transparent or Nontransparent condition.
- *"Number of GN"* variable(s). The aim of these variables was to model the dependency of the entropy-like measure (H_r) on the number of events. Given that a non-linear relationship was observed between H_r and number of errors, I entered first, second, third, fourth and fifth degree expansions of this GN term. To avoid a problem of multicollinearity in the regression model, I used orthogonal polynomials.
- *Participant* variables. The SGNT had a within-subject design but not all participants displayed GN across both conditions of transparency. To avoid a mixture of within- and between-subjects data in the model, only participants who displayed GN in both conditions of transparency were used. Hence, the use of *participant variables* which coded for cases by the same individual in the two conditions. These were not necessary for the LMT because the design was between-subjects.

The independent variables were entered into the model simultaneously ("enter" method) which according to some researchers is the only appropriate method for theory testing (Studenmund & Cassidy, 1987 as cited in Field et al., 2012 p.298).

Table 3.3 Regression analysis output of the entropy-like measure in the SGNT

	B	SE B	β	p
Constant	.70	.25		.010**
Median position of GN	.08	.03	.53	.004**
Transparency factor	.03	.11	.02	.819
Number of GN polynomial 1st degree	1.16	.60	.30	.066
Number of GN polynomial 2nd degree	-1.18	.48	-.31	.023*
Number of GN polynomial 3rd degree	.70	.49	.18	.168

R^2 =.88; *p<.05, **p<.01.
B=unstandardised beta value, SE B=standard error of B, β=standardized beta value, which is the amount of standard deviations by which the dependent variable will change given the standard deviation change in the independent variable, and hence is independent of the units of measurement of the variables (Field et al., 2012).

The results for the two linear regressions can be seen in Table 3.3 and Table 3.4. In the SGNT, the median position of GN was significant at predicting clumpiness of GN which suggests that even when the number of errors was accounted for, GN was more clumpy when it occurred early compared to later on in the task. However, this finding was not replicated in the LMT, where only the number of errors was significant at predicting clumpiness (Table 3.4).

Table 3.4 Regression analysis output of the entropy-like measure in the LMT

	B	SE B	β	p
Constant	.73	.12		.000**
Median position of GN	.02	.02	.08	.494
Transparency factor	.14	.10	.14	.168
Number of GN polynomial 1st degree	1.29	.22	.56	.000**
Number of GN polynomial 2nd degree	-1.58	.21	-.68	.000**
Number of GN polynomial 3rd degree	-.26	.22	-.11	.255

R^2 =.88; *p<.05, **p<.01.
B=unstandardised beta value, SE B= standard error of B, β=standardized beta value, which is the amount of standard deviations by which the dependent variable will change given the standard deviation change in the independent variable, and hence is independent of the units of measurement of the variables (Field et al., 2012).

These results prompt another question – is it possible that individual differences such as age and fluid intelligence predict whether participants will show GN earlier or later in the

task and/or whether they predict overall clumpiness of GN? To test this I ran four separate linear regressions, two for each of the tasks SGNT and LMT. Within each task, the first regression model had *Median position of GN* as the dependent variable (DV) while the second regression model had the *Entropy-like measure* as the DV. The independent variables (IVs) were fluid intelligence, age, the *number of GN*, the *Entropy-like measure* and the *Median position of GN* (unless any of these were the DV) and the within-subjects condition of transparency. In the model for the *Entropy-like measure*, the IV of *number of GN* was translated to orthogonal polynomials to the third degree of expansion, in order to model the non-linear association between the *Entropy-like measure* and frequency of GN, in line with the method of the previous regression analysis.

See scatterplots in Figure 3.5 to Figure 3.12 for raw correlations between Fluid intelligence or age and *Median position of GN* or *Entropy-like measure*. High IQ scores were associated to more errors earlier on in the SGNT (Figure 3.5) and although there was a similar trend in the LMT (Figure 3.6), this was not statistically significant. Age did not appear to be related to the position of errors (Figure 3.7 and Figure 3.8). Instead, the *Entropy-like measure* was related to age in the SGNT, such that older subjects showed less entropy, i.e. older subjects showed clumpier errors (Figure 3.11). A similar pattern was found in the LMT, however this was not significant (Figure 3.12). There were no significant associations between fluid intelligence and the *Entropy-like measure* (Figure 3.9 and Figure 3.10). However, I was interested to test these associations whilst taking into account other variables. The results of the linear regression are presented in Table 3.5 and Table 3.6, and indicated that in the SGNT only, it was possible to predict the *Median position of GN* based on fluid intelligence, even when accounting for age, the number of GN and the entropy-like measure. This suggests that for participants with high scores of fluid intelligence, instances of GN tended to occur earlier-on in the task. In the LMT, neither age nor fluid intelligence were significant predictors. The *Entropy-like measure* regression model did not reveal any significant predictors, except for the known relationship with the number of errors (see Figure 3.2). This suggested that overall clumpiness of GN could not be predicted by any of the individual differences measures, at least when taking into account other variables such as the number of errors.
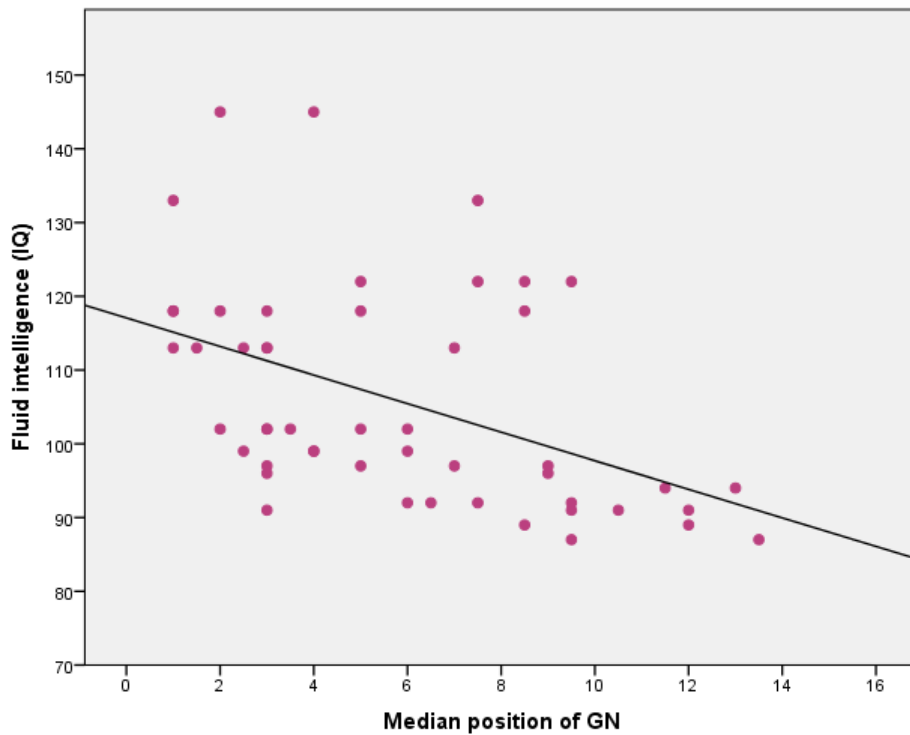
Figure 3.5 Scatterplot of Fluid intelligence and Median position of GN in the SGNT

The correlation was r=-.46, p<0.001***.
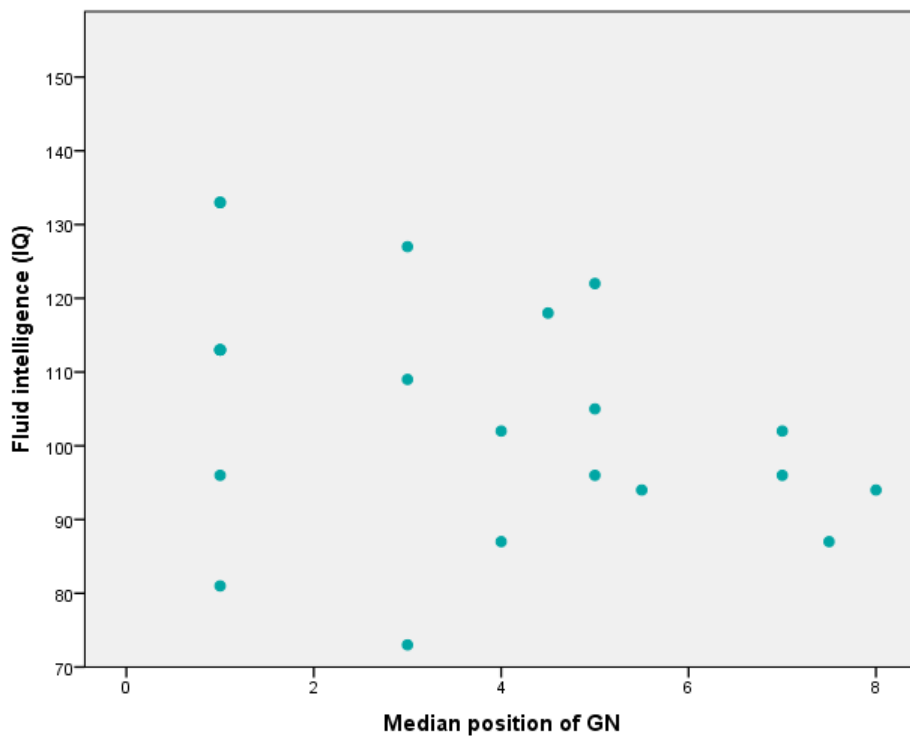
Figure 3.6 Scatterplot of Fluid intelligence and Median position of GN in the LMT

The correlation was r=-.37, p=0.09 n.s.
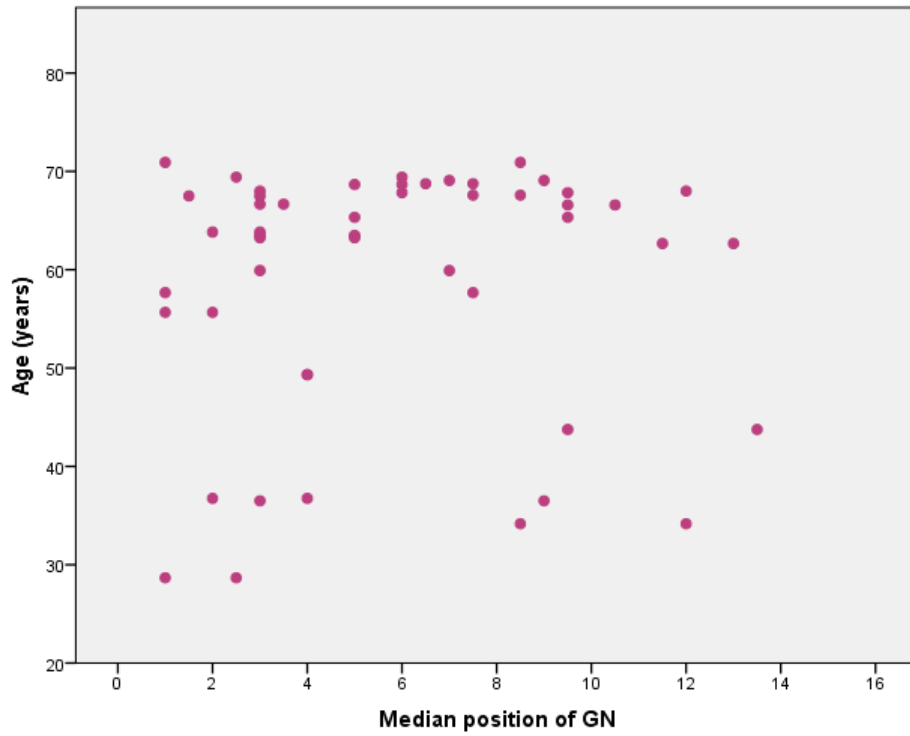
Figure 3.7 Scatterplot of Age and Median position of GN in the SGNT

The correlation was r=.05, *p*=0.74 n.s.
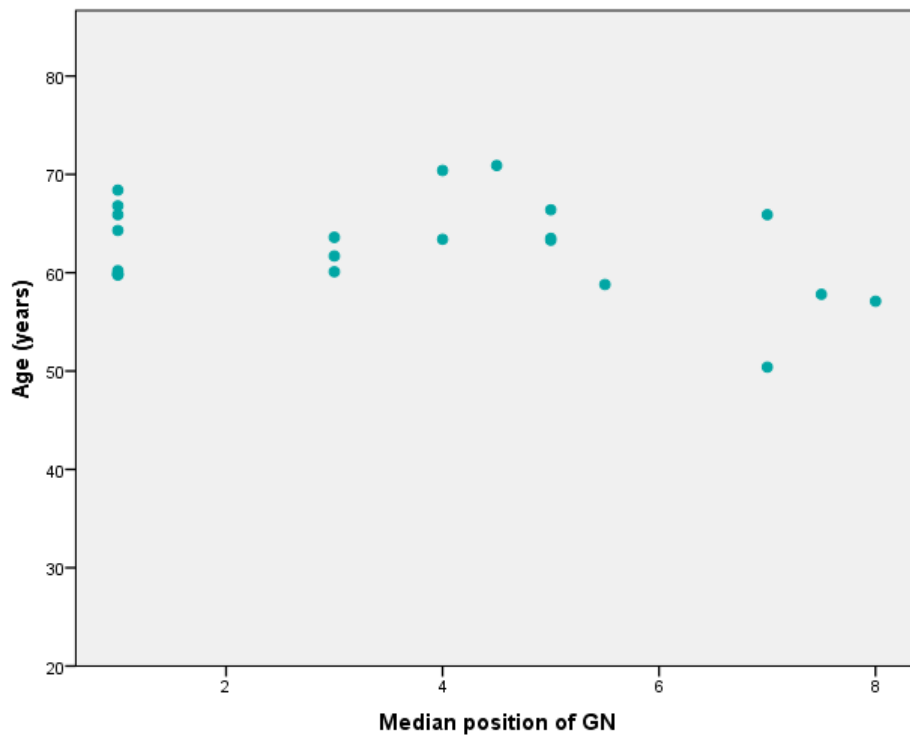
Figure 3.8 Scatterplot of Age and Median position of GN in the LMT

The correlation was r=-.30, *p*=0.18 n.s.
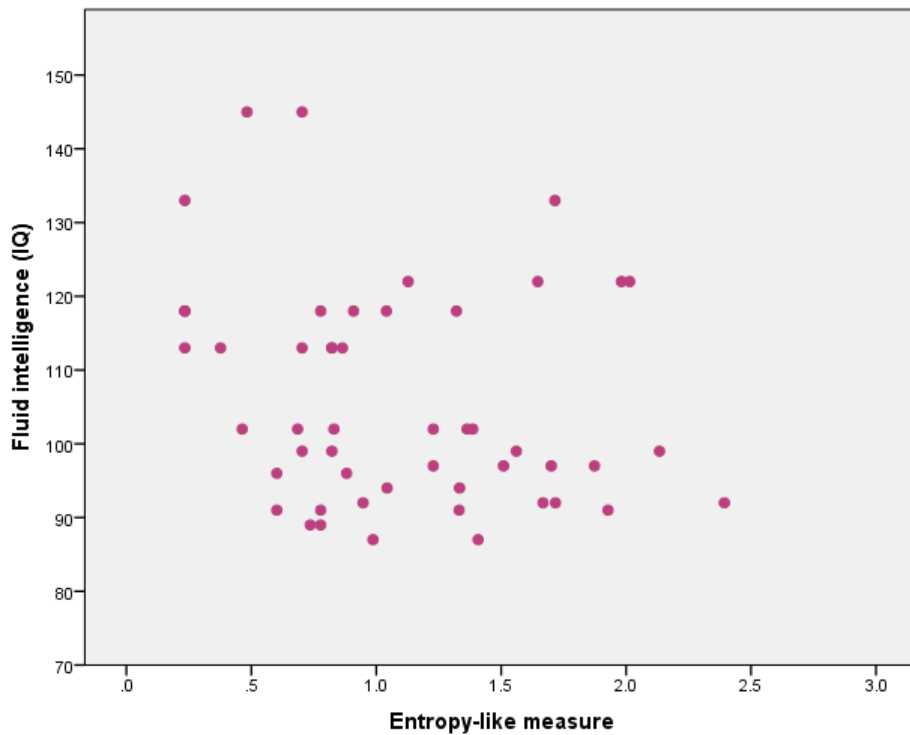
Figure 3.9 Scatterplot of Fluid intelligence and *Entropy-like measure* in the SGNT

The correlation was r=-0.24, *p*=0.09 n.s.
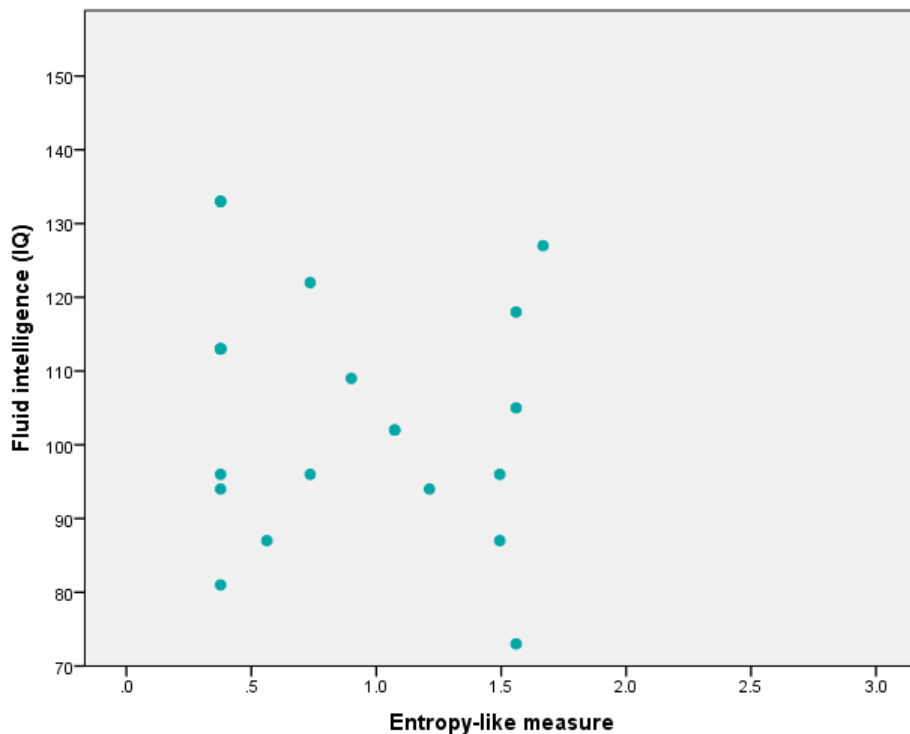
Figure 3.10 Scatterplot of Fluid intelligence and *Entropy-like measure* in the LMT

The correlation was r=-.20, *p*=0.37 n.s.
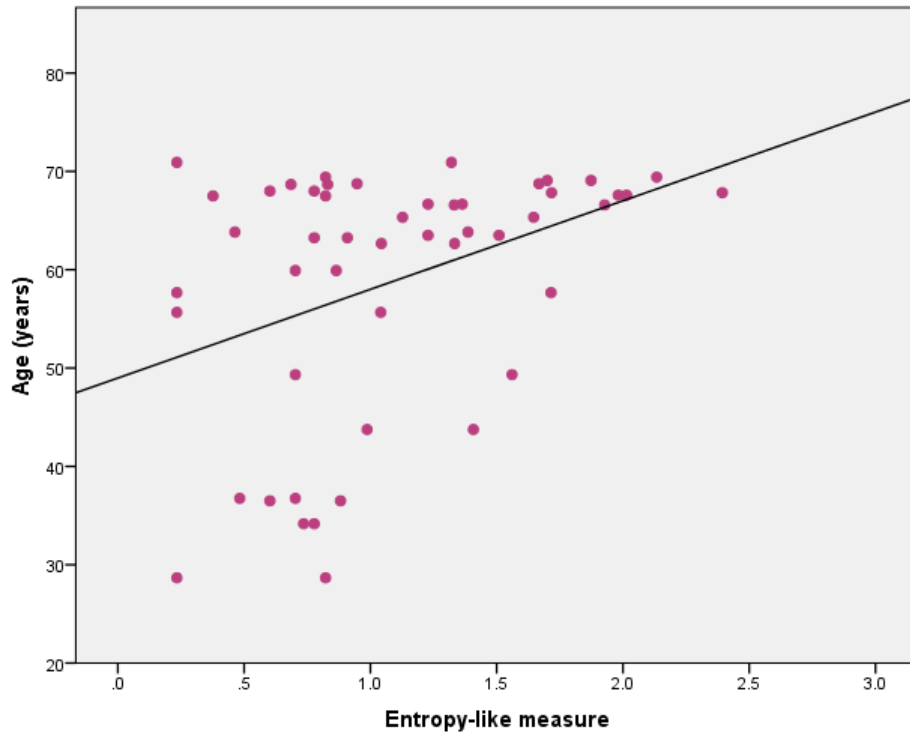
Figure 3.11 Scatterplot of Age and *Entropy-like measure* in the SGNT

The correlation was r=.39, *p*=0.005**.
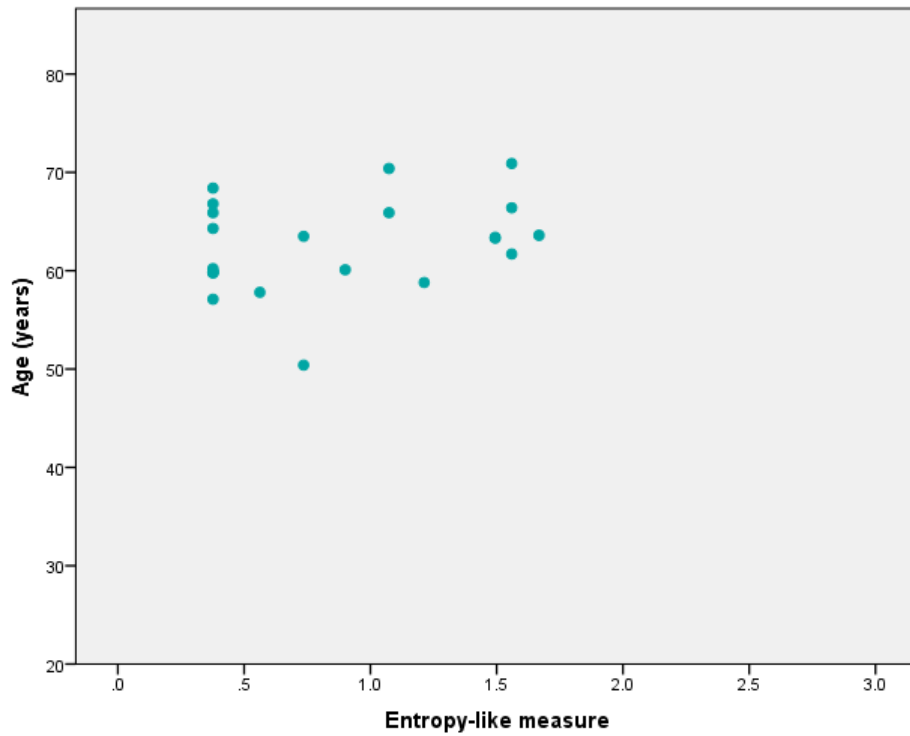
Figure 3.12 Scatterplot of Age and Entropy-like measure in the LMT

The correlation was r=0.28, p=0.21 n.s.

Table 3.5 Output from 2 regression analyses for the SGNT to predict: (1) the median position of GN and (2) overall entropy-like value including individual differences as IVs

	B	SE B	β	p
DV=Median position of GN				
Constant	15.81	3.93		.000**
Fluid intelligence	-.09	.03	-.39	.003**
Age	-.06	.04	-.20	.132
GN	.20	.18	.19	.274
Entropy-like measure	2.34	1.11	.36	.040*
Transparency	-.82	.88	-.12	.359
DV=Entropy-like measure				
Constant	.37	.56		.508
Fluid intelligence	.00	.00	.02	.870
Age	.01	.00	.15	.128
Number of GN polynomial 1st degree	2.03	.41	.53	.000**
Number of GN polynomial 2nd degree	-1.35	.35	-.35	.000**
Number of GN polynomial 3rd degree	.29	.35	.08	.403
Median position of GN	.05	.02	.34	.003**
Transparency	-.03	.11	-.03	.791

*p<.05, **p<.01; B=unstandardised beta value, SE B= standard error of B, β=standardized beta value
Model predicting *Median position of GN*, R²=.40;
Model predicting *Entropy-like measure*, R²=.70.

Table 3.6 Output from 2 regression analyses for the LMT to predict: (1) the median position of GN and (2) overall entropy-like value including individual differences as IVs

	B	SE B	β	*p*
DV=Median position of GN				
Constant	17.02	6.72		.022*
Transparency	-.76	1.05	-.16	.480
Fluid intelligence	-.02	.03	-.15	.513
Age	-.20	.11	-.40	.085
GN	.42	.35	.30	.250
Entropy-like measure	1.43	1.27	.30	.279
DV=Entropy-like measure				
Constant	.86	.83		.320
Transparency	.12	.11	.12	.270
Fluid intelligence	.00	.00	.06	.642
Age	-.01	.01	-.04	.716
Number of GN polynomial 1st degree	1.34	.27	.58	.000**
Number of GN polynomial 2nd degree	-1.65	.25	-.71	.000**
Number of GN polynomial 3rd degree	-.26	.23	-.11	.292
Median position GN	.01	.03	.07	.624

*$p<.05$, **$p<.01$; B=unstandardised beta value, SE B=standard error of B, β=standardized beta value
Model predicting *Median position of GN*, $R^2=.41$;
Model predicting *Entropy-like measure*, $R^2=.89$.

3.4 DISCUSSION

The main entropy-like measure analysis suggests that in the SGNT, GN was more clumpy if it mostly occurred earlier in the task and less clumpy if the majority of GN was observed later on, even when the number of GN was accounted for. This result is congruent with the hypothesis that participants who are mostly failing at the start are likely to be doing so primarily because of a weak or incorrect *task model.* Instead, those who are performing their worst towards the end of the task, are more likely to be failing mostly due to *monitoring-like* impairments.

The SGNT result supports the idea that GN may reflect failures of at least two separate mechanisms:

(i)   The *task model* which acts as a blueprint of the rules and actions to be performed. The *task model* is shaped early on in the task. A failure of the *task model* may be driven by the competition between sub-components which may make it more likely for a weak or incorrect version of the *task model* to destabilize, thus leading to GN.

(ii)   *Monitoring-like* mechanisms, which are attentionally-costly processes and are involved in the continuous checking of the environment for appropriate triggers (GN cues).

However, the SGNT result was not replicated in the LMT which means this interpretation should be read with caution. The lack of replication may be explained by low power issues: the sequence of neglectable trials in the LMT was half the length of the one in the SGNT and the overall frequency of GN was lower in the LMT compared to the SGNT. Hence, in the LMT, the degrees of freedom in which it was possible to test the hypotheses of temporal distribution using the entropy-like measure were radically reduced.

There may be alternate interpretations to the SGNT finding. For example, the clumpier distribution of GN at the start may simply reflect a learning curve in which more errors are observed at the start than at the end. However, this learning hypothesis would not be able to account for the cases in which GN was more frequent towards the end of the task.

According to the Bhandari and Duncan study (2014) an incorrect *task model* can explain cases in which GN occurs mostly at the end of the task. They revealed that for participants who showed major performance failures, the task dynamics were characterized by almost equal probabilities of success or failure on the first practice trial which then rapidly stabilized into a consistently incorrect pattern of performance. This finding leads to the expectation that cases of GN with a high median position should be fairly clumpy which, overall was not the case according to the result of the linear regression on my data. However, this inconsistency could be explained by the fact that the authors (Bhandari and Duncan, 2014) used a subset of the data which biases their sample to only one type of major performance failure profile and exclude other temporal signatures including more sporadic, *monitoring-like* errors. The fact that my analysis was

run on this unbiased and more heterogeneous set of temporal profiles may have cancelled out the possibility to detect a *task-model*-consistent result for cases with relative high error rates towards the end of the task. In future work, I could formally test this by sub-setting the data similarly to Bhandari and Duncan (2014).

Another finding was that participants with poorer scores of fluid intelligence showed more GN towards the end of the SGNT. A possible interpretation is that participants with lower scores of fluid intelligence are primarily neglecting because of impaired *monitoring-like* mechanisms. However, this finding could be driven by cases in which performance is generally quite poor, and tends to get worse with time which would resemble the pattern found by (Bhandari and Duncan, 2014) – one which would better fit with the idea of a poor *task model* account. Nonetheless, this result should be interpreted with caution given that (1) the result did not replicate in the LMT, and (2) both LMT and SGNT regression models only explained 45% and 37% of the variance of the median position of GN.

Neglecting participants showed intact memory recall of the instructions both before and at the end of the task. Therefore, any "lost" or "inactivated" task components of the *task model* are lost or inactivated only temporarily. How could this occur? My proposed how-possible explanation is that the *task model* may have two modes, *dormant* and *active* and that retrieval processes may be crucial at transforming one mode to another. What follows is an elaboration of this *dormant-active* model. The active mode guides task performance in real-time, and because of other real-time demands (e.g. motor coordination) it necessitates being a compact version of the dormant mode, the latter being less capacity-constrained but also less readily accessible. The active mode can be thought of as form of working memory (WM), whilst the dormant form is stored in long-term memory (LTM). In this framework, the "lost" or "inactivated" component exists in the dormant *task model* where its potential to be copied/summarised to an active form may be partly dependent on internal control mechanisms (e.g. cognitive effort, fluid intelligence), but also partly on retrieval mechanisms via external cues. Research suggests that retrieval can modify memories - successful retrieval of memories can facilitate their future retrieval whilst at the same time lead to the forgetting of their unretrieved associate pairs (Anderson et al., 1994; Wimber et al., 2015). Following this principle, retrieval of a task component may facilitate its implementation in the active mode of the *task model*, which in turn enhances

the detection of relevant stimuli in the environment and so on, in a self-perpetuating way. Hence, task conditions which affect retrieval may be important in this translation, as evidenced by the transparency effect reported in Chapter 2. Transparent cues are by definition more similar to the task component they refer to hence, in line with the encoding specificity principle (Tulving and Thomson, 1973), are more likely to be retrieved than the Nontransparent cues which in turn may be suppressed and hence neglected. In other words, retrieval may be pivotal in the recovery of GN by restoring the neglected component from the dormant to the active *task model*. Of course, this *active-dormant task model* hypothesis is just a potential explanation and requires further empirical investigations.

Apart from the transparency effect, this *active-dormant task model* account could also explain how GN can co-exist with a perfect memory recall of the instructions. The cues used to test the memory of instructions at the start and end of the task are qualitatively more transparent (verbal questions asked by the experimenter) compared to the cues used in the actual task (critical cue symbols). In addition, the recall phases, are not embedded in a context that is demanding like the one during performance (in line with previous ideas, Duncan et al., 1996). In sum, the transparency of the recall cues and lack of real-time demands make it more likely that at memory recall of instructions, retrieval is of the dormant, not active, *task model.* This predicts that if it were possible to gauge the memory of task instructions *during* task performance, then this memory recall would be highly correlated to GN.

This *active-dormant task model* hypothesis originates in part from separate evidence and existing theories of WM. For example, via a monkey electrophysiological study, Stokes and colleagues (2013) have shown that during the delay period in a WM paradigm, neural activity representing these memory traces is not sustained consistently – suggesting that these memory traces may enter some latent inactive form. A recent visual working memory fMRI study (Sprague et al., 2016) revealed that both memory recall and its neural representation were enhanced depending on the quality of the "retrospective" cue which preceded the memory probe. The cue was either meaningful, that is, it was closely associated to the identity of the memory item (but not the actual spatial position to be probed) or, neutral, which was entirely uninformative. It was the former condition which showed improved performance. This finding is congruent with the transparency effect

that I report in Chapter 2 in so far that it suggests that close associations may be crucial at reactivation of latent memories. However, to be precise, the *active-dormant task model* I propose would need to be formally tested by modifying the paradigm by Sprague et al. so that the neutral retrocues are replaced with Nontransparent variants. My prediction is that the results will be similar to the ones found for the neutral retrocues, such that Nontransparent retrocues will show both decreased fidelity of fMRI signal reconstruction and, decreased behavioural performance when compared to the Transparent retrocues. Finally, there is also growing support for the idea that early on in a novel task WM is in a dynamic state and readily shaped by current task demands (Duncan, 2001; Bhandari and Duncan, 2014; Bhandari and Badre, 2016).

Is the *task model* sufficient to explain GN? Supposedly, a defining feature of *the task model* account is that at some point early on in the task it stabilizes. Looking at the temporal dynamics of the critical trials data on a case-by-case basis, suggests that even if I applied a relatively conservative criterion for *task model* stability such as "several" runs of correct critical trials, then there would still be many cases where a significant proportion of neglect occurred towards the end of the sequence. Therefore, the answer is likely to be no. The reasoning is that assuming that *task model stability* is permanent, any subsequent performance fluctuations must be driven by other control mechanisms. These could include *monitoring-like* mechanisms in line with the ones proposed by the PM research which suggest an active and continuous search in the environment for relevant cues. Given that *monitoring* mechanisms are attentionally costly (McDaniel and Einstein, 2000; Smith, 2003; Einstein et al., 2005; Scullin et al., 2010), these are likely to be competing for resources with the real-time demands of the ongoing task, in line with dual-task interference (Allport, 1980 as cited in Duncan et al., 2008). These demands may wax and wane as ongoing task items of different difficulty are presented across the task, thus explaining the unstable zig-zag like pattern of these errors. It may be more costly to *monitor* for Nontransparent cues compared to Transparent cues in a similar way to how "nonfocal" cues are more costly than "focal" cues (Einstein et al., 2005; Scullin et al., 2010). Insufficient attentional resources may not be the only limitation to *monitoring* success. One other possibility is based on the sampling frequency – there could be time windows in the task which carry more crucial information than other parts. Hence, maximizing sampling in these time windows and saving resources at other times may be

another way in which monitoring may lead to successful performance without a change in the overall amount of attentional resources used. As mentioned in Chapter 1 (p.13), fluctuations in attention prior to an event have been associated to GN (De Jong et al., 1999; West and Craik, 1999; West and Alain, 2000; Unsworth and McMillan, 2014).

Having said all this, it is possible that a *task model* does not in fact reach permanent stability (for example see Westbrook et al., 2012 for data which may point in this direction) – in which case it would be more difficult to distinguish from a *monitoring-like* account.

On a somewhat separate note, it is worth remembering the dangers of over-theorizing. The models (*task model, monitoring, active-dormant*) discussed in this chapter are abstractions shaped by data and are useful in so far as they can aid with the integration of knowledge and guide future investigations. A consequence of their abstract and novel nature is that they may reappear in different guises across the literature. For example, a theory which in part appears to merge the *task model* and *monitoring-like* accounts together is what Marsh and colleagues *(2003)* refer to as "attentional allocation policies*"*. These policies refer to explicit and implicit processes via which the participant may determine at the outset which sub-tasks to allocate more attentional resources to. The strategic-like planning component of these *attentional allocation policies* resonates with the idea of the *task model*, whilst the component referring to the specific allocation of attentional resources is more similar to *monitoring* mechanisms.

It is also possible to conceive of a unified account of all three models discussed (the *task model*, *monitoring* and the *active-dormant* model), as follows: if the task model is labile, such that the elements may be gained or lost from it over time (which would be equivalent to an *active* vs. *dormant* modality), then this might produce a zig-zag pattern of errors as predicted by the *monitoring* account. The question remains whether this unified account would be equivalent same as the *monitoring* account or distinct from it.

Finally, it may be worth mentioning here that an alternate method to explore the temporal distributions of errors is to use a Fourier series which would allow the expression of a time series of errors in terms of their component frequencies (harmonics) (see http://www.falstad.com/fourier/ for a simple demonstration of a Fourier series).

3.5   CONCLUSION

This chapter revealed that GN is unlikely to be explained by one single control mechanism. This investigation was driven by two theories of GN, the *task model* and *monitoring-like* mechanisms, which predict differing temporal distributions of errors. This was tested using a novel method based on an entropy-like measure of clumpiness of errors. Results from the SGNT indicated that cases predicted to be explained primarily by a faulty *task model* by displaying the majority of GN early on, did in fact show clumpier error distributions. Instead, those which were thought to be failing mostly because of *monitoring-like* impairments with most of their GN appearing towards the end, did show more zig-zag like patterns of error. The results did not replicate in the LMT possibly because the model is incorrect, or due to low power or both.

I conclude that initial task performance is probably best explained by what Duncan (2008) refers to as the *task model,* and add that, once this *task model* stabilizes, failures in *monitoring-like* mechanisms are likely to explain the subsequent errors. Furthermore, I propose that conditions affecting retrieval, such as cue transparency, may be key to the recovery of GN. One suggested mechanism which I refer to as the *active-dormant model*, is that the degree of cue transparency may trigger task knowledge to be transformed from a latent, to an active state and thus reinstate the significance of the critical cues. In the context of the *monitoring* account, the transparency effect can be explained as insufficient attentional resources allocated to monitor for the more demanding Nontransparent cues, in line with PM research (McDaniel and Einstein, 2000; Smith, 2003). Finally, analyses on individual differences suggested that participants with poorer fluid intelligence were more likely to show that the majority of their failures appeared later in the task. However, it is not entirely clear which model is most likely to be affected by fluid intelligence.

Chapter 4. INSTRUCTIONAL VS. ACTUAL TASK COMPLEXITY

4.1 INTRODUCTION

Experiments in Chapter 2 revealed that changing Actual Task Complexity (ATC) by manipulating the degree of transparency of the critical cues affected GN. This is in contrast to previous findings which suggested that GN is immune to real-time demands (Duncan et al., 2008; Bhandari and Duncan, 2014). This novel result raises the possibility that these previous manipulations of complexity may not have been sufficiently large to affect GN, thus the null results. Alternatively, it raises the possibility that the particular type of complexity is important. For example, Chapter 2 involved manipulating the complexity of critical-cue-processing, whereas previous experiments (Duncan et al., 2008) manipulated the complexity of the regular sub-task (ongoing task). A separate issue is whether the Instructional Complexity (IC) effect reported in the same series of published experiments using the LMT (Duncan et al., 2008) would replicate in the newly developed SGNT. These two concerns motivated the two experiments presented in this chapter. In Experiment 3, I manipulate ATC and IC in the SGNT. Results indicate the presence of an IC effect but do not reveal any ATC effects in the SGNT, thus replicating the published findings by Duncan and colleagues. To further test for any ATC effects on GN, in Experiment 4 (ATCv2), I use a different real-time manipulation of complexity in the SGNT and, again, I find no effects on GN. Together, Experiments 1-4 and an analysis on local effects in the LMT presented in this chapter, show different effects of ATC on GN. I go on to discuss potential ways to reconcile these findings and link them to associated findings in the PM literature.

4.2 EXPERIMENT 3 – SGNT WITH INSTRUCTIONAL COMPLEXITY (IC) AND ACTUAL TASK COMPLEXITY (ATC)

The task used in this experiment is the SGNT which is described in Chapter 2 (p.45). Additional modifications to this task are detailed in the *Method* section below.

4.2.1 METHOD

4.2.1.1 Participants

A total of 64 participants were tested in this study and selection criteria were the same as for Experiment 1 (p.45). To avoid confounds due to practice effects, participants never performed the SGNT in any version more than once. For distributions of Age and Fluid Intelligence please see Table 4.2.

4.2.1.2 Task and experimental manipulation

There were three experimental conditions, which are summarised in Table 2.1 and are a result of combining 2 factors (IC, ATC) each having 2 levels (Low load, High load).

Table 4.1 IC and ATC experimental conditions

Experimental Condition	IC factor	ATC factor	N
Low Low	Low load	Low load	24^
High Low	High load	Low load	24^
High High	High load	High load	16

^Originally, there were 16 participants in each experimental condition within the same testing phase. However, I collected additional data on the *Low Low* and *High Low* conditions at a subsequent testing phase and given that the main results did not change I collapsed these data together for added power.
IC (Instructional Complexity)
ATC (Actual Task Complexity)

The IC factor manipulation referred to either the inclusion or exclusion of an instruction to an extra rule in the SGNT. In the high load IC condition, participants were told that if they saw displays with three photographs they had to judge which of the lateral blurred photographs, left or right, matched the facial expression of the central one (Figure 4.1). This extra rule required the use of the same button box used for the Emotion sub-task (Figure 2.4), with the left and right buttons corresponding to choosing the left or right photograph. The low load IC condition was identical to the original SGNT in which there

was no mention of this extra three-photographs-rule. In fact, the 24 participants in the *Low Low* experimental condition are a subset of those in Experiment 1.

Figure 4.1 Extra rule in the SGNT

The extra rule referred to new target displays which featured three photographs of faces. This extra rule required judging which of the lateral blurred photographs matched the facial expression of the central photo. The display duration was of 1 second.

The ATC factor manipulation referred to the inclusion or exclusion of "extra rule" trials in the actual task. In the low load ATC conditions, the trials were identical to the ones in the original SGNT with no extra rule trials. In the high load ATC conditions, 16 neutral trials were replaced by 16 extra rule trials. These replacements occurred at fixed positions and in such a manner that ensured that the minimum gap of 3 neutral trials (6s) was maintained between response trials. The duration of the extra rule target display was of 1 second, which is 400ms longer than a standard target (see Figure 2.1). This also meant that the task in the high load ATC condition was, in total, 6.4s longer than the original (low load ATC) version.

4.2.1.3 Stimuli and procedure

To create the extra rule trials, blurred versions of the target stimuli were produced. This blurred effect was achieved by applying a "glass" filter using Power Point (Microsoft, 2012) with transparency and scaling settings of 60% and 34, respectively. In high load IC conditions, the extra rule was introduced after explaining the Emotion sub-task but before describing the critical sub-tasks. As in the original SGNT, the memory for instructions was tested at the start and the end of the task. Importantly, after the initial recall of instructions and just before running the task, participants in the *High Low* experimental

condition were told that there would be no extra rule trials and hence they should temporarily disregard this rule. At the very end, after the final cued recall of instructions, a short block of 48 trials (just under a minute) which included extra rule trials was presented to all participants. Data from this final block were not used in the main analysis. The purpose of this final block was to provide an opportunity for participants in the *High Low* condition to execute the extra rule which they had been told to ignore in the main experiment and thus avoid leaving these participants under the impression that they had been deceived.

4.2.2 RESULTS

To rule out that differences in GN across the 3 experimental conditions were attributable to differences in age or fluid intelligence, two separate one-way ANOVAs were run. As reported in Table 4.2, there were no significant between-group differences in age nor fluid intelligence. Hence, I went on to test for GN differences.

Table 4.2 Statistics and ANOVA results for age and fluid intelligence across experimental conditions

	Experimental Condition	N	Mean(SD)	*F(df)*	*p*
	Low Low	24	108.2(14.7)		
Fluid Intelligence	*High Low*	24	105.2(18.0)	0.4(2,61)	.66, n.s.
	High High	16	101.4(11.2)		
	Low Low	24	59.5(7.8)		
Age	*High Low*	24	60.5(5.9)	1.0(2,61)	.39, n.s.
	High High	16	61.5(6.3)		

SD=standard deviation, *df*=degrees of freedom, n.s=not statistically significant,
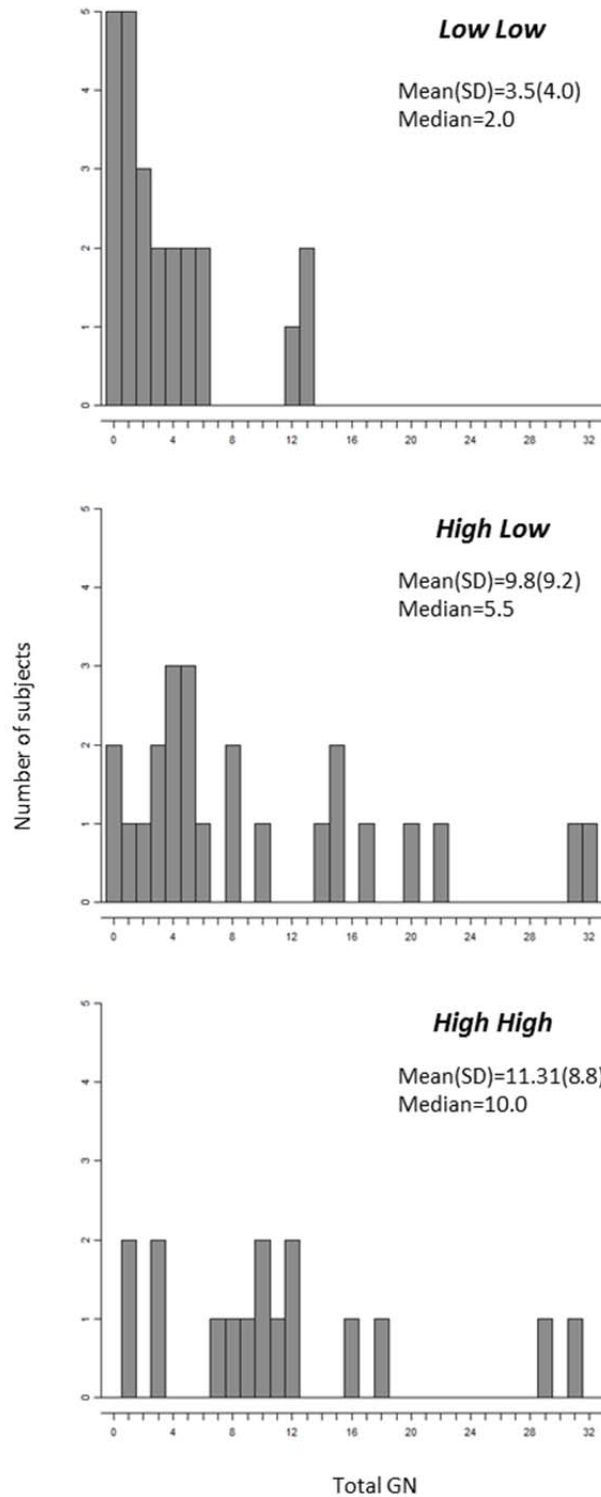
Figure 4.2 Distributions of total GN across the 3 experimental conditions of IC and ATC

The first histogram illustrates the distribution of total GN in the *Low Low* experimental conditions and similarly for *High Low* and *High High* conditions. The *Low Low* condition is when both instructions and execution involve only the basic set of instructions; the *High Low* condition is when the instructional stage involves basic + an extra set of instructions, whilst task execution only requires the basic set of rules; the *High High* condition is when both instructions and execution require both basic and extra set of instructions.

113

The distributions of GN in the three different experimental conditions are shown in Figure 4.2. Pairwise comparisons using Mann-Whitney-Wilcoxon tests revealed that *Low Low* vs. *High Low* conditions ($W=147.0$, $Z=-2.92$, $r=-.42$, $p=.007$) and *Low Low* vs. *High High* conditions ($W=76.0$, $Z=-3.22$, $r=-.51$, $p<.004$) were different. Instead, conditions of *High High* vs. *High Low* were not statistically different ($W=164.5$, $Z=-.76$, $r=-.12$, $p=.45$). All reported *p*-values are Holm corrected for multiple comparisons.

If the ATC manipulation in the High load condition was successful at increasing real-time attentional demands compared to the Low load condition, then this should be reflected in poorer performance on regular trials. To test this I compared the distributions of accuracy and reaction times on correct regular trials. across the three experimental conditions (
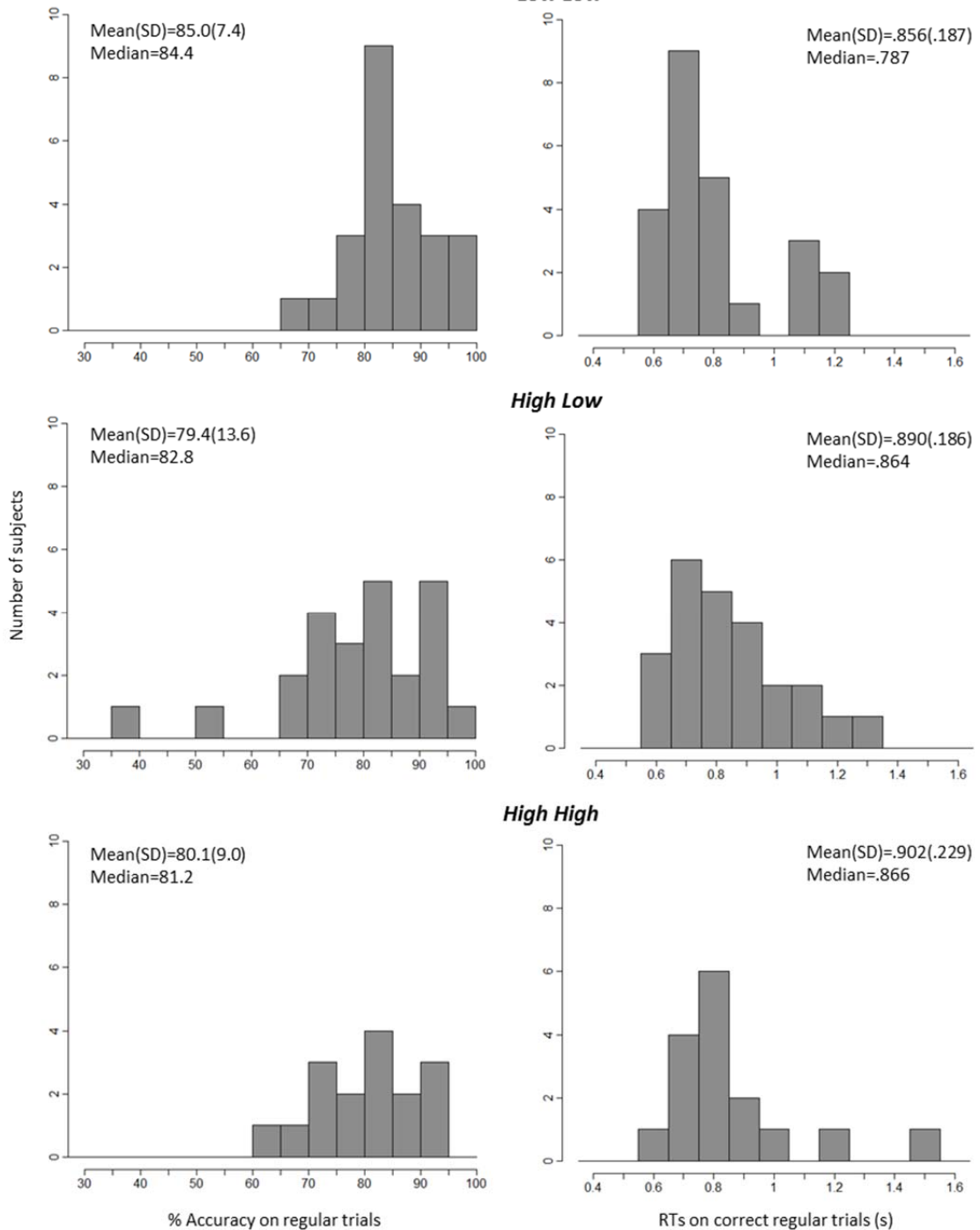
Figure 4.3). Correct regular trials were precisely defined as a sad keypress for a sad target or a happy keypress for a happy target and accuracy was expressed as a percentage out of the total of 32 regular response trials. Reaction times were mean values for latencies on correct regular response trials. Two separate Kruskal-Wallis tests revealed that there were

no between-group differences in accuracy $\chi^2(2.0)=3.1$, $p=0.21$ nor in reaction times $\chi^2(2.0)=1.1$, $p=.57$. This result suggested that the High load ATC manipulation may not have been not sufficient to increase the real-time demands of the task.

Although not central to the investigation in this current chapter, I was interested to test if the transparency effect reported in Chapter 2 was also found in the new SGNT data. Wilcoxon Signed-Rank tests revealed that GN was significantly larger in the Nontransparent conditions in both the *High Low* condition (Nontransparent GN Median score=5.0, Transparent GN Median score=1.0) $W=27.5$, $Z=-2.53$, $r=-.22$, $p=.0089$ and the *High High* condition (Nontransparent GN Median score=8.0, Transparent GN Median score=1.5) $W=6.0$, $Z=-2.94$, $r=-.26$, $p=.0021$. The test was not applied to the *Low Low* condition because, as mentioned earlier, this was part of the same dataset reported in Chapter 2. In summary, the transparency effect was replicated in these new variants of the SGNT.
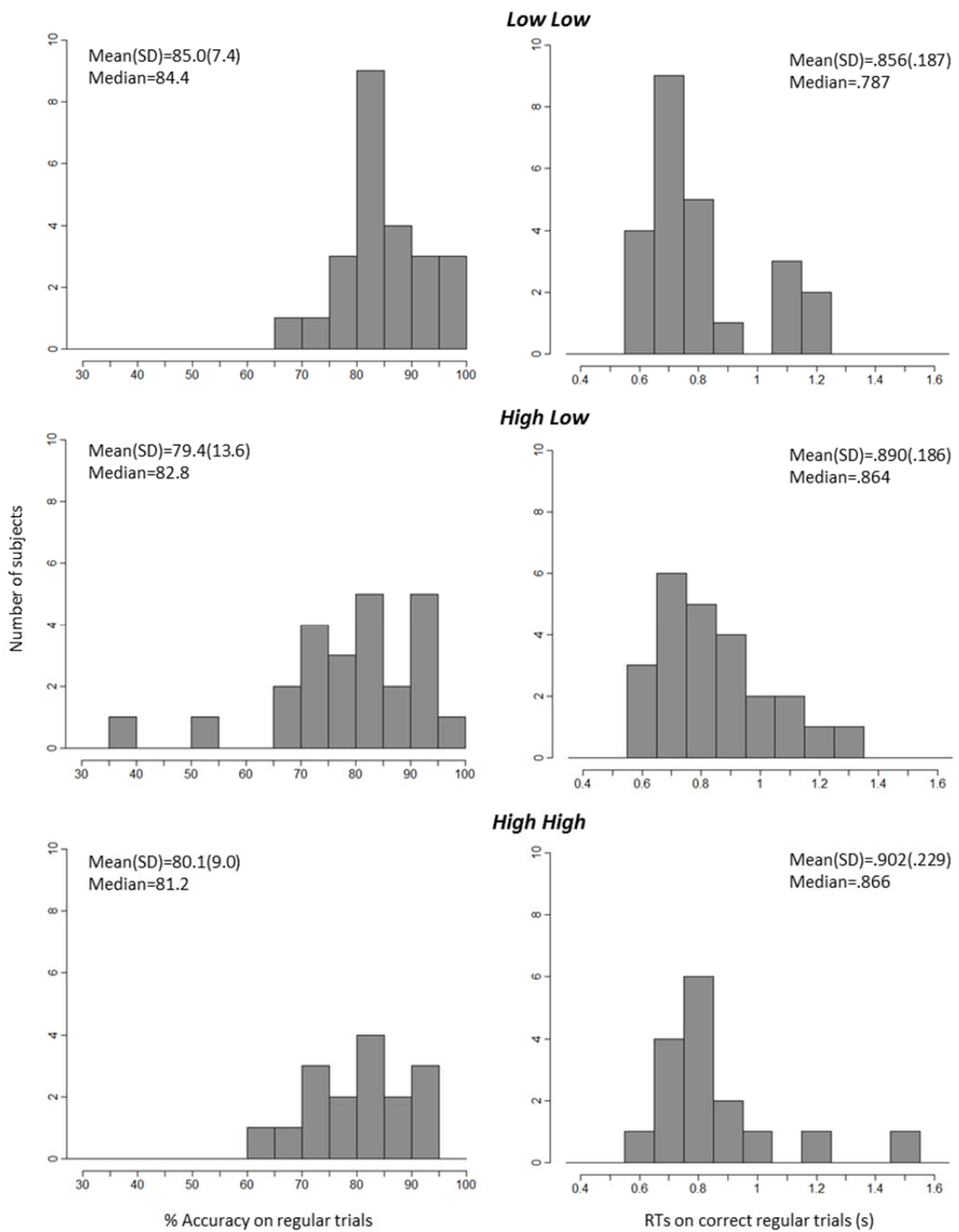
Figure 4.3 Distributions of accuracy and RTs for regular trials across the 3 experimental conditions of IC and ATC

The left column illustrates histograms of % correct regular trials whilst on the right, are distributions of RTs in seconds. The first row shows data from the *Low Low* condition; the second row *High Low* condition and the bottom row *High High* condition.

4.2.3   DISCUSSION

The results of this experiment indicated that it was sufficient for the instructional load to be *High* to observe a higher amount of GN errors. Even when the actual task execution was in a condition of *Low* load, but the instructions were in *High* load (*High Low* condition) then, the errors were significantly higher than the *Low Low* condition but, not different to the *High High* condition. In terms of IC, this experiment replicated previous results (Duncan et al., 2008) which suggest that GN is sensitive to the instructional load. Moreover, it also replicated the finding that increasing the real-time demands of the task by adding an extra trial type with a different rule did not affect GN. However, the ATC manipulation did not appear to have an effect on the regular trials, raising the question of whether the ATC load was large enough to cause any detectable effect on GN. For this reason, a second experiment was run in which a more effortful ATC intervention was used.

4.3   EXPERIMENT 4 – SGNT WITH ACTUAL TASK COMPLEXITY VERSION 2 (ATCV2)

4.3.1   METHOD

4.3.1.1   Participants

A total of 32 participants were tested in this experiment and selection criteria were the same as for previous experiments (see p.45). For distributions of Age and Fluid Intelligence see Table 4.3.

4.3.1.2   Task and experimental manipulation

The original SGNT was used which is described in Chapter 2 (p.45). There were two experimental conditions of ATC, *High load* and *Low load*, with 16 participants per group. The ATC manipulation was based on the perceptual degradation of targets (for a brief discussion on whether perceptual difficulty classifies as a manipulation of complexity refer to the General Discussion in this Chapter, p.128). In the *High load* condition 87.5% of the trials had degraded targets whilst in the *Low load* condition this was set at 12.5%.

What follows is a description of the distribution of degraded trials in the *High load* condition. Out of the 32 regular response trials, 28 were degraded and half of these were preceded by degraded neutral trials. Out of the 16 Transparent critical trials, 14 were degraded and half of these were preceded by degraded neutral trials. Similarly for the

Nontransparent critical trials. Finally, out of the 320 neutral trials, a total of 280 were degraded. Runs of consecutive degraded targets never exceeded 16 trials. The order of trials was fixed across participants. The distribution of degraded trials in the *Low load* condition was the inverse of the one in *High load* condition – the trials which were degraded in the *High load* condition were intact in the *Low load* condition and vice-versa.

4.3.1.3 Stimuli and procedure

The aim of the perceptual degradation in this experiment was primarily to increase the difficulty of the Emotion sub-task by making the facial expression harder to read. Degraded versions of the target stimuli were produced by masking the eye and mouth areas using 2 black rectangles (Figure 4.4).

Figure 4.4 Degraded target in the ATCv2

Two black rectangles were used to mask the face in degraded trials of the ATC, version 2.

The size of the 2 rectangles was kept approximately the same across degraded targets. There were a few faces that were visibly larger/smaller than average and hence required rectangular masks which were slightly larger/smaller to cover the mouth/eye areas in similar ways. After instructing participants on the regular and critical sub-tasks, they were told that some photos would appear masked and one example was used as practice.

Half the sample (n=16) was allocated to the *High load* condition, and the rest to the *Low load* condition.

4.3.2 RESULTS

Age and fluid intelligence did not differ between groups (see Table 4.3).

Table 4.3 Descriptive statistics and *t-test* results for age and fluid intelligence across experimental conditions in the ATC version 2

	Experimental Condition	N	Mean(SD)	*t(df)*	*p*
Fluid Intelligence	Low load	16	107.5(12.5)		
				.6(30)	.56, n.s.
	High load	16	104.8(13.8)		
Age	Low load	16	63.0(6.2)		
				.4(30)	.69, n.s.
	High load	16	62.0(7.9)		

SD=standard deviation, *df*=degrees of freedom, n.s.=not statistically significant

To ensure that task demands were overall higher in the *High load* condition, I compared the difference in accuracy on regular trials between the 2 experimental conditions (Figure 4.5).

A Mann-Whitney-Wilcoxon test indicated that accuracy was significantly lower in the *High load* condition $W=195.5$, $Z=-2.56$, $r=-.45$, $p=.01$. This established that the *High load* condition was generally more difficult than the *Low load* condition. Accordingly, I went on to test whether this increase in demand affected GN. A t-test revealed that there was no difference between the groups, $t(30)=-.18$, $r=.03$, $p=.86$. Although Shapiro-Wilk tests did not find GN to be non-normally distributed ($W=.96$, $p=.60$ for *Low load* and $W=.90$, $p=.09$ for *High load)*, the distributions appeared to be platykurtic (Figure 4.6) and hence I also ran a non-parametric version of the t-test. A Mann-Whitney-Wilcoxon test confirmed the results of the parametric test indicating no significant difference between the 2 groups: $W=128.5$, $Z=-.02$, $r=-.00$, $p=.98$.
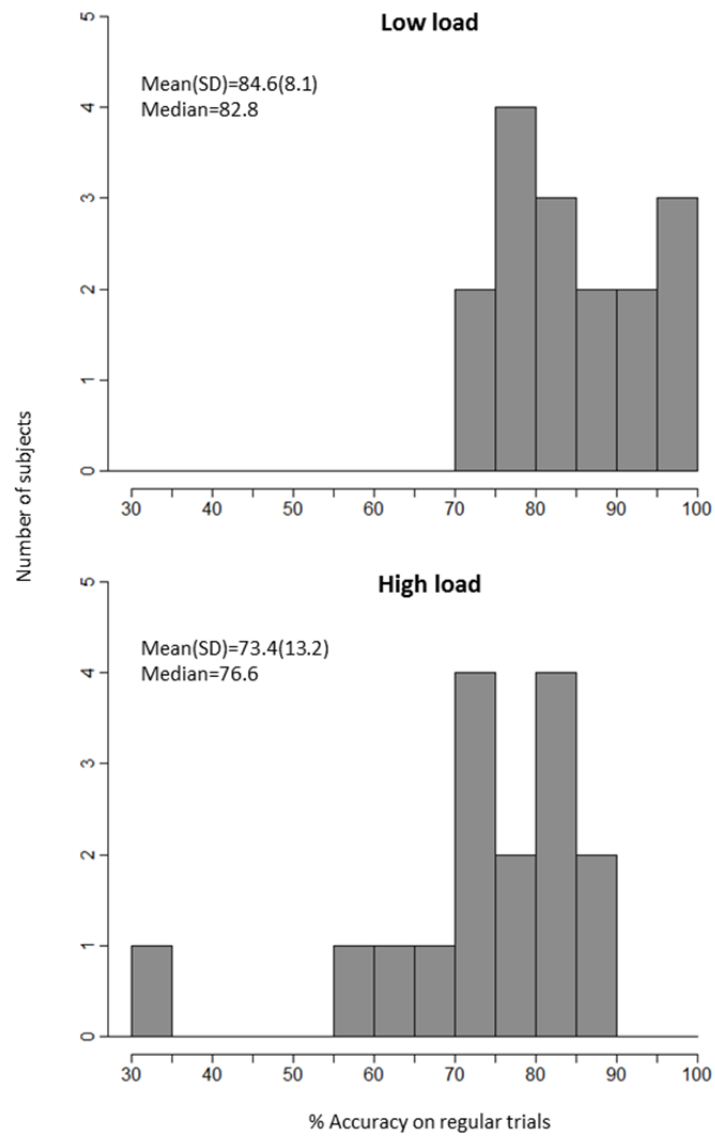
Figure 4.5 Distributions of accuracy data on regular trials in the SGNT, with ATC version 2

The *x*-axis indicates the percentage accuracy on regular trials, either in the High load condition (bottom histogram) or Low condition (top histogram). From these plots it appears that more errors were made under conditions of High load.
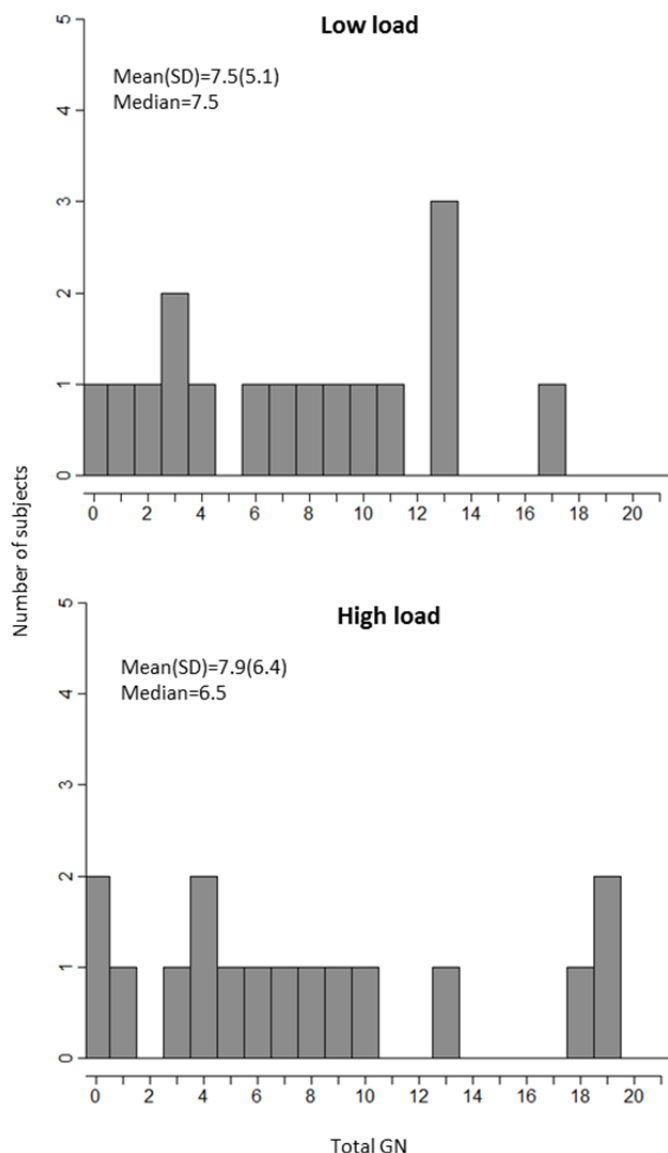
Figure 4.6 Histograms of Total GN across experimental conditions in the SGNT in ATC, version 2

There were 16 participants per condition of ATC: High or Low load. The maximum possible GN was 32, although the observed maximum in this experiment was 19. A t-test revealed no differences in the amount of GN across conditions.

This result indicates that despite regular trials being significantly more difficult in the *High load* condition, this did not appear to affect performance on critical trials in terms of GN. A more conservative version of this critical trials' analysis is one based on correct rather than neglected performance, where correct is defined as pressing the precisely correct button. Congruent with the GN analysis, a t-test did not reveal any significant between-group difference in terms of correct critical trials, $t(30)=1.42$, $r=.25$, $p=.16$ (*Low load* Mean(SD)=48.4(18.1)% correct vs. *High load* Mean(SD)=39.3(18.4)% correct), or

reaction times on these correct trials, $t(30)=.87$, $r=.16$, $p=.39$ (*Low load* Mean(SD)=1.496(.323) seconds vs. *High load*, Mean(SD)=1.409(.240) seconds). Overall, these congruent findings suggest that while real-time demands significantly affected performance on the regular trials, GN was immune to them.

On a separate note, I was again interested to test if the transparency effect replicated in this version of the SGNT. I ran a repeated measures ANOVA with Transparency (Transparent, Nontransparent) as within-subjects factor and ATC (Low, High) as a between-subjects variable. The mean (SD) number of neglect errors for conditions of Transparent –*Low-load* and Transparent-*High-load* was 1.75(2.96) and 1.88(2.63) and for Nontransparent-*Low-load* and Nontransparent-*High-load*, 5.75(4.95) and 6.00(4.75), respectively. This replicated previous findings of a main effect of Transparency, $F(1,30)=18.13$, $r=.61$, $p<.001$ and confirmed no main effect of ATC, $F(1,30)=.03$, $r=.03$, $p=0.86$ nor any interaction with ATC, $F(1,30)=.00$, $r=.00$, $p=.95$.

Does it matter whether a degraded or intact trial precedes a critical trial? To test this I ran a repeated measures ANOVA on GN in the ATCv2 with 3 factors each having 2 levels: Transparency (Transparent, Nontransparent), Complexity of the Critical target (High, Low) and Complexity of the Preceding target (High, Low). The result confirmed a main effect of transparency $F(1,31)$, $r=.57$, $p<.001$, but no other main or interaction effects *(F* and *p* values <1) suggesting that whether the critical target or the target preceding it were degraded or not, this did not affect GN.

4.3.3 DISCUSSION

This second version of the ATC was designed to ensure an overall increase in attentional demands in the SGNT. Indeed, trials on the dominant sub-task (Emotion) were significantly more difficult in the *High load* condition compared to the *Low load* condition. However, this still did not affect GN, nor accuracy or reaction times on correct critical trials. This result replicates previous published findings in which real-time demands did not affect GN (Duncan et al., 2008).

Did the second version of ATC manipulation affect the real-time demands of the SGNT uniformly? Probably not. The use of rectangular masks over the eye and mouth area are likely to have impaired the Emotion sub-task much more than either of the Colour and Gender critical sub-tasks. This is because judging the emotion of a face strongly relies on

information around the eyes and the mouth (Kestenbaum, 1992). However, judging the colour of the photograph or the gender of the face is less likely to be affected by the rectangular masks. Also, the manipulation used in the first version of ATC (extra rule with 3 photographs per display) was relatively separate/distant to the critical sub-tasks. This was also the case for published experiments (Duncan et al., 2008), in which the ATC manipulation was achieved by increasing the number of targets or increasing the number of rules and, excluded manipulation of critical trials. This contrasts with the manipulation of complexity in Chapter 2, in which critical trials were more directly affected by changing their transparency and this showed to significantly affect GN. Of course, increasing difficulty of an operation is likely to produce more errors in that operation, so the result of doing so could be obvious. It is worth clarifying here, that in the transparency manipulation, what was manipulated was the demand *surrounding* the critical operation rather than the operation itself. In other words, the GN resulting from the transparency effect reflected not just a failure to recall and execute the neglected rule in time, but also a failure to *notice something should be* recalled and executed. Overall, it seems like a pattern may be emerging in which GN is affected by ATC, if and only if, the distance/association to the critical operation, of the manipulated events surrounding the critical operation, is small/strong enough.

One additional way to test this hypothesis is to assess whether a highly demanding event which is temporally close to the critical event affects GN. The null results from the 3-way repeated measures ANOVA were not congruent with this prediction, which instead suggests that it does not matter whether the target preceding the critical trial was degraded or not. However, this analysis has an important confound given that by design, all trials preceding critical ones were neutral targets and hence did not require a response. In fact, all versions of the SGNT had at least 3 neutral trials (a minimum of 6 seconds) between each response trial. A degraded neutral trial preceding the critical trial is still effortful in as much as the degradation stands in the way of identifying it as a neutral target in the first place. However, unlike a degraded response trial, in a degraded neutral trial once the face is identified as neutral, no further processing is required given that a neutral target requires no response (no decision about which button to press etc...). Hence, the padding of relatively low processing demands via the use of a minimum of 3 neutral trials between critical trials and the weaker interaction of degradation on neutral trials preceding critical

trials, may have washed out any local complexity effect. Instead, to better address this question, in the next section, I look at data from the LMT task reported in Chapter 2 in which half the trials had number segments preceding the second side-cue and the remaining half, letter segments.

4.4 LOCAL EFFECTS OF COMPLEXITY IN THE LMT

The LMT, as described in Chapter 2 (p.65), was made up of two types of segments, reflecting two sub-tasks – numbers and letters. For all trials, the segments surrounding the second side-cue were never of the same kind: if a number segment preceded the second side-cue this was followed by a letter segment and vice-versa. Half of the trials had number segments first. Adding numbers was previously found to be significantly more difficult sub-task than reading letters (see Experiment 2 in Duncan et al., 2008). Given this finding, I went on to test whether processing a more demanding sub-task before the appearance of the critical event affects GN.

4.4.1 METHOD

I compared the distribution of GN in trials (from Experiment 2) which had number segments preceding the second side-cue to those which had letter segments instead. Considering that the distributions of GN were non-normal, I used non-parametric Wilcoxon-Signed Rank tests. The comparison was within trial type (switch or non-switch) and within condition of transparency (Transparent or Nontransparent), which resulted in a total of four comparisons. *P*-values were corrected for multiple comparisons using the Holm method.

4.4.2 RESULTS AND DISCUSSION

The results (Table 4.4) indicated that in the switch-trials only, but independently of the degree of transparency, if number segments preceded the second side-cue, then it was significantly more likely for this side-cue to be neglected than if it were preceded by a letter segment. I did not expect this finding to be replicated in the non-switch trials given that these hardly showed any neglect in the first place. Overall, this result suggests that the level of complexity of the task component which directly precedes the critical event affects GN and this is independent of the complexity of the actual critical event itself.

Table 4.4 GN as a function of segment type preceding the critical cue in the LMT

	Experimental Condition	Segment preceding 2nd side-cue	N	Mean(SD)	Median	W	Z	r	p
Switch trials	Transparent	Letter	29	.03(.09)	.00				
		Number	29	.12(.20)	.00	74.0	-2.77	.36	.02*
	Nontransparent	Letter	29	.28(.38)	.00				
		Number	29	.38(.37)	.20	131.5	-2.62	.34	.03*
Non-switch trials	Transparent	Letter	29	.01(.04)	.00				
		Number	29	.00(.00)	.00	.0	-1.63	.21	.20
	Nontransparent	Letter	29	.18(.23)	.00				
		Number	29	.16(.21)	.00	54.5	-.70	.09	.48

GN is based on the Mean Side Error (MSE) score, as described in Chapter 2 p.67
W=Wilcoxon Signed Rank statistic, Z=z-score, r=Pearson's r a measure of effect size.
*p-value<.05 (2-tailed)
All reported p-values are Holm corrected for multiple comparisons.

A possible explanation for this finding derives from the phenomenon referred to as the attentional blink, whereby participants cannot report the second of two targets that are presented within around 90-540ms of each other (Shapiro et al., 1997). However, this is unlikely given that there is a 600ms gap between the expiry of the number display and the onset of the side-cue display. Therefore, a more likely explanation is that GN is dependent, at least in part, on processes that require attentional resources which are sensitive to competing real-time demands - such as preparatory or monitoring mechanisms. Nonetheless, it is worth noting that this result is confounded by an order effect given that the allocation of number/letter segments were in fixed positions for all participants i.e. the first trial was always a number-second-side-cue-letter trial. In addition, switch trials had more number-second-side-cue-letter trials (5) than letter-second-side-cue-number trials. Hence, a future experiment could test this more reliably

by randomizing and counterbalancing the position of the letter/number segment order across time and across trial type (switch/non-switch).

4.5   GENERAL DISCUSSION

Experiment 3 replicated the IC effect reported in previous experiments (Duncan et al., 2008), whereby increasing the complexity of the instructions without changing the load of the actual task, showed a significant increase in GN. A suggested explanation for this IC effect is an impaired control structure referred to as the *task model* (for details see Chapter 3, p.77). As the complexity of instructions increases, the competition between task components in the *task model* also increases. This competition affects the integrity of this control structure leading to the observed GN (Duncan et al., 2008).

Interestingly, the PM literature may provide a way to unpack the *task model* account of the IC. Such an interpretation of the IC effect would suggest that participants may be unable to fully delete an instruction once they have encoded it, even if they have been asked to disregard it. The emphasis here is that if the to-be-ignored instruction continues to exist to some degree, then it can interfere with current task performance because it is spontaneously and involuntarily activated (Experiment 5, Einstein et al., 2005) and/or, because any associated *monitoring* mechanisms (for details see Chapter 3, p.79) are also likely to continue to operate, even if involuntarily and implicitly so. Evidence of an inability to fully deactivate instructions comes from various studies. The final PM experiment reported in a paper by Einstein et al., (2005) required participants to temporarily ignore the PM instruction during an embedded lexical decision task, before they return to the ongoing imagery rating task. Participants were found to be significantly slower on the lexical decision task trials which included PM targets, suggesting that it was not possible to entirely deactivate the PM rule on command. Other evidence comes from studies reporting the Intention Superiority Effect (ISE) (Goschke and Kuhl, 1993; Freeman and Ellis, 2003). The ISE is when *faster* responding to PM-target-related stimuli is reported during a task in which the PM instruction is deactivated. It is not entirely clear why the attempted deactivation of a PM intention in one study resulted in slowing down, whilst in others, in the speeding up, of processing of PM-intention-related stimuli. However, overall, the results suggest that it is hard to unlearn instructions efficiently. One important difference between these PM findings and the GN experiments is that in the

latter, the stimuli mapped to the to-be-ignored rule were absent. For example, in the high load IC condition of Experiment 3, the to-be-ignored rule referred to special displays featuring 3 photographs, unlike the rest of the trials which involved a single photograph. These 3-photo displays never actually appeared in the *Low Low* and *High Low* conditions which were used to gauge the IC effect. Instead, in the cited PM findings (Goschke and Kuhl, 1993; Freeman and Ellis, 2003), the task which required to ignore the PM intention actually contained the PM targets (words) associated to the to-be-ignored rule. Also, the presentation of the PM targets, was not largely different from the precise way in which they were presented in the version of the task in which the PM intention was active. Hence, these two features together probably increased the likelihood for the to-be-ignored rule to be reactivated. Despite these differences, it is hard to exclude the possibility that in the GN experiments the to-be-ignored rule was completely unlearnt and not activated in any way during performance. After all, the regular single photo displays were not categorically different from the unseen 3-photo displays.

Experiment 4 involved manipulating task complexity by perceptually degrading the target stimuli. Does increasing perceptual difficulty increase complexity? I would claim that it probably does, using a similar argument to the one presented in the introductory chapter (see p.33). In Experiment 3 the assumption was that as the number of task rules increased, so did task complexity. Degrading a stimulus may not appear to explicitly "add" anything to the task per se. However, additional or different cognitive steps are likely to be required to process such degraded stimuli (e.g. rotations or searching for particular edges; for evidence of use of such mental transformations see, Jolicoeur, 1988; Graf, 2005; Risko, 2015; Risko and Gilbert, 2016). In summary, I argue that at a descriptive level, the difference between these two examples of complexity manipulations is primarily at the resolution at which the manipulation occurs – one being fine-grained (perceptual degradation) and the other at a higher level of abstraction (extra rule). However, although these complexity manipulations have a tendency to increase task demands it is not entirely clear what the underlying cognitive stages and their durations are, and how such processes may differ across perceptual vs. rule based manipulations. Recent developments in fMRI methods using multi-voxel pattern analyses and semi-hidden Markov models appear to be a promising way forward to investigate these stages of mental operations (Anderson, 2016).

The manipulations of ATC in both Experiments 3 and 4 replicated the previously published findings (Duncan et al., 2008; Bhandari and Duncan, 2014) that real-time attentional demands do not affect GN. However, these results are in contrast to two strands of research findings which indicate that real-time demands *can* affect GN. The first strand refers to the findings that I report myself in this dissertation: the transparency effect (see Chapter 2) and the local effects of complexity in the LMT described in this chapter (see section 4.4). The second strand comes from the PM literature, where typically, increasing the attentional demands of the ongoing task does interfere with PM (Einstein et al., 1997; Marsh and Hicks, 1998; Marsh et al., 2002; Meier and Zimmermann, 2015). These PM findings are consistent with a *monitoring* account of PM: given that *monitoring* processes require attentional resources (McDaniel and Einstein, 2000; Smith, 2003; Smith et al., 2007), then any additional attentional demand should reduce available cognitive resources for the PM task leading to a PM failure (PMf). Surprisingly, even PM remembering driven by what researchers refer to as *spontaneous* processes (little or negligible costs detected as a result of the PM intention) has been found to be sensitive to real-time demands (McDaniel and Scullin, 2010; Harrison et al., 2014).

Can these mixed findings be reconciled? One possibility is that the *quality* of the ATC manipulation is crucial such that direct but not indirect complexity manipulations of the critical events affect GN. For example, in Experiment 3 the ATC load involved an extra rule which did not directly influence the setup of the critical events. Similarly in Experiment 4, even though the perceptual degradation (face masks) was applied to targets of regular and critical trials in equal proportions, the type of degradation was designed to affect the judgment involved in the Emotion sub-task, but not the judgments required in the critical sub-tasks (gender and colour). In addition, the complexity of the targets preceding critical trials was also not important to GN possibly because these were relatively unengaging trials given that, being neutral trials, they did not require a response. Instead, for the transparency and local LMT analyses, the complexity manipulation arguably has a much more direct effect on the critical events. The degree of transparency of the critical cue is an obvious case, since it changes the appearance of the cue signalling the critical event. In the LMT, the load is immediately preceding the critical event and requires a response. In sum, it appears that only if the complexity

manipulation is linked to some extent to the critical event episode then is it likely to affect GN.

How does load of the ongoing task affect GN? Some PM researchers suggest that real-time demands interfere with the retrospective component of PM, that is, participants simply forget "what" to do when the PM cue appears (Meier and Zimmermann, 2015). In their study, Meier and Zimmerman increased the load of the ongoing task by making it faster, which significantly increased the rate of PMf. Instead, the reaction time analyses I ran on the transparency experiments (Chapter 2) suggested that increasing the complexity of the GN task, by reducing cue transparency, not only affects the retrospective component (cue accessibility analysis) but also affects how easily participants noticed the critical cues (cue sensitivity analysis), which is equivalent to the prospective component ("when") of PM. In fact, later in their paper, Meir and Zimmerman acknowledge that they cannot exclude the possibility that the prospective component is also compromised.

Consistent with this "fail-to-notice" view is an interpretation offered by Harrison et al. (2014), who suggest that dividing attention can interfere with PM (even when this is of the *spontaneous* type) by interfering with the full processing of the critical/PM events. A dual-task could impair retrieval of cue-related information for both strategic and associated learning retrieval routes (Moscovitch, 1994), either by directly competing with the high-demands of strategic processes, or by insufficient parsing of the critical event to trigger the associated recall. Another possibility they consider is that competing ongoing tasks may block a PM intention from being processed consciously, which is the precise aim of my cue sensitivity analyses presented in Chapter 2. This idea was first proposed by Smith (2008) who argues that preparatory attentional processes are necessary for intact PM performance and that these are more likely to take up capacity in the "periphery", rather than the "focus" of awareness. The difference between "peripheral" and "focal" attention is analogous to what Wundt (1912/1973 as cited in Smith, 2008) referred to as apprehension and apperception, where apperceived contents are ones we are fully aware of, whilst apprehended ones only minimally so. Smith goes on to argue that preparatory attentional processes may occupy capacity not so much so as directly preparing for target detection or action preparation, but rather as a process of evaluating the optimal way to respond to the environment (Smith, 2008). The threshold between peripheral and focal attention may vary depending on various factors and adding a dividing attention task may

shift this threshold, making it more difficult for cue-related thoughts to enter the focus of attention (Harrison et al., 2014). Furthermore, this idea is supported by studies which report less involuntary autobiographical memories when participants are in states of focused attention (Kvavilashvili and Mandler, 2004), and with emerging ideas that shifting attention from a current cognitive task (e.g. mind wandering, imagination, mentalising) may involve neural networks (default mode network) which are distinct from regions which allow focused attentional episodes (fronto-parietal network) (Crittenden et al., 2015).

4.6 CONCLUSION

Experiments 3 and 4 (ATC versions 1 and 2) generally replicated previously published results in which increasing the IC significantly affected neglect, whilst changing ATC did not affect GN. However, the ATCv2 results are in contrast with 2 of my other findings which suggest that real-time demands *can* affect GN: the transparency effect and local effects in the LMT. These mixed results may suggest an emerging pattern in which ATC affects GN if (1) ATC is sufficiently large, and (2) if ATC is of the type which affects events linked or surrounding critical events episodes. Future GN experiments with different ATC manipulations could attempt to address this hypothesis. Overall, my ATC findings were largely linkable to findings and interpretations reported in the PM literature, which leads me to the question of the next chapter – how are PM failures different from GN?

Chapter 5. TASK STRUCTURE

5.1 INTRODUCTION

Prospective Memory failures (PMf) appear similar to GN. PM is commonly defined as remembering to perform an intended action or thought, at some future moment (Einstein et al., 1997). PM is thought to vary on a number of dimensions, with the main division classifying PM as having an either event-based or a time-based intention. In event-based PM, an event in the environment serves as a cue to prompt the execution of the intention; whilst in time-based PM the intention is to be performed after the elapse of a specific amount of time (for a review see Gonen-Yaacovi and Burgess, 2012). One way to view this distinction is that time-based PM involves processes which are thought to be more self-initiated; instead event-based PM is comparatively less demanding given that part of its processing is delegated to the occurrence of an external event. In GN, the critical event is cued by a stimulus in the environment and hence GN paradigms most closely resemble event-based PM. Therefore in this dissertation, unless otherwise noted, PM refers to the event-based type only.

GN and PM paradigms are similar because of two main features. First, the task requires at least 2 sub-tasks, one of which is dominant and one which applies to a minority of trials. In GN terminology, these are called the "regular" and "critical" sub-tasks, respectively. In PM, these are referred to as the "ongoing task" and the "PM intention/instruction". Secondly, if participants disregard the instructions to the critical sub-task the participant can still respond by treating the critical event as a regular one and thus display what is referred to as GN or PMf.

One apparent difference between GN and PM paradigms is the delay between the encoding of the PM instructions and the start of the task which in the PM literature is referred to as the "retention interval" (Gonen-Yaacovi and Burgess, 2012). Most lab-based PM tasks involve a retention interval which usually varies in the range of a couple of minutes to an hour. However, some studies use intervals ranging in the units of days

(Hicks et al., 2000). This interval period is sometimes filled with a "retention task" which can be something like a questionnaire and is used with the intention to prevent participants from rehearsing the PM instruction. The effect of the length of the retention interval and type of retention tasks on successful prospective remembering is mixed, sometimes with counterintuitive results (Hicks et al., 2000). In any case, there is also a significant body of work in the PM literature which uses experimental tasks with retention intervals in the order of seconds to a few minutes (for e.g. Einstein et al., 2005), thus making them structurally very similar to GN tasks in which there is minimal retention interval and no retention task. Hence, this retention interval/task does not appear to be necessary to define it as a PM task.

Other differences between GN and PM tasks are the amount of practice and the frequency of critical events. In GN tasks, practice is minimal whilst in PM tasks practice tends to be relatively extended, at least for the ongoing task. Critical/PM events tend to be more infrequent in PM paradigms compared to GN tasks. GN and PM paradigms may differ in various other ways such as the regularity of the critical event, speeded vs. self-paced tasks, rule types, stimulus set, quality of the cue, the mode of delivery of instructions and so on. However, these differences are also found between task variants within each paradigm. Overall, having also considered these differences, it is apparent that there are strong commonalities between GN and PM paradigms.

This structural similarity has two main implications. First, it offers exciting opportunities to constrain theorizing of the operations underlying GN/PM, as I have attempted to do throughout previous chapters in this dissertation. Second, it raises the question – what, if anything, makes these errors different? This question is especially relevant given that GN results, at least as reported in my experiments, appear to be largely compatible with data and accounts from the PM literature.

This leads me to the next point on the problems of empirical isolation and unnecessary parsimony which is neatly presented by a talk by Popov and Reder (2016) and a talk by Poldrack (2016) who has been working on these problems for almost a decade (Poldrack et al., 2011). Since the early 20th century, the function of the rate of publications in the field of experimental psychology is exponential, almost quadratic with a current rate of around 10,000 articles per year (Popov and Reder, 2016). In addition, there is a large

volume of studies each reporting all sorts of novel effects and phenomena. Unfortunately, it is often the case that these phenomena are not appropriately integrated into a common framework which allows a clear discernment of how they are related to each other. Some examples: "induced-forgetting effect" vs. "directed forgetting"; "fan effect" vs. "priming effect"; "self-monitoring failures" vs. "strategy implementation failures"; "perseveration errors" vs. "capture errors" and so on. This results in a fragmented understanding of the field with the overall result that we "don't know what we know". On the one hand, creating a new label for an observed effect is an easy way to refer to the specific phenomenon observed. On the other hand, this approach combined with a lack of awareness of other similar pre-existing results can lead to a false parsimony, that is, the false belief that there are many more distinct effects than there actually are. Instead, these differences would be better captured as subtypes or variants, of one or more generic phenomena. These concerns are therefore further motivation to attempt working towards an integration of what are otherwise isolated domains of GN and PM.

How can this process of integration into a common framework occur? The approach I present in this chapter is based on three stages. The first stage attempts to generate the errors under investigation: GN from a typical GN-like task structure and PMf from a typical PM-like structure. The second stage applies the experimental manipulation of Instructional Complexity (IC) that I expect GN to be sensitive to and test, for the first time, whether it also affects PMf. To anticipate the results of this second stage I find that PMf are insensitive to IC. In the final stage, I morph the GN-like and PM-like tasks together by manipulating two task structure parameters: the extent of practice and the frequency of critical events. I then apply the manipulation of IC as a "litmus test" with the following hypotheses: a GN task morphed into a PM-like structure will produce errors similar to PMf and hence these will be insensitive to IC, conversely, a PM task morphed into a GN-like structure will produce errors which are GN-like and hence these will increase with conditions of High IC.

In these experiments, I had two additional interests. The first was related to understanding if and how individual differences such as age and fluid intelligence would interact with the task structure. The second regarded the experimental entropy-like measure that I introduced in Chapter 3 as a way to quantify the "clumpiness" of errors; does task

structure affect how clumpy errors are? More details on both can be found in the Methods/Hypotheses section of this chapter (p.147).

5.2   METHOD

The following experiments are based on two tasks: a typical GN task and a typical PM task, where by "typical" I mean that the task has been reported in various publications, and is reasonably representative. I will first describe these tasks before proceeding to the full experimental design.

5.2.1   THE PM TASK: THE WORD CATEGORIZATION TASK (WCT)

The PM task was a variant of the one used by Einstein et al. (2005) and others (for example Scullin et al., 2010). The original code script and stimuli of this task were kindly sourced directly from the authors. I modified this task in various ways, although all changes were minor and described in this section.
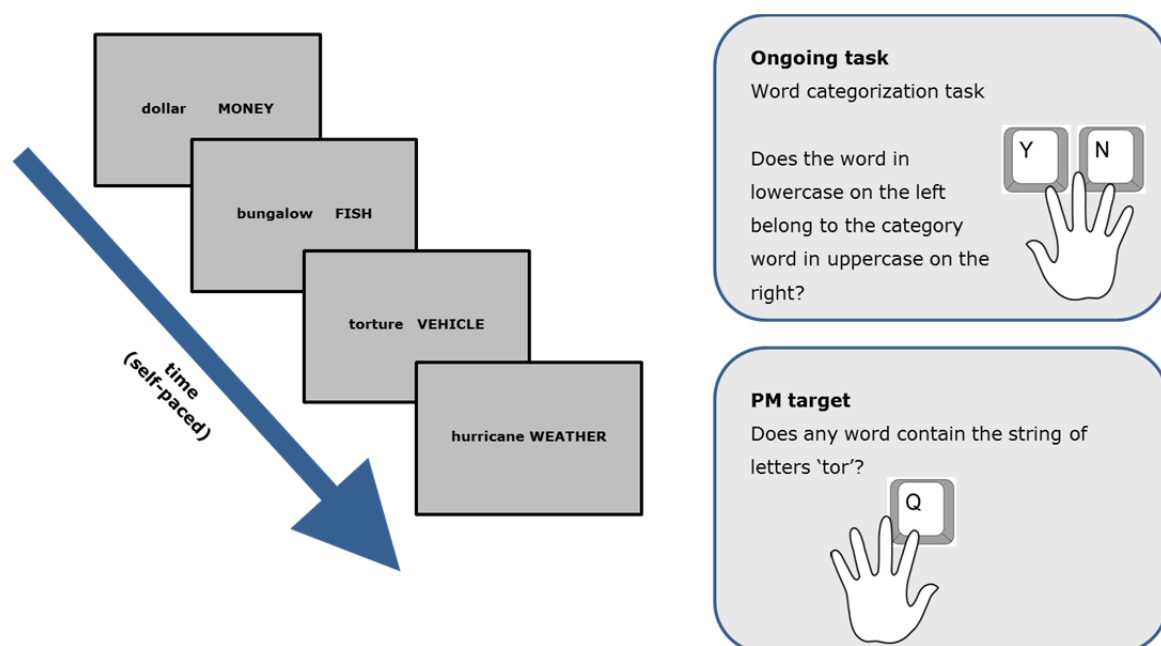
Figure 5.1 The Word Categorization Task (WCT)

The WCT required judging whether words on the left belonged to category words on the right by pressing Y or N keys. The PM target was the string of letters "tor". If participants saw this string of letters anywhere they had to press the Q key.

The task was made up of two sub-tasks, the ongoing task and the PM instruction (Figure 5.1). Participants read all task instructions directly from the screen (see Appendix for verbatim instructions). The ongoing task, a word categorization task, was introduced first. In this task, pairs of words were presented on the screen and participants had to judge whether the word in lowercase letters on the left, belonged to the category word in uppercase letters on the right. A standard QWERTY keyboard with a number pad on the right was used to make responses. The "Q" key, "5" key and "6" key (both number keys found on the number pad) were labelled using white stickers as "Q", "Y" and "N", respectively. Participants pressed either the Y key or the N key to indicate "yes" or "no" as their response. The task was self-paced such that the word pairs remained on screen until a Y or N response was made; this response triggered the next trial. Participants were instructed to respond as quickly and as accurately as possible.

WCT stages	
Instructions: Rule A	
Practice 1a: questions allowed/accuracy and speed feedback	13 trials
Practice 1b: no feedback	30 trials
Instructions: Rules A + B	
Recall instructions	√
Practice 2: all rules, no feedback	11 trials
Remind instructions	√
Main block	
non-PM trials	160 trials
PM trials	4 trials
Recall instructions	√
Final 'anti-deception' trials	√
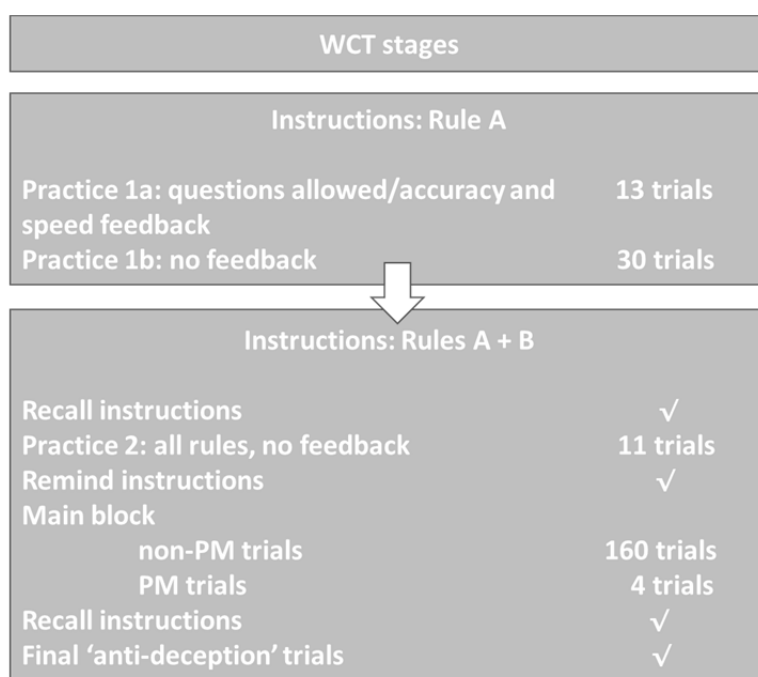
Figure 5.2 Task stages in the Word Categorization Task (WCT)

Rule A=ongoing task
Rule B=prospective memory instruction
The ongoing word categorization task was introduced first and was followed by extended practice (practice 1a and 1b). The prospective memory instruction was introduced later with some practice and followed by the main block of trials, in which only 4 trials had the PM target. The "anti-deception" trials were not used in the analyses and their function is described later in this chapter (p.137).

There were an initial 3 practice trials in which participants had the opportunity to ask questions, followed by 10 practice trials with speed and accuracy feedback per trial to encourage quick and accurate responses (Figure 5.2). This was followed by a block of 30 trials which also functioned as additional practice, however without any feedback. Next, participants were told that there was an additional task (PM instruction) which was of secondary interest. They were instructed to press the Q key whenever they saw the string of letters "tor" (PM target) in this order in any word. They were told that if they forgot to press the Q key immediately, they could do so whenever they remembered. They were further instructed that their primary goal throughout the task was to perform the word categorization task as quickly and as accurately as possible. This was followed by (1) recall of instructions, (2) a practice block of 11 trials which included one instance of the PM target, (3) reminders to the instructions including a final opportunity to ask questions and (4) the main block of trials. The main block consisted of 164 trials with the PM target appearing 4 times. At the end, participants were asked to recall instructions and went on to a few additional final trials. These final trials were referred to as the "anti-deception" trials and were not analysed. Their primary purpose was to avoid deceiving participants in conditions of Instructional Complexity (IC) which will become clearer later on in this chapter.

The word pairs for the WCT were created with the Battig and Montague (1969) norms in line with the original task by Einstein et al (2005). In total, 248 word pairs were created from 31 categories. For each category, there were 8 word pairs. Half of these were a match (they required a "yes" response because they were an exemplar of that category) and the other half were a non-match (they required a "no" response because they did not belong to the category word). The original WCT task (as received from the authors) was set in American English; consequently some words were replaced because they were unsuitable for British English speakers. Given that one of my secondary interests was to assess individual differences in PM performance, the order of the word pairs was fixed throughout the task, unlike in the original task in which the order was pseudo-randomized to remove item-effects. No word appeared twice in the task which was not strictly controlled in the original task. Four of the 164 main block word pairs included the PM target "tor". The PM targets occurred once in the words *dormitory*, *tornado*, *history* and

tortoise, in this fixed order and appeared at the following fixed positions; 40th, 80th, 120th and 160th trial.

5.2.2 THE GN TASK: THE LETTER MONITORING TASK (LMT) VERSION 2

The GN task was based on the LMT as described in Chapter 2 (p.65), with five main changes. These modifications were needed to accommodate the overall experimental design. For the reader's convenience, below I reproduce the illustration of the trial structure of this first version of the LMT.
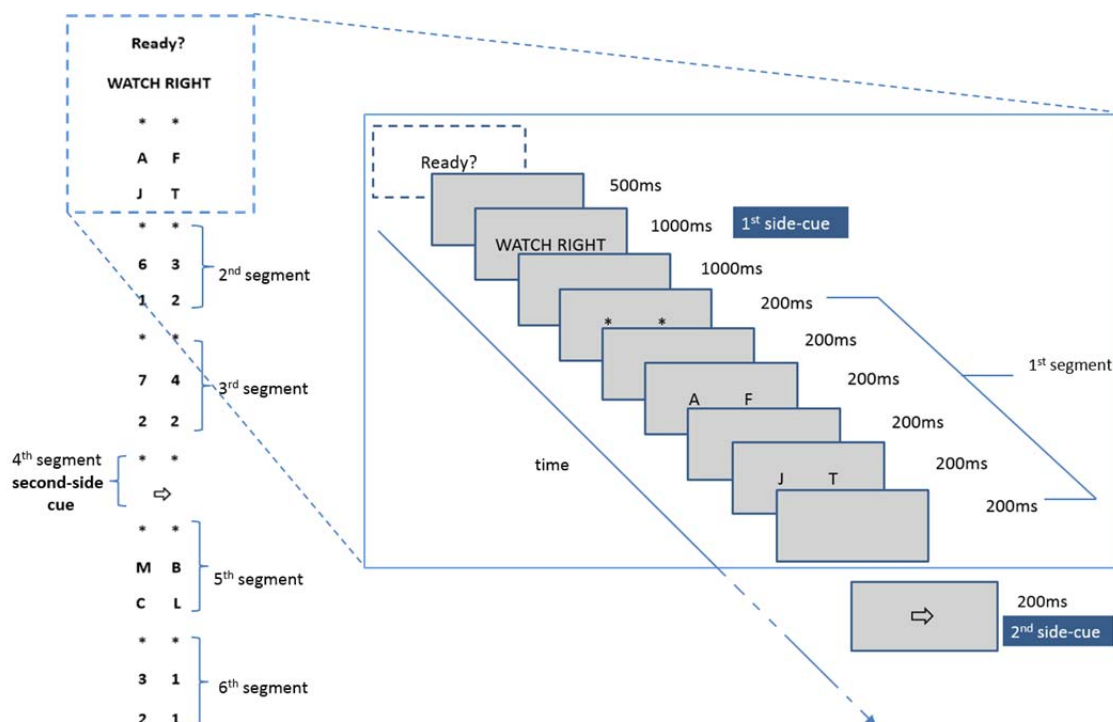
Figure 5.3 A trial from the LMT, version 1 as copied from Figure 2.15

On the left, is a sample of the sequence of stimuli in one trial out of a total of 16, in the LMT version 1. A trial starts with the first side-cue which contains a verbal instruction and is followed by a sequence of frames organized in 6 segments. On the right are details of the durations of the displays. A letter or number segment is made up 6 frames: an asterisk display and two letter (or two number) displays each preceded by a blank interval. Participants had to attend to the side they were instructed to by the side-cues, and on that side read letters aloud or say the sum of numbers within that segment. For example, in this trial the correct response is "F,T, 5, 6, B, L, 2".

What follows is a description of the five modifications to the LMT giving rise to the LMTv2 which is illustrated in Figure 5.4. First, there were no numbers, just letters for this

version of the task, because I wanted just one sub-task as the regular task, to match the PM task. Second, the pair of asterisks which appeared in the first frame of a segment were removed and replaced with a single central symbol from the following set: £, %, @, &. This served to make the second side-cue less salient. The pair of asterisks preceding the second side-cue were not replaced, but instead completely removed. Third, the set of symbols used for the second side-cue were changed to either a plus symbol "+" or minus "-" symbol representing right and left respectively (as found in Duncan et al., 1996; Duncan et al., 2008). Thus there was no longer a manipulation of transparency of the second side-cue. The fourth and fifth changes regard timings: the display duration changed from 200ms to 300ms, and the number of trials increased from 16 to 40 trials.

Ready?

WATCH RIGHT

£

A   F   — 1st segment

J   T

%

R   Z   — 2nd segment

K   E

@

M   B   3rd segment

C   L

4th segment
second-side   +
cue

P   Y   5th segment

X   N

&

Q   S   6th segment

G   H
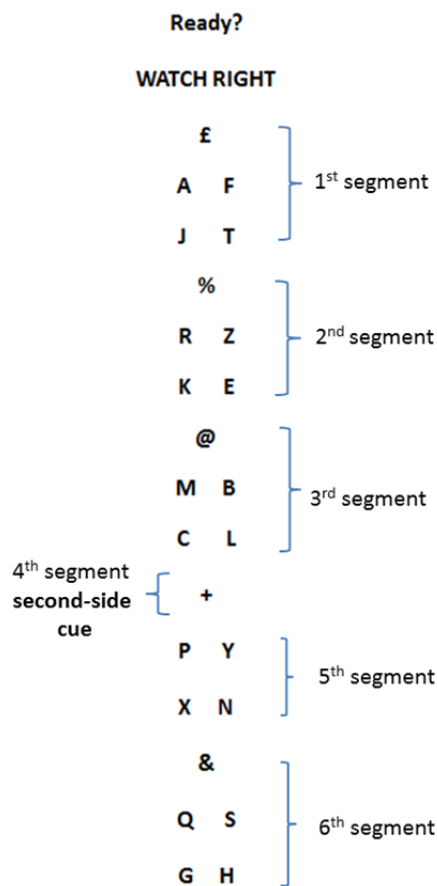
Figure 5.4 A trial from the LMTv2

The second version of the LMT, the LMTv2, was produced by applying five main modifications to the original LMT: there were no number segments and instead only letter segments, pairs of asterisks were replaced with single symbols, the duration of each frame changed from 200ms to 300ms, the number of trials increased from 16 to 40 and the symbols for the second side-cue changed to a + and -, for right and left, respectively.

5.2.3 EXPERIMENTAL DESIGN

Now that I have described the basic tasks in this set of experiments, I can proceed with describing the overall experimental design for the task structure investigation. This was a 2x2x2 factorial design with Structure (GN-like, PM-like), Task (LMTv2, WCT), and IC (High, Low) which resulted in 4 experiments and a total of 8 experimental conditions (see Table 5.1). The Tasks (LMTv2, WCT) were described in the previous section and their form (morphed or original) depended on the interaction between the Task and Structure factors. The Structure factor had two levels; GN-like or PM-like (see Figure 5.5). Each level was characterised by two parameters: (1) the frequency of the critical event and (2) the amount of practice. First, I established what these settings were in the original tasks. Second, I created morphs by swapping these settings across tasks. The result was (i) a morphed LMTv2 with extended practice and minimal critical trials and (ii) a morphed WCT with minimal practice and more frequent PM targets (Figure 5.6). The following sub-sections describe in detail the way the parameters were calculated and the morphs produced.

Table 5.1 Experimental design

Structure factor	PM-like				GN-like			
Frequency of critical events/PM targets in the main block of trials	≈2.4%				≈16.6%			
Practice	Extended				Minimal			
Task factor	WCT original WCT		LMTv2 morphed LMTv2		WCT morphed WCT		LMTv2 original LMTv2	
IC factor	High	Low	High	Low	High	Low	High	Low
Experiment #	Experiment 5		Experiment 6		Experiment 7		Experiment 8	
N	25	25	16	16	31	31	16	16

WCT=Word Categorization task
LMTv2=Letter Monitoring Task, version 2
IC=Instructional Complexity
This table summarises the 2x2x2 factorial design of Structure (PM-like, GN-like), Task (WCT, LMTv2) and IC (High, Low) which resulted in Experiments 5-8. Morphed vs. original versions of the tasks depended on the interaction between the Structure and Task factors. For example, the LMTv2 in the GN-like structure is the original LMTv2 task, whilst the morphed LMTv2 uses the LMTv2 in a PM-like structure.

5.2.4 FREQUENCY OF CRITICAL EVENTS IN THE ORIGINAL TASKS

The frequency of critical events was calculated by dividing the number of critical events by the total number of events in the main block of trials (i.e. excluding any practice). To make the measurements comparable across the LMTv2 and the WCT, I defined an "event" as 1 trial in the WCT and as 1 segment in the LMTv2. In the LMTv2, there were 40 trials each made up of 6 segments giving a total of 240 segments, or 240 events. Forty out of these 240 events contained the critical event (the second side-cue). Hence, the rate of critical events in a GN-like task was set to approximately 16.6% of the main block. In the WCT task, there were a total of 164 trials in the main block and 4 of these contained

the PM cue. Therefore, the frequency of critical events in a PM-like task was set to around 2.4%.

5.2.5 FREQUENCY OF CRITICAL EVENTS IN THE MORPHED TASKS

In the morphed LMTv2, the frequency of critical events was reduced to 8 out of 40 trials, equivalent to 8 occurrences of the second side-cue out of 240 events ($\approx$ 3.3%). These critical trials occurred at fixed positions, on every 4th trial, which alternated between switch and non-switch trials. Considering that only half were switch trials, this implied that it was unlikely that more than 4 trials would produce a GN error. For the remaining 32 trials, a symbol picked at random (but in a fixed order) from the set: £, %, @, &, replaced the second side-cue in the 4th segment. In the morphed WCT, PM target trials increased to 20 out of 164 trials which is approximately 12.2%.
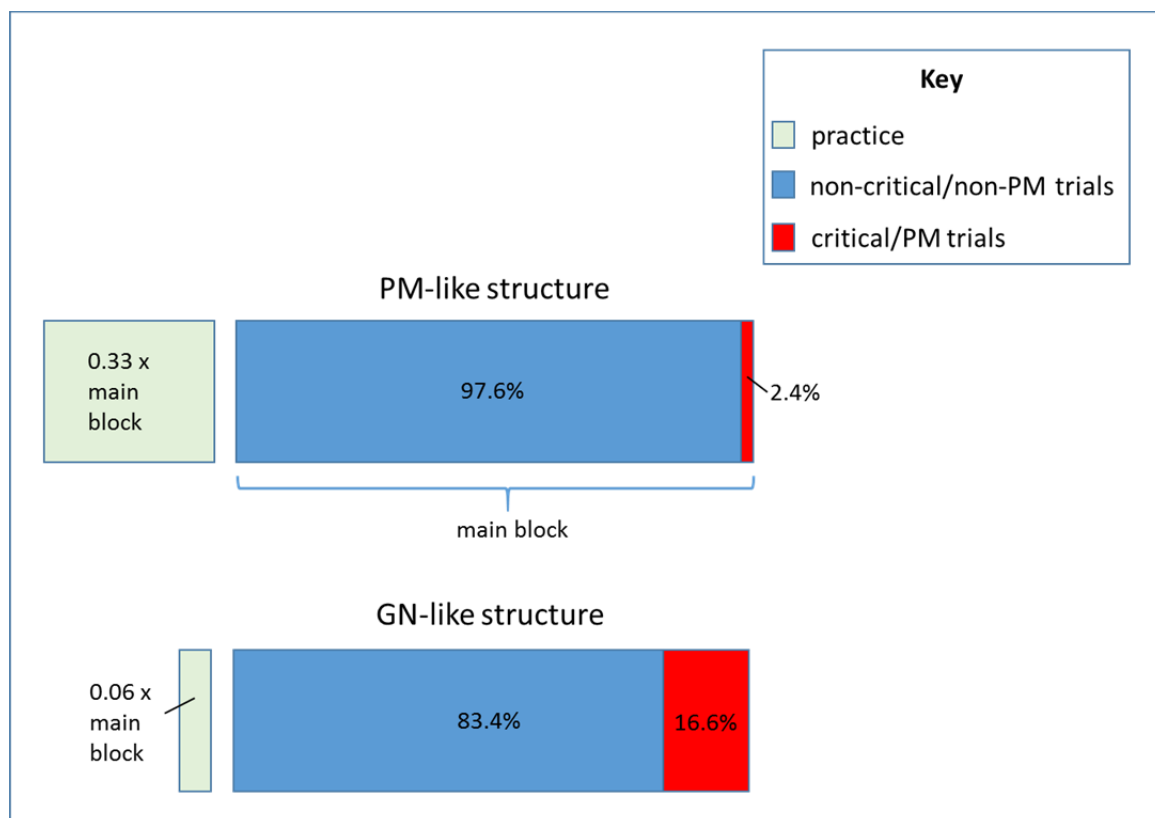
Figure 5.5 Structure Factor

A PM-like structure as modelled on the WCT, had extended practice (around 0.33 times the number of trials in the main block) and minimal PM targets ($\approx$2.4%). Instead, a GN-like structure, as modelled on the LMTv2, had minimal practice (around 0.06 times the number of trials in the main task) and relatively frequent critical events ($\approx$16.6%).

5.2.6 PRACTICE IN THE ORIGINAL TASKS

Before I proceed with describing the nature of task practice across original and morphed tasks, it is worth noting that there were a total of three rules: Rule A, Rule B and Rule C. Rule A referred to the regular/ongoing task, which in the WCT was the word categorization, and in the LMTv2 was reading letters aloud from the side indicated by the first side-cue. Rule B referred to the critical rule/PM instruction which in the WCT required detecting the string of letters "tor" in any word, and in the LMTv2 instructed participants to attend the appropriate side, right or left, depending on the second side-cue. Rule C in the WCT required participants to read aloud any words that were both in green ink and in lowercase letters. In the LMTv2, Rule C required adding pairs of numbers together and saying the result aloud (as in the LMTv1 described in Chapter 2). Rule C was not always instructed to participants, which depended on the level of the Instructional Complexity factor (IC), which was either High (instructed) or Low (not instructed), in line with previous descriptions found in Chapter 4 (p.110). Irrespective of the level of IC, Rule C was never actually required in any of the main blocks across all Experiments 5-8. In conditions of High IC, where participants were given the extra instruction (Rule C), they were also told to temporarily ignore this extra rule just before the start of the main block of trials.

Table 5.2 Rules A, B and C, a description.

Rule	Description of Rule
A	The rule about the regular/ongoing task. For the WCT this is the word categorization task. For the LMTv2, this is reading the letters aloud from the side indicated by the first side-cue.
B	The critical rule (in GN terms) or the PM instruction (in PM terms). In the WCT, this requires detecting the string of letters "tor" in any word. In the LMTv2, this involves attending the correct side, left or right, depending on the nature of the second side-cue.
C	The extra rule. In the WCT, this instructs participants to read aloud words that are presented in lowercase letters and green ink. In the LMTv2, this rule requires adding pairs of numbers together and saying the result aloud.

In the LMTv2, practice incorporated all rules at once. Typically it consisted of only one trial, with the exception of a rare few participants who did not respond at all. In the latter cases a few more practice trials were run (usually not more than 1 or 2 more trials), until a response was uttered, ignoring whether this response was correct or not. Instead, practice in the WCT was relatively extended and had two general stages. The first stage involved instructions and considerable practice on only Rule A. The first stage also provided feedback on accuracy and speed, per trial, and allowed participants to ask questions. The second stage introduced Rule B (and if applicable also Rule C) followed by brief practice which included instances of Rule B and the PM target (and if applicable also Rule C).

5.2.7   PRACTICE IN THE MORPHED TASKS

In the morphed LMTv2, Rule A was introduced first followed by 8 practice trials, which included the possibility to ask questions as well as feedback on accuracy. After this practice, Rule B (and Rule C if applicable), were introduced followed by recall of instructions and 4 practice trials, one of which contained Rule B, the critical event. The original WCT had a total of 54 practice trials; in the morphed WCT this was reduced to 6 trials in which all rules were active, the PM target appeared once and no feedback was provided.

Task stages	PM-like		GN-like	
	WCT original	LMTv2 morphed	LMTv2 original	WCT morphed
Instructions: Rule A	trials	segments	segments	trials
1. Practice 1a: questions allowed/ speed and/or accuracy feedback	13	12	X	X
2. Practice 1b: no feedback	30	48	X	X
Instructions: Rules A + B + C				
3. Recall instructions	√	√	√	√
4. Practice 2: all rules, no feedback	11	24	6*	6
5. Remind instructions **and abandon rule C**	√	√	√	√
6. Main block				
non-critical/non-PM events	160	232	200	144
critical/PM events (switch trials)	4	8(4)	40 (20)	20
7. Recall instructions	√	√	√	√
8. Final 'anti-deception' trials (Rules A, B & C)	√	√	√	√

Figure 5.6 Stages in the original vs. morphed PM-like and GN-like tasks

Morphed tasks were obtained by manipulating: i) the extent of practice, and ii) the frequency of critical/PM trials, based on the settings in the original tasks. Calculations were based on number of trials in the WCT and the number of segments in the LMTv2, where 1 trial in the LMTv2 was made up of 6 segments. The morphed LMTv2 was modelled on the structure of the original WCT, with extended practice and minimal critical events. The morphed WCT was based on the structure of the original LMTv2, with minimal practice and a higher frequency of PM targets. When conditions of IC were applied, the condition of High IC involved introducing and then deactivating Rule C (in orange). In the WCT, Rule A was the ongoing word categorization task; Rule B the PM instruction, involved detecting the string of letters "tor"; Rule C required to read aloud words that were written in green ink and lowercase letters. In the LMTv2, Rule A was the regular task which required reading letters aloud from the side instructed at the start of each trial; Rule B referred to the critical event which entailed attending right or left depending on the second side-cue; Rule C required participants to add pairs of numbers together and say the result aloud.
*In the original LMTv2 practice was typically 1 trial long (6 segments/events); however, on rare occasions participants did not give any response in this practice trial and hence further trials were administered until a response was uttered.

5.2.8 PRE-PROCESSING OF DATA

GN was measured using the MSE score as described in Chapter 2 (p.67). PMf were measured as a proportion of failures out of the total number of PM targets. No reaction time data was available for the LMTv2 and although reaction times were recorded for the WCT these data were not analysed.

Output from the WCT required some pre-processing to address cases in which there were double presses and skipped trials. In cases of double presses which involved a word

categorization response ("yes"/"no") and a PM target response ("q"), the PM target response overruled. In cases of double presses involving two or more "yes"/"no" responses, the first response was kept and any other discarded. Very occasionally (5 times in the whole sample) the stimulus presentation software skipped trials and in such cases, the trials were ignored.

Following the protocol used by the authors of the original WCT (Einstein et al., 2005), a window of up to 2 trials after the PM target trial was used in which pressing a "q" was counted as successful prospective remembering. Also, the same authors ignore the first 4 ongoing task trials after the PM target to avoid switch-cost related interference effects. However, given the limited amount of ongoing trials in the morphed version of the WCT, I reduced this threshold to 1 trial.

5.2.9 PARTICIPANTS

Experiment 5 was run in the same session as Experiment 3 (SGNT, Chapter 4) and Experiment 2 (LMT, Chapter 2). Experiments 6-8 were run in the same session with Experiment 4 (SGNT, Chapter 4). Across these two sessions, the order of the tasks was set to SGNT first, WCT second and LMT third and if required, the Culture Fair test Scale 2 Form A occurring at the very end. To avoid confounds due to practice effects, participants never performed the SGNT, WCT or LMT in any version or condition more than once. In addition, to reduce any potential cross-task effects it was ensured that no participant took part in more than one task in conditions of High IC – where the initial instruction was deactivated. A total of 50 participants were tested on Experiment 5 split across conditions of IC (see Table 5.1). For Experiments 6 and Experiment 8, 64 participants were tested, 16 in each condition. All of these 64 participants were also tested on the morphed WCT for Experiment 7, but 2 participants were removed from the latter analysis, one because he/she had previously taken part in a pilot version of the WCT, and another one (the last participant tested) from the other IC condition to keep the groups equal in size (n=31 in each condition of IC). Otherwise, selection criteria were the same as for Experiment 1 (p.45). Distribution of Age and Fluid Intelligence is reported in Table 5.3.

5.2.10 HYPOTHESES, INDIVIDUAL DIFFERENCES AND CLUMPINESS

As previously mentioned, Experiment 5 (original WCT) was run in an earlier session than Experiments 6-8. At this initial stage, my research question was, "is a PM task also sensitive to the IC effect?" Given the apparent similarity between PM and GN tasks, my hypothesis was that I would detect an IC effect in the WCT. However, as will be described in more detail in the results section, the original WCT gave a null result such that PMf in conditions of High and Low IC were not significantly different from each other. My subsequent hypotheses were that if the structure of the WCT was made more GN-like then, PMf in this morphed WCT would become sensitive to the IC effect. Conversely, if the LMTv2 was made more PM-like (morphed LMTv2) this would lose sensitivity to the IC effect.

In terms of individual differences, I expected to replicate the finding that GN was negatively associated to fluid intelligence scores as measured by the Culture Fair test, Scale 2 Form A (Cattell and Cattell, 1973). Fluid intelligence is not a typical covariate in PM studies, with a few exceptions such as the study by Salthouse (2004) who reports a moderate-strong positive association ($r \approx .69$) between the construct of PM ability as measured by 4 different PM tasks, and fluid intelligence as measured by 6 different fluid intelligence tests. However, age was not partialled out in this study and hence it is unclear if the known relationship between age and fluid intelligence, as also reported by the same author in a later study (Salthouse et al., 2008), could explain this association. In sum, I did not have a strong prediction about the relationship between PMf and fluid intelligence. In contrast, the effects of age on prospective remembering have been studied extensively and indicate a positive association between lab-based PMf and age (see meta-analyses by Henry et al., 2004; and Kliegel et al., 2008). However, the reverse association has also been reported for naturalistic PM tasks (see Kliegel et al., 2016 for a recent review on the PM-age paradox). Given these published reports and the positive associations between age and GN even after controlling for fluid intelligence reported in my own GN experiments in earlier chapters (Chapter 2 and Chapter 4), I expected to find positive associations between both age and PMf, and age and GN.

With respect to the clumpiness of errors using the entropy-like measure that I proposed in Chapter 3, my questions were: how do factors of Structure, Task and IC affect the clumpiness of errors? Do fluid intelligence and/or age predict clumpiness? Do I replicate

the finding that errors are more clumpy if they mostly occur early on in the task? In Chapter 3, I argued that two candidate mechanisms that may underpin errors of GN and PMf are the *task model* and *monitoring* accounts respectively. In line with this argument, here I hypothesize that a PM-like structure is dominated by *monitoring-like* mechanisms and hence would show decreased clumpiness compared to a GN-like structure in which I would expect the majority of errors to be driven by a faulty *task model*.

To answer these questions and test this hypothesis, I extracted entropy-like measures for each sequence of critical (switch) trials or PM target trials, which were 20 and 4 trials long for the GN-like and PM-like structures, respectively. I also calculated the median position of GN/PMf in each sequence, which I then converted to a proportional score given that the number of neglectable trials differed across experiments. The sequences were binary with "0" indicating correct and "1" incorrect. For the LMTv2, values of "0.5", which indicated equal proportion of letters from the correct and incorrect sides of the stream of characters, were recoded to "0" to accommodate for the entropy-like measure calculation which required a binary sequence. Scores of "0.5" occurred in 5% of the LMTv2 trials across the whole sample (n=64) and were primarily confined to 7 participants. The entropy-like measure was calculated according to the formula provided in Chapter 3, p.87. Clumpiness could not be calculated for participants who showed either no errors or no corrects; therefore these were removed from the analysis.

5.3   RESULTS

Considering that the experimental design was factorial, the analysis required a factorial ANOVA. However, the data were mostly non-normally distributed and hence I first present results using non-parametric pairwise tests within the Task factor.

All participants correctly recalled instructions at the end of all experiments. For the LMTv2, initial recall of instructions was repeated once except for a few participants in the original (n=3) and morphed (n=2) versions in which it was repeated twice. For the WCT, initial recall was correct and repeated once for all participants in both versions.

5.3.1   WCT

Experiments 5 and 7 involved the WCT in the original and morphed formats, respectively. As previously mentioned, Experiment 5 was run at a separate and earlier

time point than Experiments 6-8; however I will present all WCT results together in this section. Table 5.3 shows fluid intelligence and age data summaries on both morphed and original formats of the WCT. This table also displays results of pairwise comparisons within experiments revealing that participants across conditions of IC were matched on these variables. The subsequent table (Table 5.4) shows that participants generally performed well on the ongoing word categorization task, and there were no differences across conditions of IC in both original and morphed versions.

Table 5.3 Fluid intelligence and age in the original and morphed WCT

		IC condition	N	Mean(SD)	Median	W	r	p
Experiment 5 original WCT	Fluid Intelligence	*Low*	25	103.4(16.4)	99.0	286.0	-.07	.61 n.s.
		High	25	104.0(13.7)	102.0			
	Age	*Low*	25	61.7(5.8)	59.9	348.5	-.10	.49 n.s.
		High	25	60.9(5.6)	62.4			
Experiment 7 morphed WCT	Fluid Intelligence	*Low*	31	107.8(15.3)	105.0	510.5	.10	.68 n.s.
		High	31	106.4(13.8)	105.0			
	Age	*Low*	31	59.2(7.9)	58.9	432.0	-.13	.50 n.s.
		High	31	60.6(8.3)	64.4			

In Experiment 7, the data were found to be non-normally distributed using a Shapiro-Wilk test (p < .01) and hence, non-parametric Mann-Whitney-Wilcoxon tests were used instead of *t*-tests. For the sake of consistency, non-parametric tests were also run on the same variables in Experiment 5 even though these were normally distributed. All tests revealed that the groups were matched on age and fluid intelligence.

Table 5.4 Proportion correct on the ongoing task in the WCT

	IC condition	N	Mean(SD)	Median	*W*	*r*	*p*
original WCT (Exp. 5)	*Low*	25	.97(.03)	.98	352.0	.08	.45 n.s.
	High	25	.97(.02)	.97			
morphed WCT (Exp. 7)	*Low*	31	.97(.02)	.97	406.5	-.13	.30 n.s.
	High	31	.97(.02)	.98			

Shapiro-Wilk tests revealed non-normal distributions for the ongoing task data for original WCT but not for the morphed WCT. To be consistent, non-parametric tests (Mann-Whitney-Wilcoxon) were used on both sets of data which showed that conditions of IC did not affect performance on the ongoing task.

Table 5.5 Proportion of PMf in the WCT

	IC condition	N	Mean(SD)	Median	*W*	*r*	*p*
original WCT (Exp. 5)	*Low*	25	.50(.34)	.50	324.0	.04	.83 n.s.
	High	25	.47(.36)	.50			
morphed WCT (Exp. 7)	*Low*	31	.27(.23)	.20	410.5	-.17	.33 n.s.
	High	31	.32(.25)	.25			

Distributions of PMf in both the original and morphed versions of the WCT revealed non-normality according to Shapiro-Wilk tests ($p<.05$), and hence Mann-Whitney-Wilcoxon tests were used for pairwise comparisons.
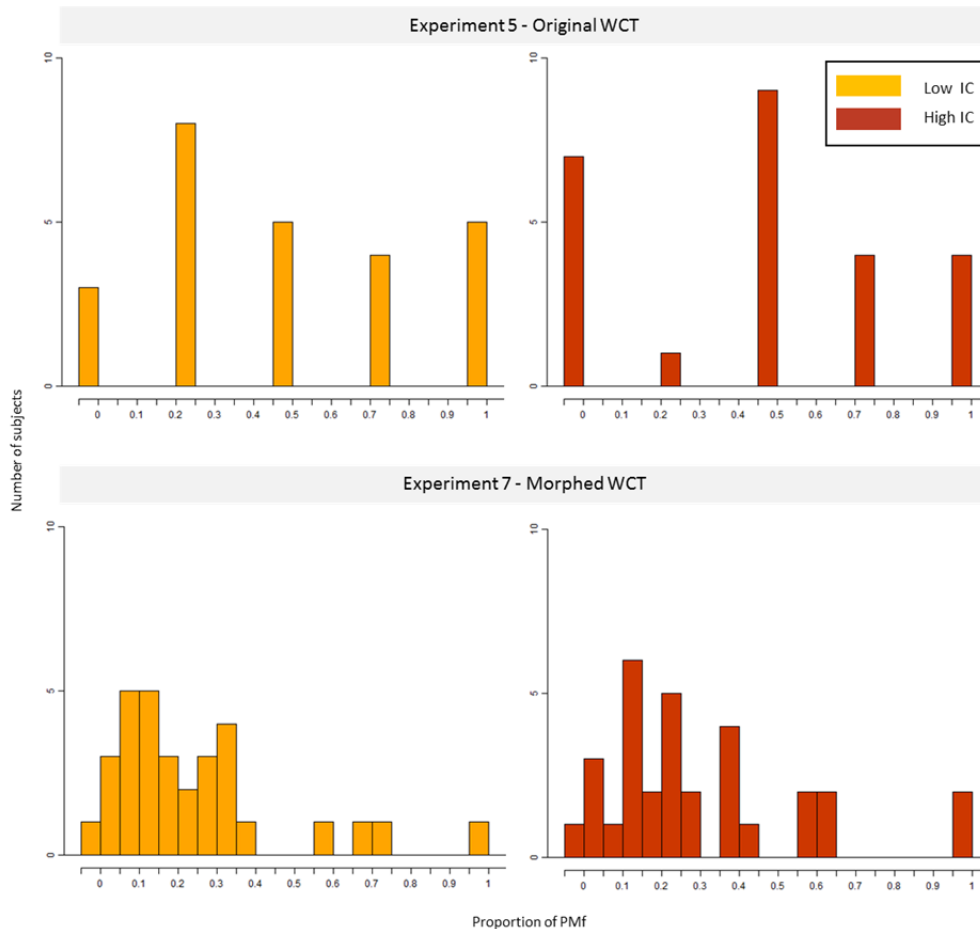
Figure 5.7 Distributions of proportion of PMf in the original and morphed WCT

The upper row shows distributions of PMf in the original WCT (extended practice, 4 PM targets) across conditions of IC (Low IC in yellow, High IC in red). The lower row displays the PMf distributions for the morphed WCT (minimal practice, 2 PM targets). Although in the morphed WCT the High IC condition showed a nominally larger amount of PMf than in the Low IC, this was not statistically significant. Statistics are reported in Table 5.5.

Contrary to my original hypothesis, PMf was not found to be sensitive to conditions of IC in the original version of the WCT (p=.83). For precise statistics and distribution data, please refer to the top sections of Table 5.5 and Figure 5.7. My subsequent prediction was that if I made the WCT more GN-like by increasing the number of PM targets and minimizing practice, then this would make any resulting PMf sensitive to the IC effect. Although the morphed WCT in High IC conditions did show a nominal increase in PMf compared to Low IC (see lower sections of both Figure 5.7 and Table 5.5), this was not statistically significant. In a similar way, although effect sizes across formats of the WCT (r=.04 vs. r=-.17) seemed to suggest that morphing the WCT introduced an IC effect,

Fisher's *r*-to-*z* transformations (Preacher, 2002) showed that this was not a significant difference (*p*=.23).

5.3.2   LMTv2

Experiments 6 and 8 referred to the LMTv2 in the original and morphed formats, respectively. Similarly to the WCT, fluid intelligence and age were matched across conditions of IC within each experiment (see Table 5.6). Performance on the regular (letter reading) task in the LMTv2 was generally high, and tests did not reveal any differences across conditions of IC in both original and morphed versions (see Table 5.7).

In the original LMTv2, GN as measured by the MSE was found to be significantly higher in conditions of High IC compared to conditions of Low IC (see top sections of both Table 5.8 and Figure 5.8). This implies that even though participants executed the exact same task, the additional instruction administered in the High IC group significantly impaired their performance on critical events. This result replicates previously published GN findings (Duncan et al., 2008) and the ones reported for the SGNT in both versions of ATC and ATCv2, in Chapter 4. In terms of task structure, I predicted that making the LMTv2 more PM-like by increasing practice trials and decreasing the frequency of critical events would remove the IC effect. This prediction was based on the finding from Experiment 5 in which IC did not affect PMf in the original format of the PM task. Congruent with this hypothesis, in the morphed LMTv2 no difference was found in GN across conditions of IC (*p*=.48), for details see lower parts of Table 5.8 and Figure 5.8. Despite these findings, using *r*-to-*z* transformations (Preacher, 2002), the IC effect size in the original LMTv2 (*r*=-.42) was not found to be significantly larger (*p*=.20) than that in the morphed version (*r*=-.12).

Table 5.6 Fluid intelligence and age in the original and morphed LMTv2

		IC condition	N	Mean(SD)	Median	*t(df)*	*r*	*p*
Experiment 6 Original LMTv2	Fluid Intelligence	Low	16	108.2(15.1)	105.0			
		High	16	102.8(9.7)	103.5	1.21(30)	.22	.24 n.s.
	Age	Low	16	60.9(7.6)	63.0			
		High	16	59.4(8.3)	59.6	.54(30)	.10	.60 n.s.
Experiment 8 morphed LMTv2	Fluid Intelligence	Low	16	109.9(15.7)	110.0			
		High	16	108.9(16.3)	105.0	.19(30)	.03	.85 n.s.
	Age	Low	16	57.9(8.5)	56.3			
		High	16	59.7(7.4)	60.3	-.64(30)	.12	.52 n.s.

Four *t*-tests were used to compare distributions across conditions of IC within each experiment. These tests revealed that the groups were matched on age and fluid intelligence.

**Table 5.7 Proportion correct on the ongoing task in the LMTv2**

	IC condition	N	Mean(SD)	Median	*W*	*r*	*p*
original LMTv2 (Exp. 6)	Low	16	.98(.02)	.99			
	High	16	.95(.12)	.99	118.0	.02	.71 n.s.
morphed LMTv2 (Exp. 8)	Low	16	.98(.02)	.99			
	High	16	.98(.03)	.99	131.5	.03	.90 n.s.

Shapiro-Wilk tests revealed non-normal distributions for the ongoing task data for both formats of the LMTv2. Null results using Mann-Whitney-Wilcoxon tests suggested that conditions of IC did not affect performance on the letter task.

Table 5.8 Measures of GN in the LMTv2

	IC condition	N	Mean(SD)	Median	*W*	*r*	*p*
original LMTv2 (Exp. 6)	*Low*	16	.05(.11)	.01	66.5	-.42	.02*
	High	16	.22(.22)	.12			
morphed LMTv2 (Exp. 8)	*Low*	16	.11(.13)	.12	110.0	-.12	.48 n.s.
	High	16	.19(.22)	.12			

Given the non-normality of the MSE distributions in both the original and morphed versions of the LMTv2 (Shapiro-Wilk tests $p<.05$), two Mann-Whitney-Wilcoxon tests were used to run pairwise comparisons across conditions of IC. *Significance at the $p<.05$ level for an IC effect was found within the original format of the LMTv2, replicating previous findings. As predicted, the IC effect was not found in the morphed version of the LMTv2.
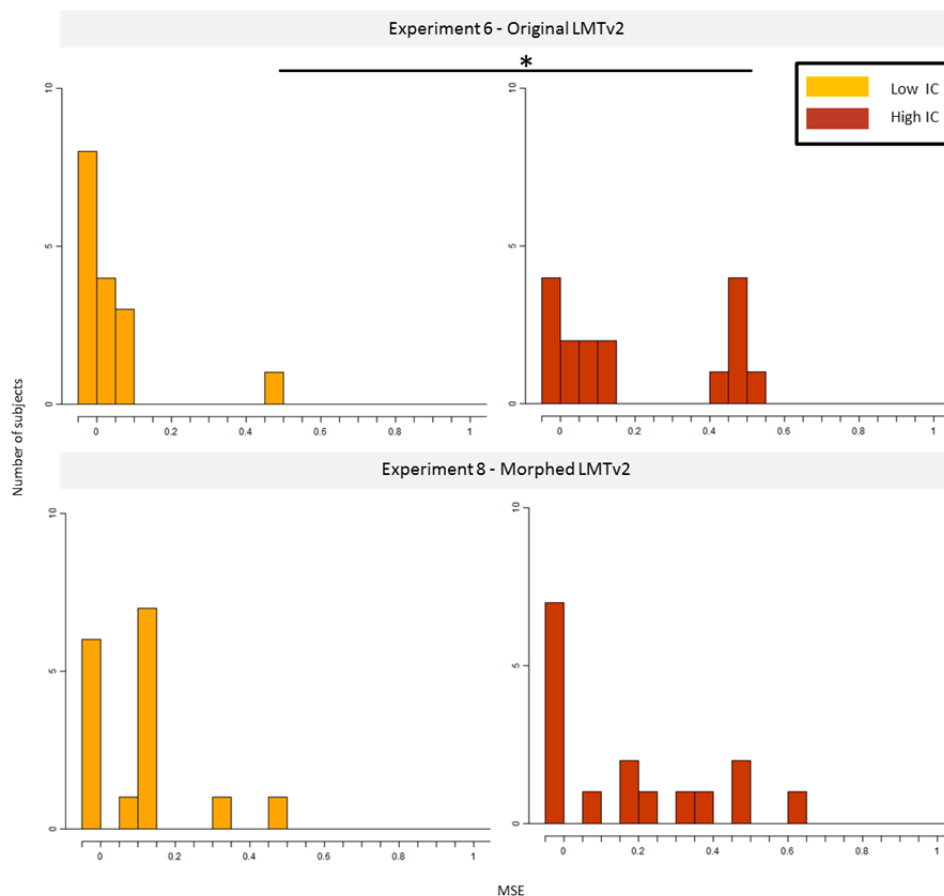
Figure 5.8 Distributions of GN in the original and morphed versions of the LMTv2

The plots show distributions of GN as measured by the MSE. The IC effect was found to be significant in the original LMTv2 (upper row), with a task structure characterised by minimal practice and a high frequency of critical events. The lower row shows the MSE distributions for the morphed LMTv2, with extended practice and low rate of critical events, and as predicted, this resulted in no detectable IC effect.

5.3.3 FACTORIAL ANCOVA AND ANOVA

The nonparametric tests revealed 2 main findings. On the one hand, congruent with my hypothesis, morphing the LMTv2 into a PM-like structure removed the IC effect. The latter finding rests on null result and hence should be read with caution. On the other hand, contrary to my prediction, morphing the WCT into a GN-like structure did not introduce an IC effect, and the interaction between task structure and IC effect was not significant. Next, I wanted to run a factorial ANCOVA to get a sense of the contributions that each of the 3 factors made to errors, including measures of individual differences as covariates. What follows are the results from a 2x2x2 ANCOVA with GN/PMf as the dependent variable; Structure (GN-like, PM-like), Task (LMTv2, WCT), and IC (High, Low) as the 3 fixed factors and age and fluid intelligence as covariates. The results should be interpreted with caution given that the data were partly non-normally distributed. To run this analysis I had to ensure that the errors (GN/PMf) on the LMTv2 and WCT were on comparable scales. Both MSE (GN) and PMf were proportional scores ranging from 0 to 1. However, typically the maximum MSE score was 0.5 rather than 1 because, as previously described in Chapter 2, half of the critical trials were non-switch trials which were unlikely to show neglect. Hence, to make the PMf comparable to the MSE, I multiplied it by 0.5. The results of the ANCOVA, including means and standard errors for all effects are presented in Table 5.9.

Table 5.9 Results of the 2x2x2 factorial ANCOVA for task structure

	$F^$	p	r	*Factor level*	*Mean(SE)*
Structure	5.16	.024*	.17	PM-like	.20(.02)
				GN-like	.14(.02)
Task	3.59	.060	.14	WCT	.20(.02)
				LMTv2	.14(.02)
IC	6.01	.015*	.19	Low	.14(.02)
				High	.20(.02)
Task x IC	5.09	.025*	.17	WCT x Low IC	.19(.02)
				LMTv2 x Low IC	.08(.03)
				WCT x High IC	.20(.02)
				LMTv2 x High IC	.20(.03)
Structure x Task	1.54	.217	.09	PM-like x WCT	.24(.02)
				PM-like x LMTv2	.15(.03)
				GN-like x WCT	.15(.02)
				GN-like x LMTv2	.14(.03)
Structure x IC	1.44	.232	.09	PM-like x Low IC	.18(.03)
				PM-like x High IC	.21(.03)
				GN-like x Low IC	.09(.02)
				GN-like x High IC	.19(.02)
Structure x Task x IC	.24	.626	.05	-	-
Age	.50	.480	.05	-	-
Fluid intelligence	9.84	.002**	.24	-	-

^ Degrees of freedom for all *F* values=9,175;
* *p*<.05, ** p<.01;
- Means for the non-significant Structure x Task x IC interaction are not presented here.

There were 2 significant main effects: structure and IC. There were more errors for the PM-like structure than the GN-like structure $F(1,175)=5.16$, $r=.17$, $p=.024$. The PM-like structure was characterized by extended practice and reduced frequency of targets. Given that it is unlikely that extended practice led to an increase in errors, by exclusion, it is likely that this was primarily driven by the low frequency of targets. The main effect of IC, suggested that conditions of High IC were generally accompanied with more errors, $F(1,175)=6.01$, $r=.19$ $p=.015$.

There was a significant interaction effect of Task and IC, $F(1,175)=5.09$, $r=.17$, $p=.025$, reflecting that the IC effect was only present in the LMTv2 and not in the WCT which is congruent with the non-parametric test results. Although the nominal results hinted at a more pronounced IC effect in the GN-like structure (mean difference=.10) compared to the PM-like structure (mean difference=.03), this was not significant $F(1,175)=1.44$, $r=.09$, $p=.232$, n.s. This latter result agrees with the findings from the non-parametric tests and suggests that contrary to my main prediction, there was no interaction between Structure and IC. There were no other significant interactions. In terms of covariates, a significant effect was found for fluid intelligence: $F(1,175)=9.84$, $r=.24$, $p=.002$ but none for age. I explore the effect of individual differences more closely in the next section.

To ensure that the covariates were not distorting the model unnecessarily, I ran the same ANOVA without fluid intelligence and age as covariates. The results are presented in the table below. The results were largely the same to the ANCOVA reported above, except for an additional main effect of Task ($p=.036$) which was trending in the ANCOVA ($p=0.060$). This main effect suggests that the type of Task (WCT or LMTv2) matters, with the WCT showing more errors compared to the LMTv2.

Table 5.10 Results of the 2x2x2 ANOVA of task structure (without covariates)

	F^	p
Structure	4.75	.031*
Task	4.46	.036*
IC	6.71	.010*
Task x IC	5.48	.020*
Structure x Task	2.75	.099
Structure x IC	1.75	.187
Structure x Task x IC	.225	.636

^ Degrees of freedom for all F values=7,175;
* p<.05

5.3.4 INDIVIDUAL DIFFERENCES

Figure 5.9 to Figure 5.12 illustrate the associations between PMf or GN, to age and fluid intelligence. Based on previous findings, my predictions were a negative association between fluid intelligence and GN, and a positive association between age and GN, hence the use of 1-tailed correlational tests. I also expected a positive association between age and PMf which justified the use of a 1-tailed test. I did not have any strong predictions about the relationship between PMf and fluid intelligence and hence the 2-tailed test.

As expected, fluid intelligence predicted the MSE score such that participants performing poorly on the Culture Fair test demonstrated more GN (ρ>-.43, p<.05). These findings were observed in both versions of the LMTv2, original and morphed, and in both conditions of IC, High and Low, except for Low IC in the original version (for plots and statistics see Figure 5.9). Partialling out age did not significantly change these results. On closer observation of these plots, it appeared that IC in the morphed LMTv2 did not change the relationship between fluid intelligence and errors (ρ=.45 vs. ρ=.55), whilst this was not the case in the original version (ρ=.23 vs. ρ=.66). To test this I ran a linear regression with GN as the dependent variable and Structure, IC, the interaction between Structure and IC, and fluid intelligence as the independent variables. This revealed that IC and fluid intelligence were significant predictors of GN, B=.141, β=.380, p=.020 (IC) and

B=-.005, β=-.386, *p*=.001 (fluid intelligence) but not Structure, nor importantly, the interaction between Structure and IC *p*=.25 and *p*=42 respectively. The lack of an interaction effect indicates that the observation that IC moderated the relationship between fluid intelligence and errors only in the original structure was not statistically significant. Instead, the finding suggests that when the LMTv2 is in the High load IC condition or participants score poorly on the Culture Fair Test A, more errors are observed.

Fluid intelligence was not found to be significantly associated to PMf in any form of the WCT or IC condition.

There was a positive association between GN and age in the low IC original version of the LMTv2 (ρ=.47, *p*=0.03), and this was at borderline significance when fluid intelligence was partialled out. I expected older participants to perform worse on the prospective component of the WCT, however, oddly, there was a negative association between age and PMf in the Low IC original WCT and this survived partialling out of fluid intelligence. Given the directionality of my hypothesis the result is not significant and is probably a fluke. I further discuss this odd result in the general discussion.
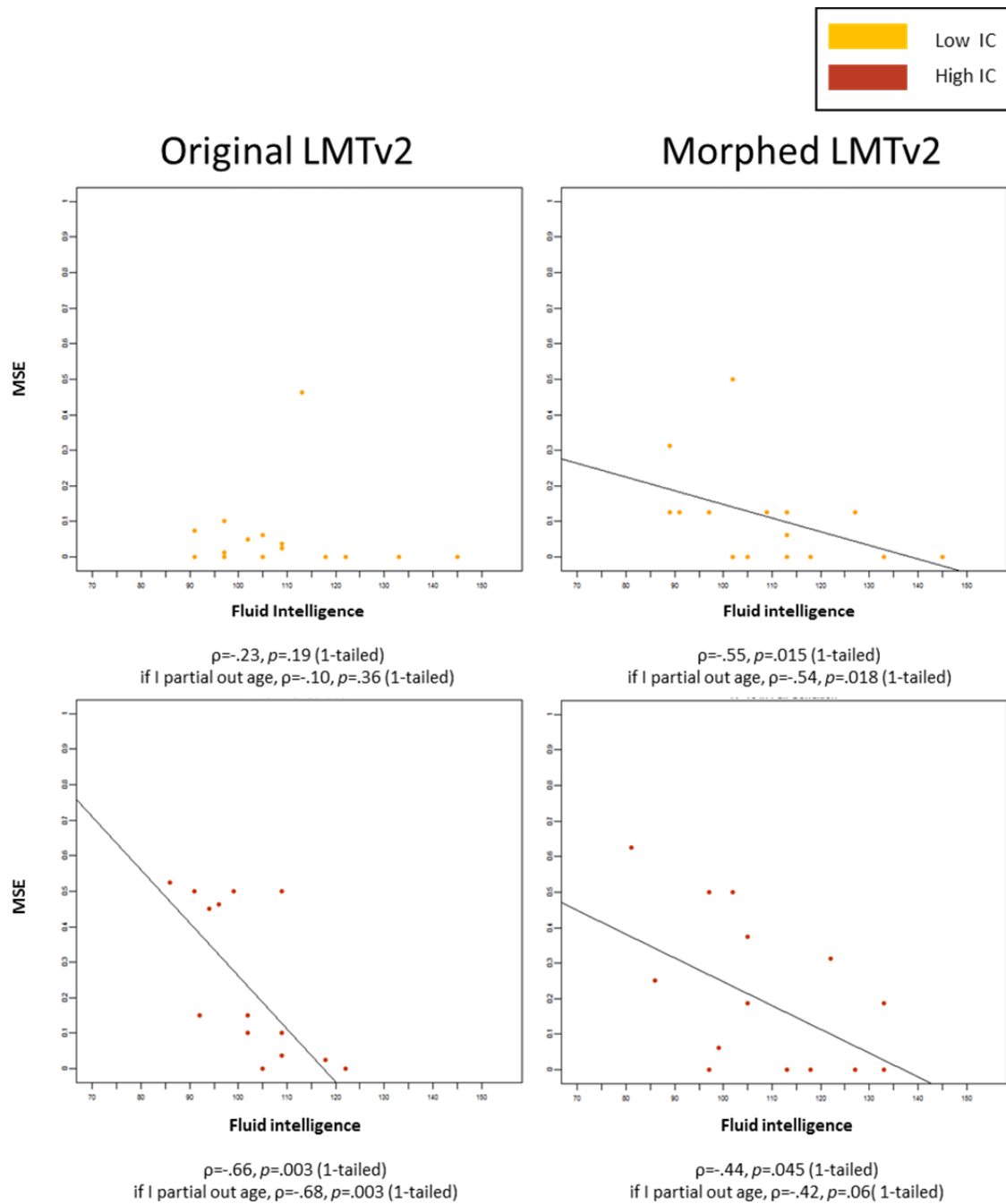
Figure 5.9 Fluid intelligence vs. GN (MSE) in the LMTv2

Associations between fluid intelligence and GN as measured by the MSE in the four combinations of the LMTv2 as a result of crossing Structure (GN-like, left column; PM-like, right-column) and IC (High, red; Low, yellow). Significant associations are marked with lines of best fit.
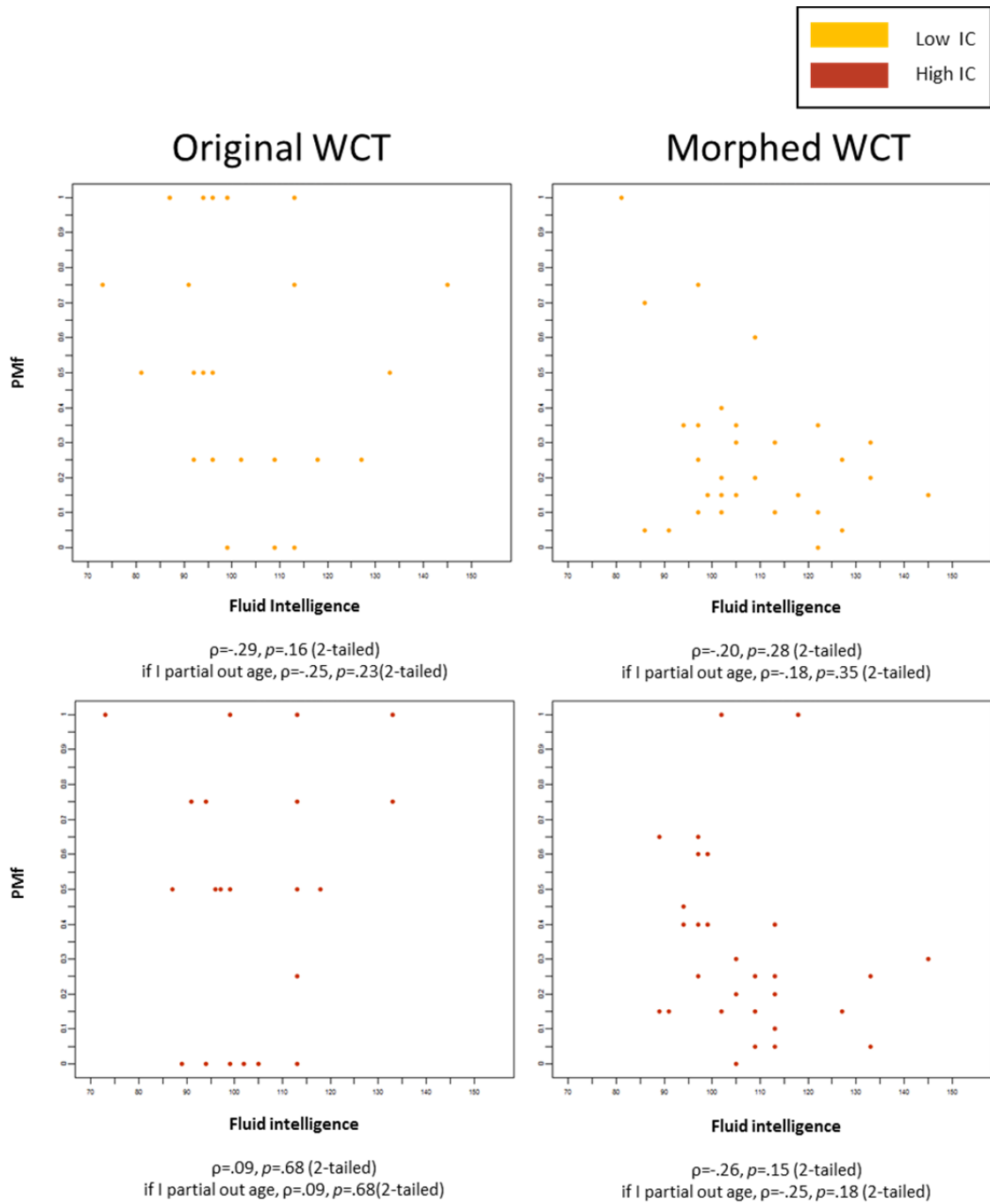
Figure 5.10 Fluid intelligence vs. PMf in the WCT

Associations between fluid intelligence and PMf in the four combinations of the LMTv2 as a result of crossing Structure (PM-like, left column; GN-like, right column) and IC (High, red; Low,yellow). There were no significant associations.
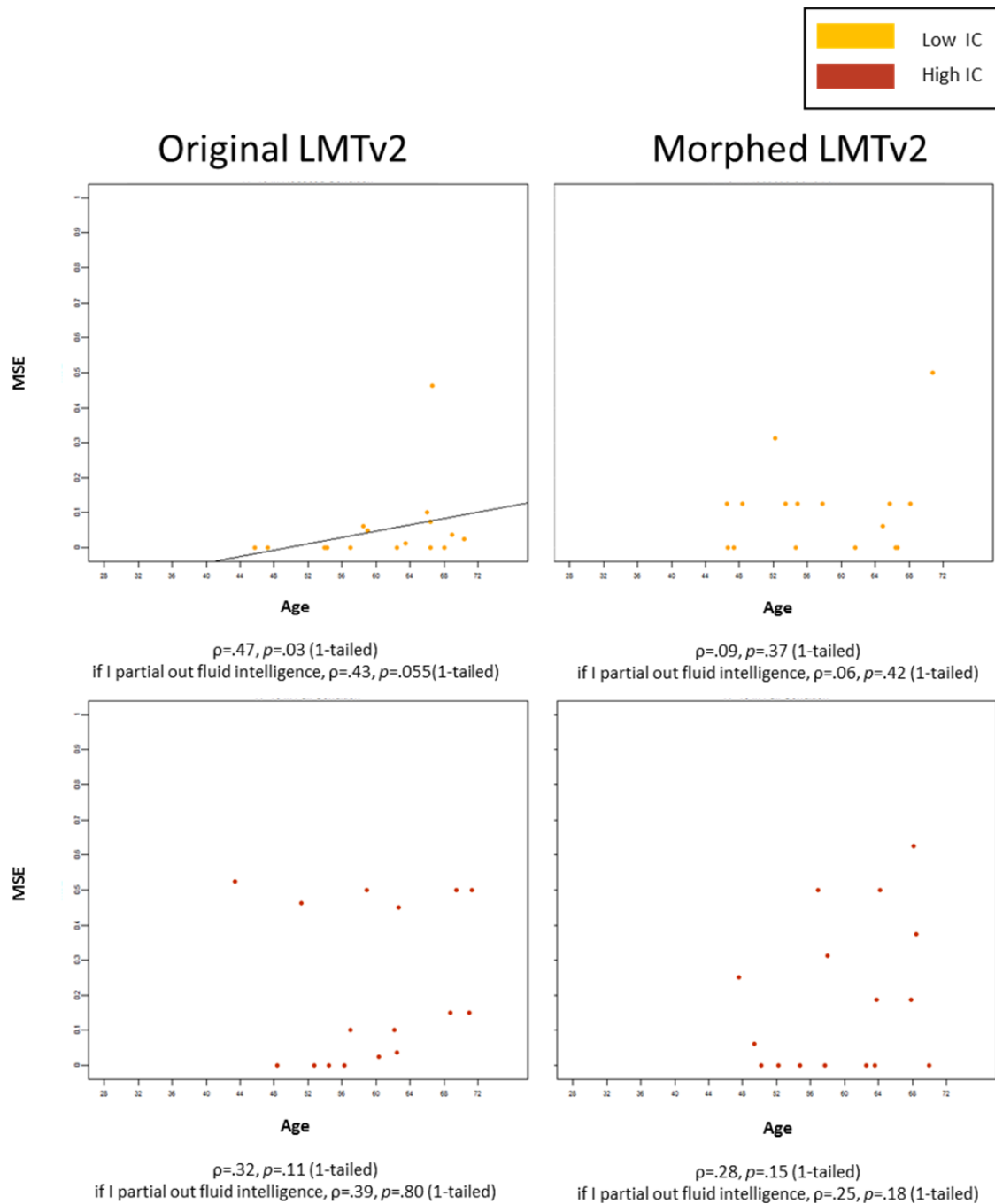
Figure 5.11 Age vs. GN (MSE) in the LMTv2

Associations between age and GN as measured by the MSE, in the four combinations of the LMTv2 as a result of crossing Structure (GN-like, left column; PM-like, right-column) and IC (High, red; Low, yellow). There was a positive association between age and MSE in the Low IC original LMTv2 and this was at borderline significance when fluid intelligence score were accounted for (top left plot).
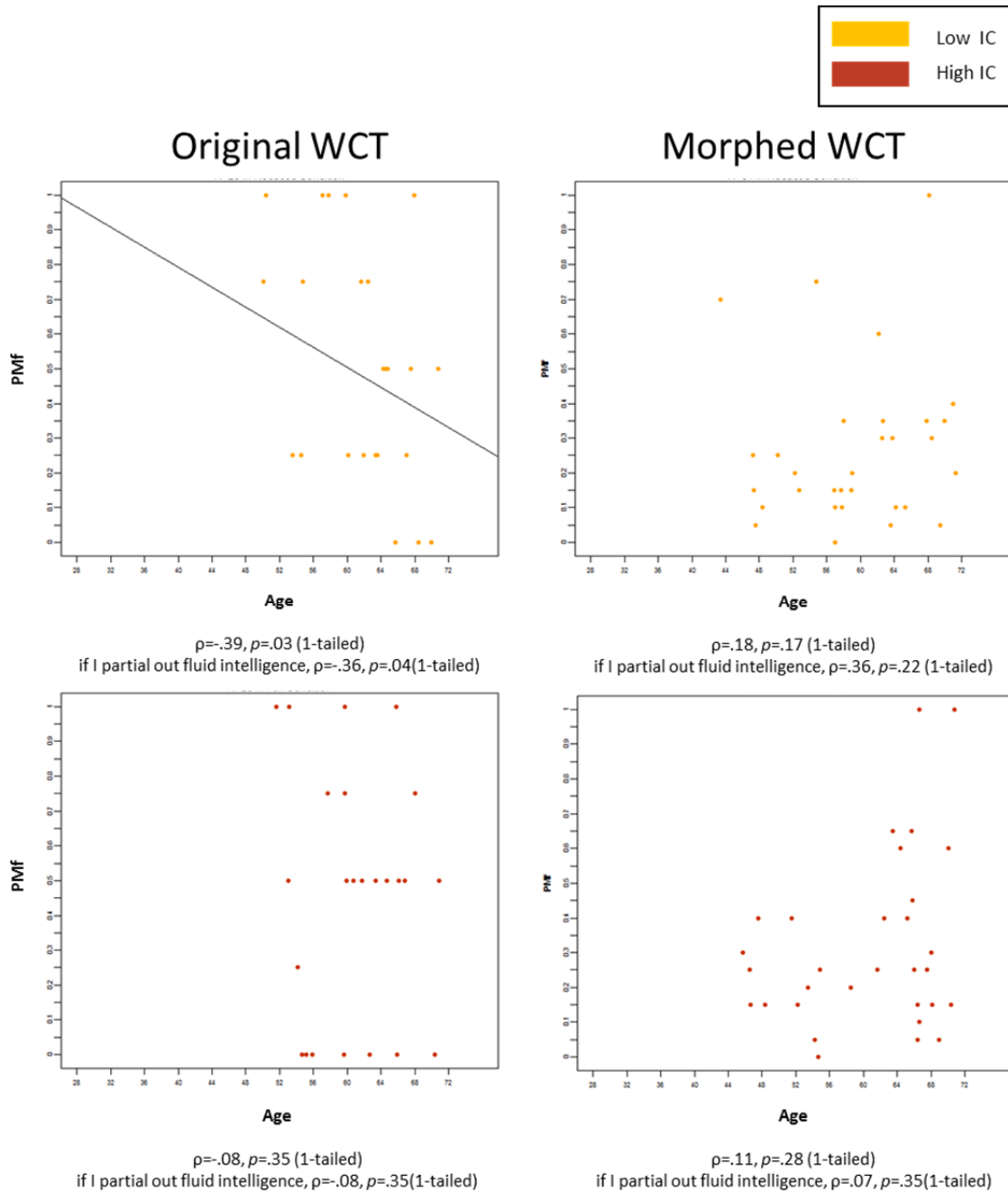
Figure 5.12 Age vs. PMf in the WCT

Associations between age and PMf in the four combinations of the WCT, as a result of crossing Structure (PM-like, left column; GN-like, right column) and IC (High, red; Low, yellow). An unexpected result was observed whereby in the Low IC condition of the original WCT, older participants showed significantly less PMf compared to younger participants (top left plot).

5.3.5   CLUMPINESS OF ERRORS ANALYSIS

As described in the method section (p.148), the entropy-like measure was calculated for eligible profiles. The updated sample sizes, entropy-like measure and proportional position of GN/PMf values are presented in Table 5.11.

Table 5.11 Entropy and position of errors in Experiments 5-8

	LMTv2 Original (GN-like)	LMTv2 Morphed (PM-like)	WCT Original (PM-like)	WCT Morphed (GN-like)
N	15	7	31	57
Number of neglectable trials per participant	20	4	4	20
Proportional position of GN/PMf				
Median	.13	.75	.75	.40
Mean(SD)	.21(.19)	.68(.28)	.65(.18)	.48(.21)
Entropy-like measure				
Median	.50	.56	1.04	1.69
Mean(SD)	.77(.61)	.77(.26)	.93(.24)	1.60(.60)

The N displayed here reflects the sample sizes for participants who showed a minimum of one failure and one correct neglectable trial, which is the necessary condition to calculate clumpiness. Positions of GN/PMf were converted into proportions within each experiment to make them comparable across experiments. The table shows for example .75 for the morphed LMTv2 and original WCT indicating that most errors occurred towards the end of the task, whilst for the original LMTv2 the value of .13 suggests that most errors occurred early on.

I used a linear regression to test if the clumpiness of errors was predictable by: the median position of errors, any of the 3 main factors, and individual differences. In this model, I entered entropy-like measure (H_r) as the dependent variable. The independent variables were the predictors of interest, their interactions, plus a few others to control for potential confounds. The exact list of independent variables was as follows:

- *"Median position of GN/PMf"* variable.

- *"Structure"* factor, which codes for whether the task was in the PM-like (1) or GN-like (0) state.

- *"Task"* factor, which codes for the task being either WCT (1) or the LMTv2 (0).

- *"IC"* factor, which is either High (1) or Low (0) Instructional Complexity condition

- Interaction factors: *"Structure*Task", "Structure*IC", "IC*Task"*

- *"Fluid intelligence"*

- "*Age*"

- *"Number of errors"* variables. These variables were introduced to capture the dependency of the entropy-like measure (H_r) on the number of events which would be a confound otherwise. A non-linear relationship was observed between H_r and number of errors and to control for this in the regression, I entered first, second, and third degree expansions of this GN term. To prevent problems related to multicollinearity, I used orthogonal polynomials.

- *Participant* variables, which coded and controlled for cases by the same individual across the WCT and LMTv2 tasks.

The independent variables were entered into the model simultaneously ("enter" method) and the results of the linear regression are show in Table 5.12.

Table 5.12 Regression analysis output of clumpiness across Experiments 5-8

	B	SE B	β	*p*
Constant	.68	.50		.176
Median position of GN/PMf	.37	.17	.15	.031*
Structure factor	-.19	.22	-.15	.395
Task factor	.32	.18	.21	.075
IC factor	.10	.18	.08	.592
Task*IC	.-.18	.22	-.14	.425
Structure*IC	.05	.15	.03	.475
Structure*Task	.05	.22	.03	.839
Fluid intelligence	.00	.00	.01	.813
Age	.00	.01	.02	.793
Number of errors polynomial 1st degree	3.75	.42	.57	.000***
Number of errors polynomial 2nd degree	-3.42	.44	-.52	.000***
Number of errors polynomial 3rd degree	-.73	.39	-.11	.063

R^2=.82; *p* < .05; ****p* < .001

B=unstandardised beta value, SE B=standard error of B, β=standardized beta value, which is the amount of standard deviations by which the dependent variable will change given the standard deviation change in the independent variable, and hence is independent of the units of measurement of the variables (Field et al., 2012). The output for the participant variables is not shown here.
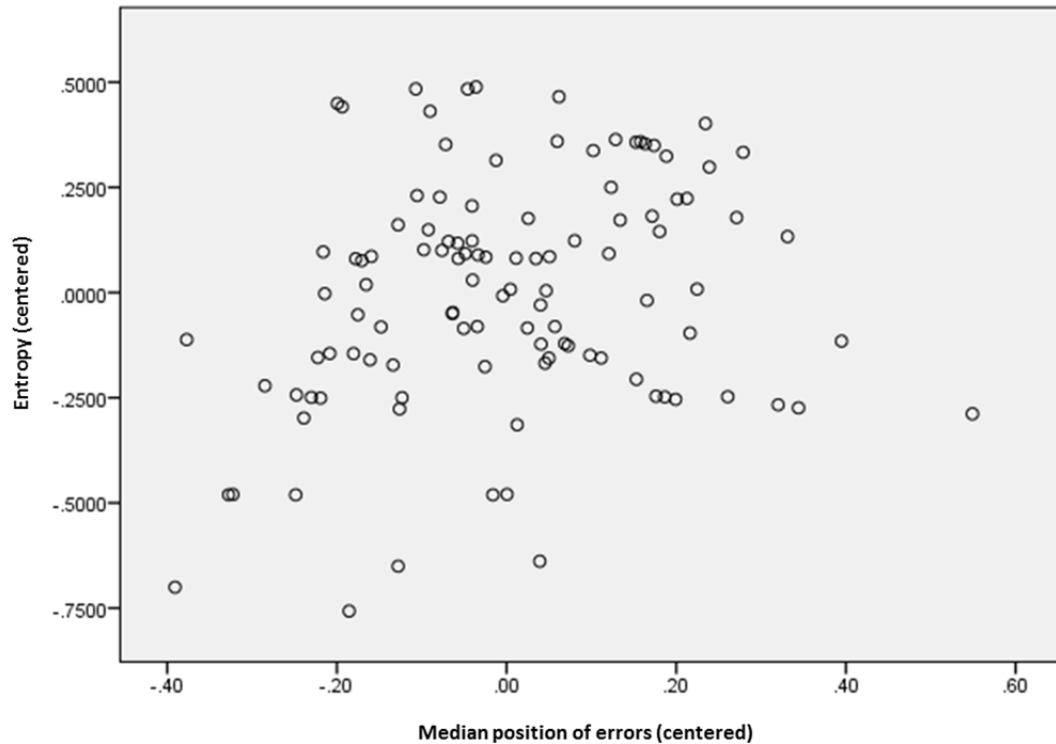
Figure 5.13 Partial regression plot for the Entropy-like measure vs. Median position of errors

The scatterplot illustrates a partial plot of Median position of errors (centered; x-axis) which significantly predicts Entropy (H_t) (centered; y-axis) β=.15,p=.031, even after taking into account factors such as number of errors and other regressors as listed in Table 5.12. This suggests that the later the errors occur, the less clumpy they are.

The median position of errors was found to be significant at predicting the clumpiness of these errors (p=.031) suggesting that even when the number of errors was accounted for, GN/PMf was more clumpy when it was observed early rather than later in the task. I expected the Structure of the task would affect the clumpiness of errors for reasons that are in part based on arguments I presented in Chapter 3. In summary, these arguments predicted that a PM-like structure would elicit largely "monitoring-like" mechanisms which I hypothesized to be associated with less clumpy distribution of errors. Instead, I expected a GN-like structure to be relatively "monitoring" free, and hence to show predominantly "task-model" based failures which I would expect to occur in clusters at the start of the task. However, Structure was not a significant predictor of clumpiness (p=.395, n.s.). Clumpiness of GN or PMf could not be predicted by any of the individual differences measured nor by any interactions across the factors.

One interesting observation of Table 5.11 is how Structure seemed to affect the median position of errors, with a PM-like structure showing errors later in the task compared to a GN-like structure. To test if this was statistically significant, I ran a linear regression with Median position of errors as the dependent variable, and the following independent variables:

- *"Structure"* factor, which codes for whether the task was in the PM-like (1) or GN-like (0) state.
- *"Task"* factor, which codes for the task being either WCT (1) or the LMTv2 (0).
- *"IC"* factor, which is either High (1) or Low (0) Instructional Complexity condition
- Interaction factors: *"Structure*Task", "Structure*IC", "IC*Task"*
- *"Fluid intelligence"*
- "*Age*"
- *"Total number of errors"*
- *"Entropy-like measure"*(H_r)
- *Participant* variables, which coded and controlled for cases by the same individual across the WCT and LMTv2 tasks.

The enter method was used for the regression and the results are shown in Table 5.13.

Table 5.13 Regression analysis output of Median position of errors, across Experiments 5-8

	B	SE B	β	*p*
Constant	.35	.32		.273
Structure factor	.50	.13	.96	.000***
Task factor	.25	.11	.41	.030*
IC factor	.06	.11	.13	.566
Task*IC	-.10	.14	-.20	.452
Structure*IC	.11	.09	.17	.211
Structure*Task	-.29	.14	-.53	.034*
Fluid intelligence	-.00	.00	-.04	.676
Age	-.00	.00	-.10	.301
Total number of errors	.01	.01	.10	.418
Entropy-like measure (H_r)	.09	.05	.22	.107

R^2=.51; *p* < .05; ****p* < .001
B=unstandardised beta value, SE B=standard error of B, β=standardized beta value, which is the amount of standard deviations by which the dependent variable will change given the standard deviation change in the independent variable, and hence is independent of the units of measurement of the variables (Field et al., 2012).

Both Structure and Task factors, and their interaction, were found to significantly predict how late errors occurred in the task (see Table 5.13 for exact statistics and Figure 5.14 and Figure 5.15 for graphical output). Particularly strong was the effect of Structure (β=.96), which is congruent with the initial observation which prompted this analysis. This result suggests that in a PM-like structure errors occur later in the task compared to a GN-like structure in which they appear earlier. This is the case even when controlling for the number of errors, which did not significantly predict if errors tended to appear later or earlier (*p*=.42 n.s.). The effect of Task suggests that errors are more likely to appear later in the WCT compared to the LMTv2. The Task*Structure interaction indicated that morphing the LMTv2 disproportionally delayed the occurrence of errors (from mean position of .21 to .68) when compared to the delay associated with the structural change within the WCT (from mean position of .48 to .65, see Table 5.11). Overall the results of this analysis suggest that the temporal position of the majority of errors differs across the two structures and a closer discussion follows in the next section.
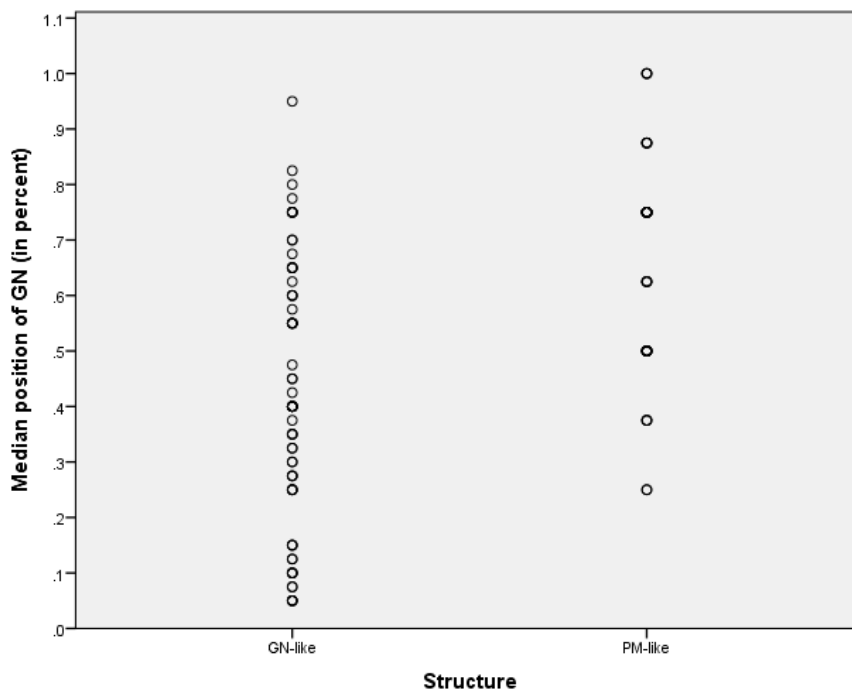
Figure 5.14 Scatterplot of Median position of GN (*y*-axis) against Structure (*x*-axis)

The scatterplot illustrates how a PM-like Structure shows errors later than a GN-like Structure.
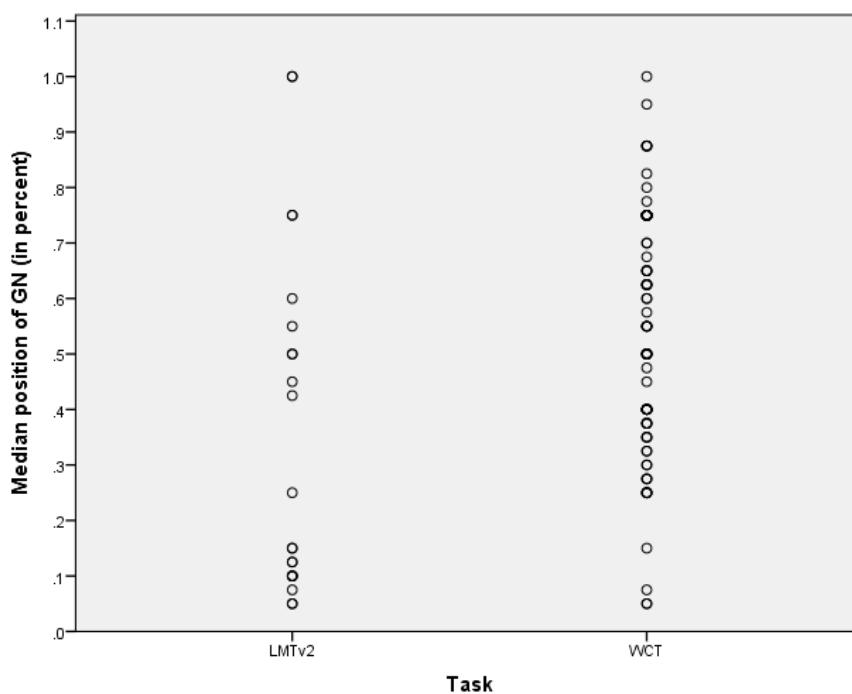
Figure 5.15 Scatterplot of Median position of GN (*y*-axis) against Task (*x*-axis)

The scatterplot illustrates how errors on the LMTv2 appear earlier than the WCT.
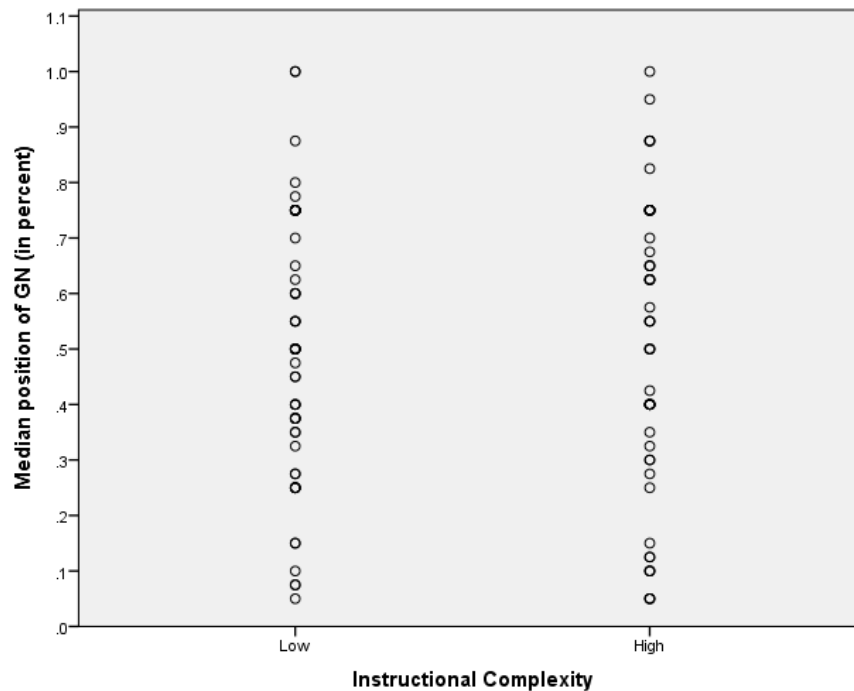
Figure 5.16 Scatterplot of Median position of GN (*y*-axis) against Instructional Complexity (*x*-axis)

The scatterplot illustrates how the Instructional Complexity of the task does not appear to associate differently the position of errors. Please note this is a null finding.

5.4 GENERAL DISCUSSION

My investigation began with the observation that GN and PM paradigms had a similar structure and hence I questioned if GN and PMf were different or were essentially describing the same phenomenon. I assumed that if they were the same then PMf would, like GN, be sensitive to the IC effect. Experiment 5 suggested that this was not the case. I then went on to predict that if I made the PM task structurally even closer to a GN task, then this would make PMf sensitive to IC. Experiment 7 revealed that this was also not the case, although making the GN task more PM-like did remove the IC effect as predicted. The latter finding rests on a null result and hence should be read with caution. I therefore directly tested if I could detect an interaction between Structure and IC using a factorial ANOVA expecting that only a GN-like structure (and not a PM-like structure) would show an IC effect, however this was not significant which converged with the results from the nonparametric tests. Another test of whether GN and PMf were the same

or not, was based on the clumpiness analysis which did not reveal any Task nor any Structure specific differences. Assuming that the entropy-like measure is valid, this finding suggests that the clustering patterns are similar for GN and PMf. The latter may further suggest that underlying mechanisms involved are not significantly different - at least in terms of *monitoring* and *task model* accounts. Again this was based on a null finding and should be interpreted with caution. Finally, an analysis on the position of errors showed that, as expected, errors occurred earlier in a GN-like compared to PM-like structure. However, in addition to this Structure effect there was a Task effect indicating that independent of other factors taken into account in the regression, errors were more likely to occur later in the WCT compared to the LMTv2.

In summary, whether GN and PMf are different or not remains inconclusive: I found a "single" as opposed to "double" dissociation between tasks with a greater IC effect in the original LMTv2 compared to the morphed version, and no IC effect in both versions of the WCT favouring the view that GN and PMf are different; the Task effect in the position-of-errors analysis indicates that they are different; instead, the absence of any difference in the clumpiness-of-errors analysis suggests that they are similar in this regard. In any case, it remains unclear what other aspects of the tasks, apart from practice and the frequency of the target, may contribute to make them more or less sensitive to the IC effect.

What are alternate explanations for the absence of the expected double dissociation between Structure and IC factors? One possibility is that PM tasks with more complex instructions can display better performance. Maylor (1993, 1996; as cited in 2016) reports studies in which superior performance (68% correct) was observed in a condition with complex instructions (name the famous face and circle the trial number if the face has a beard, or cross out the trial number if the face has a pipe) compared to simpler instructions (42% correct ; name the famous face and circle the trial number if the face has glasses). A suggested explanation for this counterintuitive effect is that given a perceived difficulty of the complex instructions, greater effort was employed to encode them and/or they may have been rehearsed more frequently.

Another possibility is due to at least a couple of limitations to the study. The first is the assumption that the IC effect is a litmus test for GN, which is not necessarily the case

although it has been observed in four GN tasks: the LMT, the feature match task (Duncan et al., 2008), the panel-decision task (Bhandari and Duncan, 2014) and the SGNT (Chapter4). Another limitation which is not new to PM research (Risko, 2015), is related to how investigating infrequent events can lead to unreliable results due to decreased precision and power. This is especially the case in the analysis involving a comparison across levels of Structure, where the PM-like structure only had 4 data points compared to the 20-trial GN-like structure. However, data for the PM-like structure were powerful enough to reveal significant correlations with IQ in the LMTv2, and with age in the original WCT.

An important finding in this study was the effect of Structure on both the number of errors and their temporal position in the task. The ANOVA revealed a main effect of Structure indicating that a PM-like structure produced more errors than a GN-like one. Of the two parameters manipulated in the PM-like structure, (i) extended practice, and (ii) low frequency of PM targets/critical events, it is likely that the increased error rate was driven by (ii) rather than (i). How could the probability of the target affect performance? One possibility is that as the frequency of the critical target decreases, the repeated non-appearance may decrease expectations that it will occur in the future and as a consequence less attentional resources are allocated to it. The less the attentional resources, the more likely the failure. Instead, a higher occurrence of critical events may encourage active rehearsal of this sub-task, thus improving its performance. Similar findings have been reported by various studies. For example Kane and Engle (2003) manipulated the proportion of congruent and incongruent trials in a Stroop task and when the incongruent trials were rarer more errors were reported. Similarly in a study by Park *et al.,*(1997) in which, within the same timeframe, one experimental condition had 6 PM targets whilst the other 12 PM targets. Although not statistically significant, it was the 6 PM target condition which nominally had more errors and was considered harder in terms of strategic demands. Some PM researchers report attempting to avoid "active rehearsal" by keeping the frequency of PM cues between 5-10% and ensure at least five ongoing task trials between PM events (West, 2008). If this guideline were valid, then GN-like structure with around 16% of critical events would be likely to show more *monitoring* for the critical sub-task compared to the 2-3% of the PM-like structure. However, the only way to measure if this was the case in this study, was via the experimental measure of

clumpiness: if the entropy-like measure was larger in the PM-like compared to the GN-like structure, then this might suggest increased reliance on monitoring in the former format. However, Structure was not a significant predictor of clumpiness.

Nevertheless, even though Structure did not seem to predict the clustering of errors, it strongly and significantly predicted their temporal position. This indicated that PM-like and GN-like structures have different learning functions, with a typical learning curve characterising the GN-like structure (more errors at the start) whilst a PM-like structure showed relatively inefficient learning/forgetting of the critical rule with most errors occurring late in the task. A possible interpretation of this finding is similar to the one mentioned earlier –critical targets that are more sparse may decrease the perceived likelihood that they will re-occur to the extent that they may be forgotten and hence neglected more consistently later in the task. The work on GN presented by Bhandari and Duncan (2014) also suggested that temporal dynamics of errors may follow a pattern characterized by initial instability (a mix of correct and incorrect), followed by a stable pattern of behaviour (mostly correct or mostly incorrect). My results add to this by suggesting that the Structure of the task is important in determining the kind of temporal signature that the errors will make. So, Structure predicts position of errors, position of errors predicts clumpiness, but Structure does not predict clumpiness. This inconsistency suggests that other factors are important in the relationship between Structure, position of errors and entropy-like measure, alternatively the absence of a Structure-clumpiness finding was a false negative or one/both of the other 2 significant findings was a false positive. The next paragraph discuss other factors that may be of importance.

The lack of a "double" dissociation between Structure and IC also indicates that there must be other characteristics of the tasks which interact significantly with IC. What could these characteristics be? There are many possibilities, but two interesting candidates I propose are (i) the focality of the critical/PM cue, and (ii) the speed of the task. In the WCT, the cue is "non-focal", which means that the processes involved in the ongoing task (taking semantic decisions) do not trigger nor overlap with processes required to detect the PM cue (string of letters "tor"). Instead, if the PM cue were the word "tornado" then simply reading this word would have served as a reminder of the PM task, and hence been a "focal" cue. Non-focal cues are thought to require more active efforts and strategies to "monitor" them successfully throughout the task (Einstein et al., 2005 ; see

also Chapter 3, section 3.1.2 for more on monitoring vs. spontaneous processes). The critical cues in the LMTv2 (the second side-cue symbols, "+" or "-") were not focal in this sense either, but arguably, they were relatively less non-focal compared to the PM cues because when these "+" and "-" symbols appeared, the regular task (letter reading) was not their immediate competitor that would otherwise serve to decrease the salience of the cues. Furthermore, the impact of a non-focal cue may be amplified in this particular study considering that the mean age was approximately 60 years and that older participants were found to be disproportionally worse on lab-based PM tasks with non-focal cues (Kliegel et al., 2008).

The other factor that I suggest as potentially influential in determining the density of GN/PMf is the timed nature of the task. The WCT is self-paced whilst the LMT is a speeded task. It is worth noting that the IC effect has generally been observed in speeded tasks like the LMT, feature match task (Duncan et al., 2008) and SGNT (Chapter 4) but not in the self-paced WCT. Hence, it is possible that the task speed interacts with IC (Iveson et al., 2017). Related to this is a study by Chuderski (2016) who suggests that if speeded intelligence tests compared to non-speeded ones affect the relationship between constructs of working memory and fluid intelligence; these are isomorphic in speeded versions, but their relationship is only moderate in non-speeded variants. Perhaps, the self-paced nature of the WCT may have buffered off the load introduced in conditions of High IC. Having said this, an IC effect has also been observed in the panel-decision task of Bhandari & Duncan (2014), which had a very generous time limit of 20 s compared to the mean reaction time of less than 6 s.

The relationship between individual differences and GN/PMf produced both expected and unexpected results. As found in previous published studies (Duncan et al., 1996; Duncan et al., 2008; Bhandari and Duncan, 2014), fluid intelligence predicted GN in most conditions of the LMTv2. An additional analysis suggested that the relationship between fluid intelligence and errors in the LMTv2 increased with IC and was also higher when the structure was in the original GN-like form. Instead, fluid intelligence was not significantly associated to PMf in any version of the WCT which suggests that the extent of practice and frequency of PM targets are not sufficient to mediate this association. Other studies also report an absence of correlation between PM and fluid intelligence (Shallice and Burgess, 1991; Roca et al., 2014), however these tend to be multitasking

tasks which are closer to the time-based PM category than the event-based PM type. Multitasking tasks such as the Six Elements Task (Shallice and Burgess, 1991) or the Hotel Task (Manly et al., 2002) measure the ability to self-initiate the whole set of 6 available sub-tasks, which are open ended and within a time limit. So, the question remains, what aspects of the WCT contribute to this lack of association? The focality of the cue? The speed of the task? Maybe the speed of the task, given that processing speed has been long found to be strongly associated to fluid intelligence (Conway et al., 2002). Or perhaps it's related to the specific types of rules, such that the left/right rules (LMT) are more dependent on fluid intelligence compared to semantic judgments (WCT)? Data from this study cannot address these questions.

An unexpected result, which is technically insignificant, in the individual differences analysis was the negative relationship of PMf and age. Studies have previously reported the so called "age-PM paradox" (Henry et al., 2004; Schnitzspahn et al., 2011; Kliegel et al., 2016), whereby older participants appear to have an advantage over younger participants in naturalistic PM tasks, but a disadvantage in lab-based PM tasks. Although more focused investigations are required, a suggested explanation for this paradox is that compared to younger participants, older participants in naturalistic settings may show (i) relatively less stress and absorption in daily tasks and (ii) higher motivation, compared to younger participants (Schnitzspahn et al., 2011; Kliegel et al., 2016). Nevertheless, this account of the age-PM paradox cannot directly explain the odd finding I report because this showed an age advantage in a lab-based task, not a naturalistic one. Perhaps similar motivational factors may explain this finding, for example the WCT task may have been particularly preferable to older participants. This remains an unexplained and puzzling result.

One promising result is the replication that clumpiness of errors decreases as the majority of the errors occur later on in the task, replicating findings from Chapter 3. This finding supports the hypothesis that if the errors are largely driven by a failure in the *task model* then most of these failures should occur at the start of the task and in a relatively clumpy way. Instead, if the observed GN/PMf are mostly due to *monitoring* impairments then these are predicted to wax and wane across the task as the limited attentional resources fluctuate; so, errors would tend to have a higher median position and be less clumpy.

5.5 CONCLUSION

My initial research question was whether GN and PMf were essentially the same phenomenon or not. If sensitivity to the IC effect is a true indicator of GN and considering the Task by IC interaction in determining the number of errors, then the answer is no - GN and PMf are different. However, as my investigation unfolded I realized that understanding whether GN and PM were equivalent was not necessarily the most interesting inquiry after all. Instead, the more interesting question became: given a set of task parameters, what pattern of errors can we predict and what does this pattern suggest of the underlying cognitive mechanisms? In line with this framework, the second research question attempted to understand which task factors were important for the lack of an IC effect in the PM task. The method involved manipulating the Structure of the task by changing 2 parameters: the extent of practice, and the frequency of targets. The results indicated that although increasing practice was enough to remove the IC from the LMT, decreasing practice and increasing the frequency of targets were not sufficient to introduce the IC effect in the PM task. Overall, the non-significant interaction between Structure and IC suggested that frequency of targets and extent of practice are not sufficient to drive the IC effect.

However, the most important finding was that Structure could predict the number of errors and their temporal position: a PM-like structure produced more errors and these occurred later in the task, whilst a GN-like structure was associated with both fewer and earlier failures. Analyses using the experimental measure of clumpiness did not reveal that the clustering of errors was different across these structures, which was interpreted as no evidence to reject the null hypothesis that similar cognitive mechanisms were being used. Further focused investigations are required to refine which, and how, other task structure parameters influence goal-directed behaviour which may also explain the IC effect more fully. Candidate parameters include the speed of the task and the focality of the target cue. In addition, issues related to power should be addressed by using longer tasks and larger samples.

Chapter 6. CONCLUSIONS

6.1 WHAT KNOWLEDGE WAS GAINED?

This dissertation aimed to add to our understanding of cognitive control by characterising how a particular failure of performance, GN, is affected by different forms of complexity manipulations. In Chapter 2, I developed a new task to test GN and unlike previous studies, I manipulated complexity qualitatively by altering the semantic transparency of the instructional cue. GN was sensitive to this kind of complexity manipulation and additional analyses indicated that this could not be solely explained by retrieval demands but instead was linked to a failure in recognizing the significance of the instructional cues. In Chapter 3, I presented two theories, the *monitoring* and the *task model* accounts of how GN may occur. I then proposed a new entropy-like measure to quantify the temporal clustering or, "clumpiness", of GN and used this to test the differential temporal patterns that the two different accounts predict. The results suggested that both models are likely to be operant, but with their relative dominance being different across time: GN which appears early on in the task is mostly driven by failures which are *task model* like, whilst GN which appears later is better aligned with *monitoring* deficits. Chapter 2 also revealed that GN can be sensitive to manipulations of complexity *during* task performance (ATC), which motivated the question of whether previously published studies suggesting the contrary were perhaps due to insufficient complexity. Hence, in Chapter 4, using the new GN task, I tested for two possibilities: i) the absence of a complexity effect *during* the task (ATC) and, ii) the presence of a complexity effect *before* the start of the task (IC). Results replicated both the published null findings of ATC and, a significant IC effect on GN. How does this reconcile with the positive ATC result from Chapter 2? Additional analyses indicate that ATC does not affect GN unless this manipulation is of a qualitatively different type (e.g. semantic manipulation of the instructional cue) and/or, is more closely linked/more proximal to the critical event episode.

In Chapters 2-4, I consistently referred to models and empirical evidence from the PM literature given the apparent similarity between PM and GN experimental paradigms thus provided a unique opportunity to constrain theorizing on these phenomena. In Chapter 5, I took a closer and systematic approach to this comparison to further understand how PM failures and GN are different, if at all, with the broader aim to integrate what are otherwise isolated domains. I tested this by developing morphs of typical paradigms from these two domains and applied IC as a litmus test for GN. A positive result would have provided strong evidence that they are similar. However, I found a mixture of null findings which, although technically inconclusive because they could be driven by insufficient power to detect the effects (and assuming that the litmus test was valid) are otherwise congruent with an account that GN and PMf reflect different capacities. Further support to this view comes from the finding that fluid intelligence was associated with GN but not PMf. Perhaps GN and PMf reflect one general type of error failure, which is heterogeneous in nature. This would be similar to the phenomenon of hemispatial neglect which manifests in various forms. The variety of hemispatial neglect is thought to reflect combinations of spared and impaired space representations of different types, for example, egocentric space vs. allocentric space, near vs. far space and personal vs. peripersonal space (Ward, 2015). In sum, this heterogeneity is thought to reflect different ways in which attentional processes operate across different spatial maps in the brain. Similarly, it would be interesting to test if the differences across GN and PMf reflect impairments across categorically different "task maps".

Whilst investigating the differences between GN and PMf, a much more interesting question emerged with respect to what structural features of a task predict different signatures of GN/PMf-like errors. This theory-neutral approach focused on two structural task parameters: the amount of practice and frequency of the critical event. The key finding was a general rule about task structure: a combination of extended practice and low frequency of critical events predicts both a larger amount of errors and with more of these occurring late in the task. The design of the task did not permit to disentangle the separate contributions of these two structural factors; however common sense suggests that the increased error rate was mostly driven by the low frequency of targets rather than increased practice. Of course, this remains to be tested empirically and further work needs to be done using this systematic approach to assess other task characteristics and their

resulting error signatures. Other task characteristics include the focality of the cue, the speed of the task and task features which may affect the perceived difficulty of the task hence leading to differential allocation of attentional resources.

Overall, like previous reports (Bhandari and Duncan, 2014), complexity effects were found to be bounded by the task structure. However my novel finding was with respect to a different aspect of task structure. Whilst in Bhandari's study task structure referred to the task set, which was clearly set out *before* the start of the task, in my work task structure refers to i) the duration of practice and, ii) the frequency of targets which, at least for the latter, is only discovered *during* task performance.

6.2 WHAT ARE THE LIMITATIONS OF THESE STUDIES?

One possible caveat to my work is the validity of the method I used to assess both comprehension and memory of the task instructions before and at the end of the tasks. The method involved the use of prompts to facilitate the recall of instructions (see Appendices for verbatim task instructions). Does it suffice for participants to correctly answer these prompts in order to ensure that they have understood the instructions? Does their language match their conceptual knowledge? This is hardly a new question at all and is a central debate in philosophy and psychology (e.g. see Jean Piaget, Vygotsky, Chomsky). Critics of the developmental psychologist, Jean Piaget, clearly demonstrated that rewording the task instructions and/or probes used to test conceptual knowledge washed out the effects that Piaget had previously claimed (although robust counterarguments also followed, for example see Lourenço and Machado, 1996). If, for a moment, I consider cued recall an invalid way to test comprehension of task instructions, such that it may have included cases in which participants did *not* effectively understand what they were meant to do, then the resulting GN could be interpreted differently. GN could be simply interpreted as being entirely a result of a faulty plan without any claim to any executive deficits, in other words, the plan of action is followed exactly, but the plan is inadequate to achieve the goal as intended by the experimenter. James Reason refers to this kind of error as a "mistake", and contrasts it to "slips" which is when the plan of action may be entirely appropriate, but the actions do not go as planned (Reason, 2013). The point of this consideration is that it highlights how measuring the discrepancy between what is reported verbally in the cued recall and what is actually understood may

actually sit at the core of the phenomenon of GN. Not surprisingly, improving the way to measure the comprehension and memory of task instructions without introducing a practice effect and washing out GN, is far from straightforward.

Some other limitations concern the way I measured responses in the actual tasks. Generally speaking, GN/PMf were based on a simple binary measure (did or did not incorrectly perform the inappropriate sub-task) and therefore is likely to include both cases of successful and unsuccessful triggering of the critical response at some other level of implicit processing. The RT analyses (cue sensitivity and cue accessibility) that I ran in Chapter 2 were a way to detect these implicit processes. However, one way this could be improved is to use a response which allows for a continuous measurement instead of the binary outcome of a button press. In addition, measuring confidence scores on participant's responses via self-report could be a way to further gauge dissociations between knowledge/intention and action.

A second concern related to measuring responses is that the complexity of stimulus-response mappings differed across the tasks that I used. For example, in the SGNT, GN was determined if on critical events the task registered any response from the Emotion sub-task response box (instead of the other response box). Hence, although the scoring is based on a binary decision (Emotion response box, or not?), the participant is in effect faced with 6 button press options plus the option to withdraw response (see Figure 2.4). Instead, in the LMT the response options are 3: read from Right, read from Left or do not read at all and similarly for the WCT: press yes, press no, press q (withdrawing response is not an option in the WCT given that it is a self-paced task). Hence, considering that the more response alternatives, the more likely to choose incorrectly just by chance (Duncan, 1980), then in this respect the SGNT might be more likely to show errors than the WCT and LMT.

One thought that has puzzled me is the fact that it is not always clear how to disentangle the IC effect from an ATC effect. For example, the transparency manipulation used in Chapter 2 loads on both IC and ATC, such that it affects both initial "encoding" and later "retrieval"(a nontransparent cue is both more difficult to learn to start with and to decode during task performance). My RT analyses indicated that the effect was driven by complexity other than just the one affecting immediate retrieval and that instead it also

affects process involved in noticing that something should be retrieved to start with. However, what remains unresolved is how to gauge the separate contributions of these (assumed) separate factors.

Two additional concerns are with respect to validity. First, is construct validity: future experiments should involve various PM and GN tasks, rather than just single exemplars of these potentially different constructs. Secondly, is ecological validity: If overlooked this might lead to false inferences (for example, see Burgess et al., 1998). How does GN as measured in the lab map to everyday functioning? Is the abstract nature of the tasks used in these experiments introducing a confound to the kind of capacity used in everyday tasks? Does this confound affect individuals differently? The effectiveness of any cue, stimulus and task is never absolute, but rather is relative to the context and the interaction of this context with the cue/stimulus/task itself (Marsh et al., 2008). One related question is whether the association or lack of association I find between performance on the 3 tasks used (LMT, SGNT and WCT) and fluid intelligence (high in the LMT, less strong in the SGNT and hardly any association in the WCT) may be mediated at least in part by the ecological validity of these tasks. Notably, there's a gradient of abstractedness, across these tasks, with LMT being the most abstract, and WCT the least abstract. By abstractedness here I specifically mean how far removed they are from real-world-activities. The way people approach abstract lab-based tasks may be different to the way they approach every day problem solving. For instance, context may affect the motivation people have towards solving the problem. Motivation and mood factors have been shown to affect cognitive control (Pessoa, 2009; Lagner et al., 2014) but were not measured in any of my experiments. In summary, it would be interesting to test GN using more ecologically valid tasks which also take into account motivational aspects.

6.3   FUTURE DIRECTIONS

The objective of this dissertation was to advance *mechanistic* accounts of cognitive control. To qualify as a mechanistic account this should accurately represent the real parts or processes producing the behaviour (e.g. brain components) (Craver, 2006; Bechtel and Abrahamsen, 2010; Wilson and Golonka, 2016). However, the advances I make in the dissertation are more appropriately defined as *functional models*. Although functional models have some common features with mechanistic models, they are fundamentally

different because of the way they link to the underlying system whose organization they aim to represent (Weiskopf, 2011; Wilson and Golonka, 2016). The components that functional models refer to are not real components (actual tangible structures, e.g. brain parts); instead they typically refer to sub-capacities which may only vaguely map onto real components (Weiskopf, 2011). One way to evaluate a model is to distinguish it across the range of "how-possibly", "how-plausibly", and "how-actually" the behaviour happens, with the former being descriptive and the latter, more mechanistic in that the model is able to produce the behaviour in question (Craver 2007 as cited in Weiskopf, 2011). Although the knowledge gained in this dissertation has helped explain how certain performance failures reflect distinct capacities and which task parameters may be important to elicit them, the current model still leans heavily on the "how-possible" end of the spectrum. In fact, not everyone agrees that functional models can provide explanations or even offer a trajectory towards mechanistic accounts (Wilson and Golonka, 2016). However, if building a mechanistic model of a behaviour involves first empirically characterising the mechanism and then identifying the real components that actually make the mechanism leading to the production of the behaviour in question (Wilson and Golonka, 2016), then this dissertation has added to the first step.

A next step could map this characterisation to biologically plausible components. Neuroimaging studies are an obvious candidate to further this end. Some work has already been done as previously mentioned in Chapter 1, with indirect links between GN and the fronto-parietal (MD) network (Woolgar et al., 2010; Duncan, 2013). However, more incisive research is required to develop direct associations between GN and its underlying capacity to brain mechanisms. Recent developments in fMRI methods using multi-voxel pattern analyses and semi-hidden Markov models appear to be a promising way forward to investigate such mental operations and their temporal durations (Anderson, 2016). Other examples of mechanistic accounts can be found elsewhere for example, Badre's and Frank's work on hierarchical reinforcement learning (Badre and Frank, 2012; Frank and Badre, 2012) or research on circadian rhythms (Bechtel and Abrahamsen, 2010). A methodology which may be particularly useful in mapping GN is EEG. For example, error-related negativity, an electrophysiological marker of early error monitoring can differentiate between error types of different significance (Maier et al., 2011; Maier and Steinhauser, 2013). In addition, other ERP components such as the

lateralized readiness potential could be useful to disentangle perceptual from motor components (Szűcs et al., 2009). However, others warn that mapping cognition to the brain may be limited in principle:

> *While evidence of the lower level structure of a system can inform, constrain, and guide the construction of a theory of its higher level structure, lower level structures are not simple maps of higher level ones. Thus in psychology we have the obvious, if depressing, truth that the mind cannot simply be read off of the brain. Even if brains were less than staggeringly complex, it would still be an open question whether the organization that one discovers in the brain is the same as the one that structures the mind, and vice versa.(p.328 Weiskopf, 2011)*

To conclude, a major challenge for future research is to provide mechanistic accounts of cognitive control which can explain why and how certain types of processing must be controlled, how these operations occur in the brain and how these processes interact with internal factors such as age, pathology/injury and external factors such as the way a task is presented. The outcome of such research could support cognitively and biologically informed interventions targeted at improving cognitive control by directly affecting internal mechanisms (e.g. drug therapy) or by modifying the environment to alleviate effects of complexity (e.g. organizing information differently). Finally, my last chapter involved a switch in focus to task parameters and their resulting error signatures, as opposed to a focus on task labels ("GN" vs "PM"). This highlighted the importance of supporting theory-neutral approaches to research (exemplified by the Cognitive Atlas, Poldrack et al., 2011) in order to integrate and maximize our existing and future scientific knowledge.

6.4 CONCLUSIONS

Although it is far from clear how the cognitive processes work mechanistically to allow us to parse complex information into successful goal-directed behaviour, this body of work adds to our knowledge of how different task conditions can affect our attentional processes in different ways resulting in different error signatures.

BIBLIOGRAPHY

Altamirano LJ, Miyake A, Whitmer AJ (2010) When Mental Inflexibility Facilitates Executive Control. Psychological Science 21:1377-1382.

Anderson JR (2016) Hidden Stages of Cognition Revealed in Patterns of Brain Activation. Psychological science 27:12-15.

Anderson MC, Bjork RA, Bjork EL (1994) Remembering can cause forgetting: Retrieval dynamics in long-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 20:1063-1087.

Arbuthnott KD, Woodward TS (2002) The influence of cue-task association and location on switch cost and alternating-switch cost. Canadian Journal of Experimental Psychology 56:18-29.

Aron AR, Monsell S, Sahakian BJ, Robbins TW (2004) A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. Brain 127:1561-1573.

Astle DE, Scerif G (2009) Using developmental cognitive neuroscience to study behavioral and attentional control. Developmental Psychobiology 51:107-118.

Atkinson RC, Shiffrin RM (1968) Human memory: A proposed system and its control processes. In: The psychology of learning and motivation: Advances in research and theory (Spence KW, Spence JT, eds). New York: Academic Press.

Baddeley AD (1986) Working memory. Oxford: Oxford : Clarendon Press.

Baddeley AD, Thomson N, Buchanan M (1975) Word length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behavior 14:575-589.

Badre D (2015) Prefrontal cortex and the hierarchical control of behavior. In. Cognition and Brain Sciences Unit, Cambridge.

Badre D, D'Esposito M (2009) Is the rostro-caudal axis of the frontal lobe hierarchical? Nature Reviews Neuroscience 10:659-669.

Badre D, Frank MJ (2012) Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: Evidence from fMRI. Cerebral Cortex 22:527-536.

Battig WF, Montague WE (1969) Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. Journal of Experimental Psychology 80:1-46.

Bechtel W, Abrahamsen A (2010) Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. Studies in History and Philosophy of Science Part A 41:321-333.

Bhandari A, Duncan J (2014) Goal neglect and knowledge chunking in the construction of novel behaviour. Cognition 130:11-30.

Bhandari A, Badre D (2016) A Nimble Working Memory. Neuron 91:503-505.

Bianchi L (1922) The Mechanism of the Brain and the Function of the Frontal Lobes. Edinburgh: Livingstone.

Bishop SJ, Fossella J, Croucher CJ, Duncan J (2008) COMT val158met genotype affects recruitment of neural mechanisms supporting fluid intelligence. Cerebral Cortex 18:2132-2140.

Bisiach E (1996) Unilateral Neglect and the Structure of Space Representation. Current Directions in Psychological Science 5:62-65.

Botvinick MM, Carter CS, Braver TS, Barch DM, Cohen JD (2001) Conflict monitoring and cognitive control. Psychological Review 108:624-652.

Bowling A, Hindman E, Donnelly J (2012) Prosaccade errors in the antisaccade task: differences between corrected and uncorrected errors and links to neuropsychological tests. Experimental Brain Research 216:169-179.

Brainard DH (1997) The Psychophysics Toolbox. Spatial Vision 10:433-436.

Broadbent DE (1958) Perception and communication. Oxford: Oxford : Pergamon.

Burgess PW, Alderman N, Evans J, Emslie H, Wilson BA (1998) The ecological validity of tests of executive function. Journal of the International Neuropsychological Society 4:547-558.

Campoy G (2008) The effect of word length in short-term memory: Is rehearsal necessary? Quarterly Journal of Experimental Psychology 61:724-734.

Cattell RB (1971) Abilities: Their structure, growth, and action. Abilities: Their Structure, Growth, and Action.

Cattell RB, Cattell A (1973) Measuring intelligence with the culture fair tests. Institute for Personality and Ability Testing.

Chatham CH, Badre D (2015) How to test cognitive theory with fMRI. In: arXiv:1404.2917.

Cipolotti L, Spanò B, Healy C, Tudor-Sfetea C, Chan E, White M, Biondo F, Duncan J, Shallice T, Bozzali M (2016) Inhibition processes are dissociable and lateralized in human prefrontal cortex. Neuropsychologia 93:1-12.

Conway ARA, Cowan N, Bunting MF, Therriault DJ, Minkoff SRB (2002) A Latent Variable Analysis of Working Memory Capacity, Short-Term Memory Capacity, Processing Speed, and General Fluid Intelligence. Intelligence 30:163-183.

Craver CF (2006) When mechanistic models explain. Synthese 153:355-376.

Crittenden BM, Mitchell DJ, Duncan J (2015) Recruitment of the default mode network during a demanding act of executive control. eLife 2015.

Cullen B, Brennan D, Manly T, Evans JJ (2016) Towards Validation of a New Computerised Test of Goal Neglect: Preliminary Evidence from Clinical and Neuroimaging Pilot Studies. PLoS ONE 11:e0148127.

De Jong R, Berendsen E, Cools R (1999) Goal neglect and inhibitory limitations: dissociable causes of interference effects in conflict situations. Acta Psychologica 101:379-394.

Deary IJ, Penke L, Johnson W (2010) The neuroscience of human intelligence differences. Nat Rev Neurosci 11:201-211.

Demakis GJ (2003) A Meta-Analytic Review of the Sensitivity of the Wisconsin Card Sorting Test to Frontal and Lateralized Frontal Brain Damage. Neuropsychology 17:255-264.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18:193-222.

Dillard MB, Warm JS, Funke GJ, Funke ME, Finomore VS, Matthews G, Shaw TH, Parasuraman R (2014) The Sustained Attention to Response Task (SART) Does Not Promote Mindlessness During Vigilance Performance. Human Factors 56:1364-1379.

Duncan J (1980) The demonstration of capacity limitation. Cognitive Psychology 12:75-96.

Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. Nature Reviews Neuroscience 2:820-829.

Duncan J (2010a) How intelligence happens. York, PA: Yale University Press.

Duncan J (2010b) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in Cognitive Sciences 14:172-179.

Duncan J (2013) The Structure of Cognition: Attentional Episodes in Mind and Brain. Neuron 80:35-50.

Duncan J, Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends in Neurosciences 23:475-483.

Duncan J, Burgess P, Emslie H (1995) Fluid intelligence after frontal lobe lesions. Neuropsychologia 33:261-268.

Duncan J, Schramm M, Thompson R, Dumontheil I (2012) Task rules, working memory, and fluid intelligence. Psychonomic Bulletin and Review 19:864-870.

Duncan J, Chylinski D, Mitchell D, Bhandari A (2017) Complexity and compositionality in fluid intelligence. Proceedings of the National Academy of Sciences of the United States of America 114:5295-5299.

Duncan J, Emslie H, Williams P, Johnson R, Freer C (1996) Intelligence and the frontal lobe: The organization of goal-directed behavior. Cognitive Psychology 30:257-303.

Duncan J, Seitz RJ, Kolodny J, Bor D, Herzog H, Ahmed A, Newell FN, Emslie H (2000) A neural basis for general intelligence. Science 289:457-460.

Duncan J, Parr A, Woolgar A, Thompson R, Bright P, Cox S, Bishop S, Nimmo-Smith I (2008) Goal Neglect and Spearman's g: Competing Parts of a Complex Task. Journal of Experimental Psychology: General 137:131-148.

Duncker (1945) On problem-solving. Psychological Monographs 58.

Einstein GO, McDaniel MA (2010) Prospective memory and what costs do not reveal about retrieval processes: A commentary on Smith, Hunt, McVay, and McConnell (2007). Journal of Experimental Psychology: Learning, Memory, and Cognition 36:1082-1088.

Einstein GO, Smith RE, McDaniel MA, Shaw P (1997) Aging and prospective memory: The influence of increased task demands at encoding and retrieval. Psychology and Aging 12:479-488.

Einstein GO, Thomas R, Mayfield S, Shank H, McDaniel MA, Morrisette N, Breneiser J (2005) Multiple processes in prospective memory retrieval: Factors determining monitoring versus spontaneous retrieval. Journal of Experimental Psychology: General 134:327-342.

Ellis J, Freeman J (2008) Ten years on: realizing delayed intentions. In: Prospective memory: cognitive, neuroscience, developmental, and applied perspectives (Kliegel M, McDaniel MA, Einstein GO, eds). New York: Lawrence Erlbaum Associates.

Farooqui AA, Mitchell D, Thompson R, Duncan J (2012) Hierarchical organization of cognition reflected in distributed frontoparietal activity. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:17373-17381.

Fedorenko E, Duncan J, Kanwisher N (2013) Broad domain generality in focal regions of frontal and parietal cortex. Proceedings of the National Academy of Sciences of the United States of America 110:16616-16621.

Field A, Miles J, Field Z (2012) Discovering Statistics Using R: SAGE Publications.

Fougnie D, Marois R, xe (2006) Distinct Capacity Limits for Attention and Working Memory: Evidence from Attentive Tracking and Visual Working Memory Paradigms. Psychological Science 17:526-534.

Frank MJ, Badre D (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. Cerebral Cortex 22:509-526.

Freeman JE, Ellis JA (2003) Aging and the Accessibility of Performed and To-Be-Performed Actions. Aging, Neuropsychology, and Cognition 10:298-309.

Friedman NP, Miyake A (2017) Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex 86:186-204.

Gonen-Yaacovi G, Burgess PW (2012) Prospective memory: The future for future intentions. Psychologica Belgica 52:173-204.

Goschke T, Kuhl J (1993) Representation of intentions: Persisting activation in memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 19:1211-1226.

Graf M (2005) Orientation Congruency Effects for Familiar Objects: Coordinate Transformations in Object Recognition. Psychological Science 16:214-222.

Harrison TL, Mullet HG, Whiffen KN, Ousterhout H, Einstein GO (2014) Prospective memory: Effects of divided attention on spontaneous retrieval. Memory and Cognition 42:212-224.

Hebb DO (1940) Human Behaviour After Extensive Bilateral Removal from the Frontal Lobes Archives of Neurology And Psychiatry 44:421.

Henry JD, MacLeod MS, Phillips LH, Crawford JR (2004) A Meta-Analytic Review of Prospective Memory and Aging. Psychology and Aging 19:27-39.

Henson RNA (1996) Unchained Memory: Error Patterns Rule out Chaining Models of Immediate Serial Recall. The Quarterly Journal of Experimental Psychology Section A 49:80-115.

Hicks JL, Marsh RL, Russell EJ (2000) The properties of retention intervals and their affect on retaining prospective memories. Journal of Experimental Psychology: Learning, Memory, and Cognition 26:1160-1169.

Hynes SM, Fish J, Evans JJ, Manly T (2015) Developing a Computerised Multiple Elements Test for Organisational Difficulties. International Journal of Developmental Science 9:85-94.

IBM Corp. (2013) IBM SPSS Statistics for Windows. In, 22.0 Edition. Armonk, NY: IBM Corp.

Iveson MH, Della Sala S, Anderson M, MacPherson SE (2017) Goal neglect, fluid intelligence and processing speed: Manipulating instruction load and inter-stimulus interval. Acta Psychologica 176:1-10.

Jolicoeur P (1988) Mental rotation and the identification of disoriented objects. Canadian Journal of Psychology/Revue canadienne de psychologie 42:461-479.

Jost K, De Baene W, Koch I, Brass M (2013) A review of the role of cue processing in task switching. Zeitschrift fur Psychologie / Journal of Psychology 221:5-14.

Kahneman D (1973) Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.

Kane MJ, Engle RW (2003) Working-Memory Capacity and the Control of Attention: The Contributions of Goal Neglect, Response Competition, and Task Set to Stroop Interference. Journal of Experimental Psychology: General 132:47-70.

Kestenbaum R (1992) Feeling happy versus feeling good: The processing of discrete and global categories of emotional expressions by children and adults. Developmental Psychology 28:1132-1142.

Khanacadamy (2016) Information entropy. In: Journey into information theory (Khanacademy, ed): Khanacademy.

Kievit RA, Davis SW, Mitchell DJ, Taylor JR, Duncan J, Henson RNA (2014) Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. 5:1-10.

Kleiner M, Brainard D, Pelli D (2007) What's new in Psychtoolbox-3? Perception 36.

Kliegel M, Jäger T, Phillips LH (2008) Adult age differences in event-based prospective memory: A meta-analysis on the role of focal versus nonfocal cues. Psychology and Aging 23:203-208.

Kliegel M, Ballhausen N, Hering A, Ihle A, Schnitzspahn KM, Zuber S (2016) Prospective Memory in Older Adults: Where We Are Now and What Is Next. Gerontology 62:459-466.

Kvavilashvili L, Mandler G (2004) Out of one's mind: A study of involuntary semantic memories. Cognitive Psychology 48:47-94.

Kvavilashvili L, Kornbrot DE, Mash V, Cockburn J (2009) Differential effects of age on prospective and retrospective memory tasks in young, young-old, and old-old adults. Memory (Hove) 17:180-196.

Lagner P, Kliegel M, Phillips LH, Ihle A, Hering A, Ballhausen N, Schnitzspahn KM (2014) Mood effects on memory and executive control in a real-life situation. Cognition and Emotion:1-10.

Lee IA, Preacher KJ (2013) Calculation for the test of the difference between two dependent correlations with one variable in common. In.

Logan GD, Bundesen C (2004) Very clever homunculus: Compound stimulus strategies for the explicit task-cuing procedure. Psychonomic Bulletin and Review 11:832-840.

Lorenz K (1970) Studies in animal and human behaviour. London Methuen.

Lourenço O, Machado A (1996) In Defense of Piaget's Theory: A Reply to 10 Common Criticisms. In: (Bjork RA, ed), pp 143-164.

Lundqvist D, Flykt A, Öhman A (1998) The Karolinska Directed Emotional Faces - KDEF. In: Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.

Luria AR (1966). New York: New York : Basic Books.

Maier M, Steinhauser M (2013) Updating Expected Action Outcome in the Medial Frontal Cortex Involves an Evaluation of Error Type. Journal of Neuroscience 33:15705-15709.

Maier ME, Yeung N, Steinhauser M (2011) Error-related brain activity and adjustments of selective attention following errors. NeuroImage 56:2339-2347.

Manly T, Hawkins K, Evans J, Woldt K, Robertson IH (2002) Rehabilitation of executive function: Facilitation of effective goal management on complex tasks using periodic auditory alerts. Neuropsychologia 40:271-281.

Marcovitch S, Boseovski JJ, Knapp RJ, Kane MJ (2010) Goal Neglect and Working Memory Capacity in 4- to 6-Year-Old Children. Child Development 81:1687-1695.

Marsh RL, Hicks JL (1998) Event-based prospective memory and executive control of working memory. Journal of Experimental Psychology: Learning Memory and Cognition 24:336-349.

Marsh RL, Hancock TW, Hicks JL (2002) The demands of an ongoing activity influence the success of event-based prospective memory. Psychonomic Bulletin and Review 9:604-610.

Marsh RL, Hicks JL, Cook GI (2008) On beginning to understand the role of context in prospective memory. In: Prospective memory: cognitive, neuroscience, developmental, and applied perspectives (Kliegel M, McDaniel MA, Einstein GO, eds). New York: Lawrence Erlbaum Associates.

Marsh RL, Hicks JL, Cook GI, Hansen JS, Pallos AL (2003) Interference to Ongoing Activities Covaries with the Characteristics of an Event-Based Intention. Journal of Experimental Psychology: Learning Memory and Cognition 29:861-870.

Maylor EA (1993) Aging and forgetting in prospective and retrospective memory tasks. Psychology and Aging 8:420-428.

Maylor EA (1996) Age-related impairment in an event-based prospective-memory task. Psychology and Aging 11:74-78.

Mayr U, Kliegl R (2000) Task-Set Switching and Long-Term Memory Retrieval. Journal of Experimental Psychology: Learning Memory and Cognition 26:1124-1140.

McDaniel MA, Einstein GO (2000) Strategic and Automatic Processes in Prospective Memory Retrieval: A Multiprocess Framework. Applied Cognitive Psychology 14:S127-S144.

McDaniel MA, Scullin MK (2010) Implementation intention encoding does not automatize prospective memory responding. Memory & Cognition 38:221-232.

McDaniel MA, Einstein GO, Guynn MJ, Breneiser J (2004) Cue-focused and reflexive-associative processes in prospective memory retrieval. Journal of Experimental Psychology: Learning Memory and Cognition 30:605-614.

Meier B, Zimmermann TD (2015) Loads and loads and loads: The influence of prospective load, retrospective load, and ongoing task load in prospective memory. Frontiers in human neuroscience 9.

Microsoft (2012) Microsoft Office PowerPoint 2013. In. Seattle, WA: Microsoft Corporation.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. In, pp 167-202.

Milner B (1963) Effects of different brain lesions on card sorting. Archives of Neurology 9:90-100.

Miyake A, Emerson MJ, Padilla F, Ahn JC (2004) Inner speech as a retrieval aid for task goals: The effects of cue type and articulatory suppression in the random task cuing paradigm. Acta Psychologica 115:123-142.

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD (2000) The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. Cognitive Psychology 41:49-100.

Morey CC, Elliott EM, Wiggers J, Eaves SD, Shelton JT, Mall JT (2012) Goal-neglect links Stroop interference with working memory capacity. Acta Psychologica 141:250-260.

Morris RG, Miotto EC, Feigenbaum JD, Bullock P, Polkey CE (1997) The effect of goal-subgoal conflict on planning ability after frontal- and temporal-lobe lesions in humans. Neuropsychologia 35:1147-1157.

Moscovitch M (1994) Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. Neuropsychology 8:524-534.

Norman DA, Shallice T (1986) Attention to action. In: Consciousness and self regulation (Davidson RJ, Schwartz GE, Shapiro D, eds). New York: Plenum Press.

Park DC, Hertzog C, Kidder DP, Morrell RW, Mayhorn CB (1997) Effect of age on event-based and time-based prospective memory. Psychology and Aging 12:314-327.

Pashler H (1994) Dual-task interference in simple tasks: Data and theory. Psychological Bulletin 116:220-244.

Pereira A, Ellis J, Freeman J (2012) Is prospective memory enhanced by cue-action semantic relatedness and enactment at encoding? Consciousness and Cognition 21:1257-1266.

Pessoa L (2009) How do emotion and motivation direct executive control? Trends in Cognitive Sciences 13:160-166.

Poldrack Russell A (2016) Chaucer Club talk at the MRC Cognition and Brain Sciences Unit, Cambridge, UK. In.

Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, Parker DS, Sabb FW, Bilder RM (2011) The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. Frontiers in Neuroinformatics 5.

Popov V, Reder M (2016) Not knowing what we know: A call for a theory-neutral database for empirical results in psychology. In. Talk presented at the Workshop on Memory and Skill, Duke University, Durham, NC http://venpopov.com/presentations/duke-presentation-final.pdf.

Prabhakaran V, Smith JAL, Desmond JE, Glover GH, Gabrieli JDE (1997) Neural Substrates of Fluid Reasoning: An fMRI Study of Neocortical Activation during Performance of the Raven's Progressive Matrices Test. Cognitive Psychology 33:43-63.

Preacher KJ (2002) Calculation for the test of the difference between two independent correlation coefficients. In. http://quantpsy.org.

R Core Team (2013) R: A language and environment for statistical computing. In, ISBN 3-900051-07-0 Edition. Vienna, Austria: R Foundation for Statistical Computing.

Reason J (2013) A life in error: From little slips to big disasters. Boca Raton, FL: CRC Press.

Reverberi C, Lavaroni A, Gigli GL, Skrap M, Shallice T (2005) Specific impairments of rule induction in different frontal lobe subgroups. Neuropsychologia 43:460-472.

Risko EF (2015) Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. Consciousness and Cognition: An International Journal 36:61-75.

Risko EF, Gilbert SJ (2016) Cognitive Offloading. Trends in Cognitive Sciences 20:676-688.

Roberts G, Anderson M (2014) Task structure complexity and goal neglect in typically developing children. Journal of Experimental Child Psychology 120:59-72.

Robertson IH, Manly T, Andrade J, Baddeley BT, Yiend J (1997) `Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 35:747-758.

Robinson G, Shallice T, Bozzali M, Cipolotti L (2012) The differing roles of the frontal cortex in fluency tests. Brain 135:2202-2214.

Roca M, Torralva T, Gleichgerrcht E, Woolgar A, Thompson R, Duncan J, Manes F (2011) The role of Area 10 (BA10) in human multitasking and in social cognition: A lesion study. Neuropsychologia 49:3525-3531.

Roca M, Manes F, Cetkovich M, Bruno D, Ibáñez A, Torralva T, Duncan J (2014) The relationship between executive functions and fluid intelligence in schizophrenia. Frontiers in Behavioral Neuroscience 8.

Roca M, Parr A, Thompson R, Woolgar A, Torralva T, Antoun N, Manes F, Duncan J (2010) Executive function and fluid intelligence after frontal lobe lesions. Brain 133:234-247.

Rogers RD, Monsell S (1995) Costs of a Predictable Switch Between Simple Cognitive Tasks. Journal of Experimental Psychology: General 124:207-231.

Salthouse TA, Berish DE, Siedlecki KL (2004) Construct validity and age sensitivity of prospective memory. Memory and Cognition 32:1133-1148.

Salthouse TA, Pink JE, Tucker-Drob EM (2008) Contextual analysis of fluid intelligence. Intelligence 36:464-486.

Schneider W, Shiffrin RM (1977) Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review 84:1-66.

Schnitzspahn KM, Ihle A, Henry JD, Rendell PG, Kliegel M (2011) The age-prospective memory-paradox: An exploration of possible mechanisms. International Psychogeriatrics 23:583-592.

Scullin MK, McDaniel MA, Shelton JT, Lee JH (2010) Focal/Nonfocal Cue Effects in Prospective Memory: Monitoring Difficulty or Different Retrieval Processes? Journal of Experimental Psychology: Learning Memory and Cognition 36:736-749.

Shallice T (1982) Specific Impairments of Planning. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences (1934-1990) 298:199-209.

Shallice T, Burgess PW (1991) Deficits in strategy application following frontal lobe damage in man. Brain 114:727.

Shapiro KL, Raymond JE, Arnell KM (1997) The attentional blink. Trends in Cognitive Sciences 1:291-296.

Smith RE (2003) The cost of remembering to remember in event-based prospective memory: Investigating the capacity demands of delayed intention performance. Journal of Experimental Psychology: Learning, Memory, and Cognition 29:347-361.

Smith RE (2008) Connecting the past and the future. Attention, memory, and delayed intentions. In: Prospective memory: cognitive, neuroscience, developmental, and applied perspectives (Kliegel M, McDaniel MA, Einstein GO, eds). New York: Lawrence Erlbaum Associates.

Smith RE, Hunt RR, McVay JC, McConnell MD (2007) The Cost of Event-Based Prospective Memory: Salient Target Events. Journal of Experimental Psychology: Learning Memory and Cognition 33:734-746.

Spearman C (1904) "General Intelligence," Objectively Determined and Measured. The American Journal of Psychology 15:201-292.

Sprague Thomas C, Ester Edward F, Serences John T (2016) Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91:694-707.

Stokes Mark G, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron 78:364-375.

Stroop JR (1935) Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18:643–662.

Stuss DT, Levine B, Alexander MP, Hong J, Palumbo C, Hamer L, Murphy KJ, Izukawa D (2000) Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia 38:388-402.

Szűcs D, Soltész F, Bryce D, Whitebread D (2009) Real-time Tracking of Motor Response Activation and Response Competition in a Stroop Task in Young Children: A Lateralized Readiness Potential Study. Journal of Cognitive Neuroscience 21:2195-2206.

Towse JN, Lewis C, Knowles M (2007) When knowledge is not enough: The phenomenon of goal neglect in preschool children. Journal of Experimental Child Psychology 96:320-332.

Tulving E, Thomson DM (1973) Encoding specificity and retrieval processes in episodic memory. Psychological Review 80:352-373.

Unsworth N, McMillan BD (2014) Fluctuations in pre-trial attentional state and their influence on goal neglect. Consciousness and Cognition 26:90-96.

Ward J (2015) The student's guide to cognitive neuroscience, Third edition.. Edition: London : Psychology Press, Taylor & Francis Group.

Weiskopf DA (2011) Models and mechanisms in psychological explanation. Synthese 183:313.

West R (2008) The cognitive neuroscience of prospective memory. In: Prospective memory: cognitive, neuroscience, developmental, and applied perspectives (Kliegel M, McDaniel MA, Einstein GO, eds). New York: Lawrence Erlbaum Associates.

West R, Craik FIM (1999) Age-related decline in prospective memory: The roles of cue accessibility and cue sensitivity. Psychology and Aging 14:264-272.

West R, Alain C (2000) Evidence for the Transient Nature of a Neural System Supporting Goal-directed Action. Cerebral Cortex 10:748-752.

Westbrook A, Martins B, Yarkoni T, Braver T (2012) Strategic Insight and Age-Related Goal-Neglect Influence Risky Decision-Making. Frontiers in Neuroscience 6.

Wilson AD, Golonka S (2016) Mechanisms and Models of Mechanisms. In: Notes from Two Scientific Psychologists.

Wilson B, Alderman N, Burgess P, Emslie H, Evans J (1996) Behavioural Assessment of the Dysexecutive Syndrome. In: (Assessment OP, ed).

Wimber M, Alink A, Charest I, Kriegeskorte N, Anderson MC (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. Nat Neurosci 18:582-589.

Woolgar A, Hampshire A, Thompson R, Duncan J (2011) Adaptive coding of task-relevant information in human frontoparietal cortex. Journal of Neuroscience 31:14592-14599.

Woolgar A, Parr A, Cusack R, Thompson R, Nimmo-Smith I, Torralva T, Roca M, Antoun N, Manes F, Duncan J (2010) Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. Proceedings of the National Academy of Sciences of the United States of America 107:14899-14902.

Zeintl M, Kliegel M, Hofer SM (2007) The role of processing resources in age-related prospective and retrospective memory within old age. Psychology and Aging 22:826-834.

Zelazo PD (2004) The development of conscious control in childhood. Trends in Cognitive Sciences 8:12-17.

Zelazo PD, Frye D, Rapus T (1996) An Age-Related Dissociation Between Knowing Rules and Using Them. Cognitive Development 11:37-63.

Zhang Y, Bradlow ET (2016) Entropy-like measure of clumpiness In: (Biondo F, ed). -personal communication-.

Zhang Y, Bradlow ET, Small DS (2013) New measures of clumpiness for incidence data. Journal of Applied Statistics 40:2533-2548.

APPENDIX

Instructions to the LMT (Nontransparent version) as delivered by the experimenter (see Chapter 2).
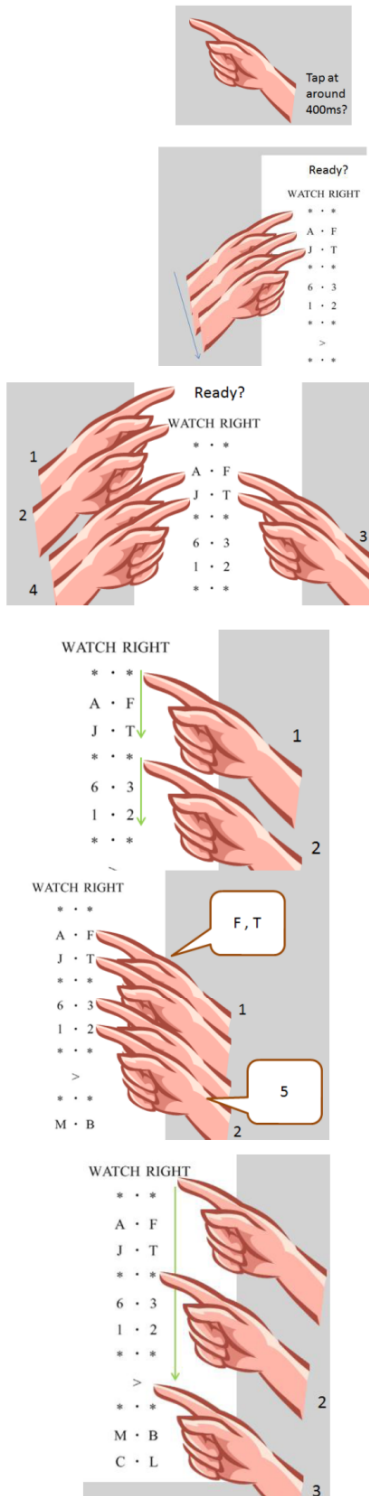
Indirect

In this task you will see pairs of symbols – asterisks, letters and numbers – coming up, one after the other, in the middle of the screen. (POINT TO SCREEN AND TAP AT APPROPRIATE RATE.)

Here is a written example to give you the idea (SHOW EXAMPLE).
Imagine these pairs coming up, one after the other (POINT AND TAP).

1. As you see, the run begins with the word READY (POINT) in the middle of the screen,

2. followed by an instruction that is flashed up WATCH LEFT or WATCH RIGHT. This tells you which side to watch for the symbols that follow.

3. So here you watch right and see **F, T** and so on (POINT)

4. At the same time these (POINT) will be coming up on the left, but you ignore these.

1. The symbols on the side you're watching are always organised in blocks of 3, an asterisk followed by 2 letters (POINT TO F AND T) or an asterisk followed by 2 numbers (POINT TO 3 AND 2).

1. When you see <u>letters</u>, just repeat them out loud. So here (POINT) you would say **"F, T".**

2. when you see <u>numbers</u> add the 2 together from the side you are watching and say the answer aloud. So here (POINT) you would see 3 and 2 and answer "5". You should <u>not</u> repeat the numbers themselves, as you do for the letters – that will count as a mistake.

1. <u>Always remember just to watch the side you were told at the start, and to ignore the other side (POINT).</u>

2. The asterisks just mark the spaces between letters and numbers – you don't need to say anything for them.

3. There is a final part to this. The sequence of letters and numbers lasts about 5 seconds. Towards the end (POINT), you see a symbol which <u>may</u> tell you to switch sides, from right to left or from left to right. It works like this.
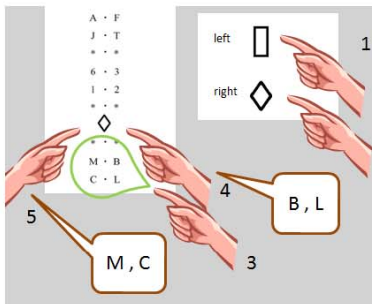
Indirect

1. The symbol is a shape which you see in the middle of the screen. The shape is either a rectangle or a diamond (SHOW EXAMPLE OF BOTH ON SHEET OF PAPER)

2. The shapes show which side to watch for the last part of the run (SHOW END OF RUN).

3. In the example (POINT) you see a diamond, which means 'watch right', so in this case you keep watching the right side and repeat **B** and **L**.

4. but if this had been a rectangle, which means 'watch left', you would have switched at once to the left (POINT) and repeated **M** and **C**.

5. To help you remember these, I'll leave them on the table like this (PLACE).

PRACTICE! Until they utter a stimulus, even if incorrect.

Is that clear? Let's try a practice run.

You will see that you are told at the beginning to watch the <u>left</u>.

Remember: repeat letters and add numbers.

The whole thing will be fairly quick, so don't be surprised at that.

Are you ready?

On-screen instructions to the Word Categorization Task (see Chapter 5.2.1)

Introduction of Ongoing Task + Practice 1

1

Welcome!

When you are ready to begin,
press any key to continue.

2

In this experiment you will use a total of 4 keys:

1. ENTER
2. Y (Yes key)
3. N (No key)
4. Q

Please visually identify each of the keys listed above - they are marked with white stickers. You can use either hand to press the keyboard, but please rest your LEFT hand on the table and NOT on the keyboard.

Please ask the experimenter if you have any questions. Otherwise, press the Y key to proceed

3

In this experiment, we are primarily interested in how individuals categorize words. One of your tasks in this experiment is to determine whether a word fits into a category or not.

Press ENTER to continue.

4

Two words will be presented on the screen.

Your task is to determine whether the word in lowercase letters on the left side of the screen fits into the category represented by the word in capitalized letters on the right side of the screen.

If the word is a member of the given category, press the Y key. If the word is not a member of the category, press the N key.

Press ENTER to continue.

5

For example:

brown   COLOUR

The correct answer is Y because brown is a colour.

If you have any questions, you may ask the experimenter now. Otherwise, press ENTER to move on to a few practice trials.

6

The task is loading...
Be ready!

Practice 1
(3 trials; ongoing only)

Practice 2 with Feedback (ongoing only) + noPM mini block1

7

Well done!

Getting the correct answer and a quick response are equally important. It is critical that you respond to the word as fast as you can without sacrificing accuracy.

Next are more practice trials that will give you feedback on how quickly and accurately you respond and, your overall performance.

If you have any questions, you may ask the experimenter now.

Otherwise press ENTER to begin this practice with feedback.

8

The task is loading...
Be ready!

Practice 2
(11 trials; ongoing only
with feedback)

9

Total Correct:

Press ENTER to continue

10

Good job!

Next is the real task.

Before you start, to ensure that you understand this task, please summarize the instructions you have received aloud to the experimenter.

Free Recall 1
Ongoing task

11

You are now ready to begin the real task.

Please locate the N key and press it to begin.

12

The task is loading...
Be ready!

noPM block 1
(30 trials; ongoing only)

197

PM instructions (+extra rule) + Practice

13

Very good!

In this experiment we have an additional, but secondary, interest. This is your ability to remember to perform an action at a given point in the future.

During the word-category task, we would like you to perform a special action whenever you see a particular string of letters.

This string of letters is "tor", in this order.

Press ENTER for more instructions

14

Whenever you see "tor" anywhere, press the Q key. For example, you may see "tor" in the middle of a word such as in "motorbike". You may press the Q key before or after responding with Y or N. If you forget to press Q right away, you may press Q as soon as you remember.

Do not forget that your primary goal throughout this experiment is to respond to the word-category task.

Press ENTER to continue.

15

There is another special action for you to perform and this is also of secondary interest.

Some words may be coloured in green ink. Whenever you see a green word which is also in lowercase letters, please read it out aloud. You may read it before, during or after responding with the appropriate keypress.

If you forget to say the word out aloud, you may do so as soon as you remember. Do NOT read green words in capitalized letters.

Press ENTER to continue.

16

Next is some practice. This time remember to press the Q key if you see "tor" and also remember to read words aloud if they are in green ink and in lowercase letters.

You will not be given feedback on accuracy and speed this time but remember that your primary goal is to respond as quickly and as accurately as you can to each trial.

Press ENTER to continue.

17

To ensure that you understand this task, please summarize the instructions aloud to the experimenter.

Free Recall 2

- Ongoing task
- 'tor'
- Green rule (if applicable)

18

You are now ready to begin the final practice trials.

Please locate the Q key and press it to begin.

The task is loading... Be ready!

Practice 3
(11trials, 1PM target,2 green*)

PM reminders, drop extra rule

19

This is the end of the practice session.

Did you make your decision, pressing Y or N as fast as you could?
Did you remember to press the Q key when "tor" appeared?
Did you remember to read the words aloud if they appeared in green and in lowercase?
Do you understand the instructions up to this point?

If you have any questions please ask the experimenter.

Otherwise press ENTER to continue

20

A few reminders:

If you forget to press Q when you see "tor", but remember it on a later trial, you may press Q when you remember. Similarly, if you forget to read the words in green ink and lowercase letters, you can do so when you remember.

Also, whenever you press Q, you may not have the opportunity to do the word-category task (by pressing Y or N). Do not worry about this, it will not count as a mistake. We will only score the trials in which you respond to the word-category task.

Press ENTER to continue.

21

Next you will be doing the real task which is a longer version of the practice you just completed.

Because this task is long, please try to maintain your focus, and ensure that you are seated comfortably.

Do not forget that your primary goal throughout this experiment is to respond to the word-category task as quickly and as accurately as possible.

This is your last chance to ask questions. Once the experiment begins, the experimenter will not answer any questions. Press ENTER to continue.

22

One final thing before we begin.

For the time being, the task will be a bit simpler.

All the words will always appear in black ink, never in green - so for now, you can forget about the green ink rule. In other words, until the experimenter tells you otherwise, you will not need to ever read words aloud.

Is that clear? Please tell the experimenter.

23

Press Y to start the task!

24

The task is loading...
Be ready!

PM
(164 trials; 4 PM targets)

noPM mini-block 2 & Final block with all rules

25

Well done!!

Before proceeding to doing more of this task,
please summarize the instructions to the experimenter

Free Recall 3
- Ongoing task
- 'tor'
- (ask them if remember inactive green rule, if applicable.)

26

Next you will be doing more of the word-categorization task, but this time the string of letters "tor" will not appear anywhere, so you do not need to remember to press the Q key at any point.

Your only task is to press Y if the word on the left is a member of the category on the right and N if it is not.

Press the ENTER to continue.

27

The task is not as long as before, but still try to maintain your focus, making sure your responses are quick and accurate.

Before you start, please summarize the instructions to the experimenter.

Just a confirmation that they've dropped the 'tor' intention

28

Press N to start the task!

The task is loading...
Be ready!

noPM block 2
(30 trials; ongoing only)

29

Very good! We have now reached the final part of the experiment.

In this part you will be doing more of the word-categorization task but now all the rules are active. What follows is a reminder of these rules. You will need to press Y or N if the word on the left is a member of the category on the right. You will also need to remember to press Q whenever you see "tor". And finally, you will need to remember to read words aloud if they appear in green ink and are in lowercase letters.

If you have any questions please ask the experimenter now. Otherwise press ENTER to start the final part of this experiment.

The task is loading...
Be ready!

30

Great!

This is the end of this task.

Final block
(11 trials; 1PM,2green*)