The Psychology of Learning and Motivation

Volume 59

THE PSYCHOLOGY OF **LEARNING AND MOTIVATION**

Series Editor

BRIAN H. ROSS
Beckman Institute and Department of Psychology
University of Illinois, Urbana, Illinois

VOLUME FIFTY NINE

THE PSYCHOLOGY OF LEARNING AND MOTIVATION

Edited by

BRIAN H. ROSS
Beckman Institute and Department of Psychology
University of Illinois, Urbana, Illinois

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK
32 Jamestown Road, London NW1 7BY, UK

ISBN: 978-0-12-407187-2
ISSN: 0079-7421

Printed and bound in USA
13 14 15 10 9 8 7 6 5 4 3 2 1

CONTENTS

This page intentionally left blank

CONTRIBUTORS

Branden Abushanab
Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, USA

Simona Buetti
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Rick Dale
Cognitive and Information Sciences, University of California Merced, Merced, CA, USA

Peter F. Delaney
Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, USA

Nicholas D. Duran
Cognitive and Information Sciences, University of California Merced, Merced, CA, USA; Division of Psychology and Language Sciences, University College London, London, UK

Nathaniel L. Foster
Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, USA

Andrew J. Furman
Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

Riccardo Fusaroli
Cognitive and Information Sciences, University of California Merced, Merced, CA, USA; Interacting Minds Center and Center for Semiotics, Aarhus University, Aarhus, Denmark

Christina L. Gagné
Department of Psychology, University of Alberta, Edmonton, AB, Canada

Stephen D. Goldinger
Department of Psychology, Arizona State University, Tempe, AZ, USA

P. N. Johnson-Laird
Department of Psychology, Princeton University, Princeton, NJ, USA; Department of Psychology, New York University, New York, NY, USA

Sangeet S. Khemlani
Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA

Alejandro Lleras
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Steven A. Marchette
Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

J. Toby Mordkoff
Department of Psychology, University of Iowa, Iowa City, IA, USA

Megan H. Papesh
Department of Psychology, Louisiana State University, Baton Rouge, LA, USA

Daniel C. Richardson
Division of Psychology and Language Sciences, University College London, London, UK

Lili Sahakyan
Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, USA

Amy L. Shelton
Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

Thomas L. Spalding
Department of Psychology, University of Alberta, Edmonton, AB, Canada

CHAPTER ONE

Toward a Unified Theory of Reasoning

P. N. Johnson-Laird[*,†,1], **Sangeet S. Khemlani**[‡]
[*]Department of Psychology, Princeton University, Princeton, NJ, USA
[†]Department of Psychology, New York University, New York, NY, USA
[‡]Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA
[1]Corresponding author: E-mail: phil@princeton.edu

Contents

Abstract

This article describes a theory that uses mental models to integrate deductive, inductive, and probabilistic reasoning. It spells out the main principles of the theory and illustrates them with examples from various domains. It shows how models underlie inductions, explanations, estimates of probabilities, and informal algorithms. In all these cases, a central principle is that the mind represents each sort of possibility in a separate mental model and infers whatever holds in the resulting set of models. Finally, the article reviews what has been accomplished in implementing the theory in a single large-scale computer program, *mReasoner*.

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00001-0

1. INTRODUCTION

The capacity to reason underlies mathematics, science, and technology. It is essential for coping with everyday problems—without it, social life would be almost unimaginable. The challenge to psychologists is to explain its underlying mental mechanisms. Since Störring's (1908) pioneering study, they have discovered several robust phenomena. Perhaps the most important is that naive reasoners—those with no training in logic—can make *valid* deductions, that is, inferences in which the conclusion is true in all the cases in which the premises are true (cf. Jeffrey, 1981, p. 1). And they are happy to do so about abstract matters with no ecological validity, as in Sudoku puzzles (Lee, Goodwin, & Johnson-Laird, 2008).

Fifty years ago, psychologists took for granted that human reasoning was rational. Individuals developed deductive competence during childhood, and the psychologists' task was to pin down the nature of the formal logic underlying this ability. As Inhelder and Piaget (1958, p. 305) wrote, "Reasoning is nothing more than the propositional calculus itself". There may be vagaries in performance, but faulty reasoning does not occur (Henle, 1962) or is attributable to local malfunctions in the system—a spanner in the works rather than an intrinsic flaw (Cohen, 1981). Indeed, theories of deduction in cognitive psychology began with accounts based on formal logic (e.g. Braine, 1978; Johnson-Laird, 1975; Osherson, 1974–1976). These views, of course, echo those of Enlightenment philosophers. But, another robust phenomenon that psychologists discovered is that individuals differ in their ability to reason. A few are very good, a few are very bad, and most are somewhere in the middle. Differences in ability are vast, and correlate with the tests of academic achievement, as proxies for measures of intelligence (Stanovich, 1999). For everyone, however, failures are inevitable: complex inferences are computationally intractable.

Nowadays, a consensus exists that that the psychology of reasoning has undergone a deep change—even, some say, a paradigm shift. The accessibility of digital computers was a license for theorizing, and psychological theories of reasoning have multiplied at a startling rate. Quite what the new foundations of reasoning should be is controversial. One view is that humans are rational but in terms of the probability calculus rather than logic (e.g. Oaksford & Chater, 2007; Tenenbaum & Griffiths, 2001). One view is that natural selection has equipped the mind with modules for reasoning about special topics, such as social exchange (Cosmides & Tooby, 2005). One view is that rationality presupposes a normative system, and psychologists should

abandon norms in favor of descriptions (Evans, 2012). One revenant is that logic, or logics, provides the inferential machinery (Rips, 1994; Stenning & van Lambalgen, 2008). We will not try to assess these views, but for those who espouse such a theory, we recommend the answers to two questions as a recipe for resipiscence: Has the theory been implemented in a computer program, and does it predict most of the 60 or more experimental results reported here? The goal of the article, however, is not polemical, but to describe a different theory.

Craik (1943) postulated that thinking was based on making mental simulations of the world to anticipate events. This idea in turn has historical antecedents, although Craik was unlikely to have known them (see Johnson-Laird, 2004, for the history of mental models). Oddly, however, Craik did not consider reasoning, other than to make a casual remark that it was based on "verbal rules" (Craik, 1943, p. 81). In the spirit of Craik, we argue that the mind is neither a logical nor a probabilistic device, but instead a device that makes mental simulations. Insofar as humans reason logically or infer probabilities they rely on their ability to simulate the world in mental models. The application of simulation to reasoning is based on mental models of the possibilities to which the premises refer, and a valid deduction has a conclusion that holds in all these models. This idea was first proposed a generation ago (Johnson-Laird, 1975). Since then, its proponents and critics have revised and extended it in hundreds of publications.

The theory of mental models—the model theory for short—is controversial, as are all current theories of reasoning, and the only way to put it beyond controversy calls for two crucial steps. The first step follows Leibniz (1685, 1952), who dreamt of replacing argument with calculation. It is to implement a unified theory of reasoning in a computer program that, for any inferential task, outputs the responses that human reasoners should make, the respective likelihoods and latencies of these responses, the processes underlying them, and, where relevant, valid or ideal responses. The second step is to show in stringent experiments that the program's predictions are correct. We are a long way from the two steps. Researchers have applied the model theory to many sorts of inferential task, implemented computational models of these applications, and tested the theory experimentally. But, until recently, the work has been piecemeal rather than unified.

What are the essentials of the model theory, and what counts as a mental model? This article is going to answer these questions step by step, and it aims above all to enable readers to understand the model theory without having to read anything else. It illustrates the theory's application to most sorts of reasoning. Its plan mirrors these aims. It begins with an outline of the

main sorts of reasoning. It follows with sections that elucidate each of the theory's main principles. It then considers the role of models in inductive reasoning, explanatory reasoning, reasoning about probabilities, and reasoning that yields informal algorithms. The final section of the article reviews what has been accomplished in unifying the theory, and in implementing it in a single computer program to achieve Leibniz's (and our) dream.

2. WHAT IS REASONING?

Suppose you infer:

If the ink cartridge is empty then the printer won't work.
The ink cartridge is empty.
So, the printer won't work.

You are making a deduction: your inference is valid because your conclusion holds in any case in which the premises hold. Suppose instead you infer:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
So, the ink cartridge is empty.

You are making an induction. Your inference isn't valid because there may be another reason that the printer won't work. Yet, your conclusion may be true, especially if the printer is producing blank pages. For many theorists—Aristotle for one, all inferences fall into one of these two categories: deduction and induction.

Aristotle defined induction as an inference from a particular assertion to a universal one (*Topics*, 105a13). But, inductions are often from the particular to the particular, as is your induction about the printer. Hence, a better way to distinguish between the two sorts of inference is in terms of semantic information (Johnson-Laird, 1983, chap. 2). The more possibilities that an assertion rules out, the more information it conveys (Bar-Hillel, 1964). An inference to a conclusion that refers to all the same possibilities as the premises do, or at least includes them all in what it refers to, is a *deduction*. Consider again your earlier deduction:

If the ink cartridge is empty then the printer won't work.
The ink cartridge is empty.
So, the printer won't work.

The premises refer to just one possibility:

The ink cartridge is empty and the printer won't work.

Hence, your inference is valid because its conclusion holds in the one possibility to which the premises refer. The conclusion therefore does not

increase information. But semantics should not be confused with epistemology: a conclusion may be news to the person who draws it, bringing to mind a novel proposition. An inference to a conclusion that refers to only some of the possibilities to which the premises refer, although it may add some new possibilities too, is an *induction*. Consider again your earlier induction:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
So, the ink cartridge is empty.

The premises refer to two possibilities:

The ink cartridge is empty. The printer won't work.
The ink cartridge isn't empty. The printer won't work.

Your conclusion, however, refers to only one of these two possibilities. It goes beyond the information in the premises, and it is consistent with them, that is, it is possible that the ink cartridge is empty. But, the conclusion does not follow validly. A special case of induction is one that also introduces new ideas to explain something, and this sort of reasoning is known as *abduction*, for example:

If the ink cartridge is empty then the printer won't work.
The printer won't work.
Hence, there's a fault in the connection between the computer and the printer.

Inferences either maintain or throw information away (deductions) or they increase information (inductions). One other relation between the premises and conclusion remains. If they refer to disjoint possibilities, they contradict one another. In logic, any conclusion whatsoever follows validly from a contradiction: the premises don't refer to any possibility in which the conclusion fails to hold because the premises don't refer to any possibility. Naive reasoners, however, reject inferences from self-contradictions. For them, the definition of validity has a rider: a valid inference is one in which the conclusion holds in every possibility to which the premises refer, and there is at least one such possibility.

Reasoners aim to draw conclusions that are true, or at least plausible. But, they also aim to draw novel and parsimonious conclusions, and so they would feel silly just to form a conjunction of all the premises even though such an inference is valid. They know more when they know:

It's raining

than when they know:

It's raining or it's cold, or both.

Yet, the disjunctive conclusion follows validly from the categorical premise. Hence, not all valid deductions are sensible, and it would be silly to make this particular inference because it throws away information by adding a disjunctive

alternative to a premise. So, a theory of deductive competence—of what the inferential system computes—assumes that individuals have the potential to be rational and an awareness of this potential. They abide by the foregoing constraints. In sum: "To deduce is to maintain semantic information, to simplify, and to reach a new conclusion" (Johnson-Laird & Byrne, 1991, p. 22). These constraints are not easy to embody in a theory based on formal logic, and this difficulty explains why such theories, like automated theorem provers in artificial intelligence, focus on the evaluation of *given* conclusions. Inductive competence also aims for parsimony and novelty, but it goes beyond the information given and ultimately aims to explain phenomena.

3. MODELS OF POSSIBILITIES

The fundamental assumption of the model theory is that each mental model represents what is common to a distinct set of possibilities. Hence, an assertion such as:

A triangle is on the right of a circle

has a single mental model, which we depict in this diagram:

The left-to-right axis of the model corresponds to the left-to-right axis of a scene, and the disposition of the triangle and circle in the model corresponds to their disposition in a scene for which the assertion is true. The model represents an indefinite number of possibilities that have in common only that a triangle is on the right of a circle. Of course, the relative sizes of the figures in the model, their distance apart, and so on, play no role in reasoning from the model, but we defer an explanation of how their irrelevance is represented until Section 11.

Everyone prefers to think about just one possibility at a time. Intuitions work in this way. And the theory postulates two separate systems for reasoning, one for intuitions and one for deliberations—a familiar distinction in "dual process" theories of reasoning (see, e.g. Evans, 2008; Kahneman, 2011; Reitman, 1965; Sloman, 1996; Stanovich, 1999; Verschueren, Schaeken, & d'Ydewalle, 2005). The model theory distinguishes between the two systems in computational power, and we have implemented both of them in computer programs (Khemlani & Johnson-Laird, 2012a; Khemlani, Lotstein, & Johnson-Laird, 2012). The intuitive system, which is sometimes known as "system 1", has no access to working memory, and so it can represent only one mental model at a time (Johnson-Laird, 1983, chap. 6), and it

cannot carry out recursive processes, including arithmetical operations such as counting. It lacks even the computational power of a finite-state automaton (Hopcroft & Ullman, 1979) because it can carry out a loop of operations for only a small finite number of times—a restriction that is built into its computer implementation. In contrast, the deliberative system, which is sometimes known as "system 2", has access to working memory, and so it can search for alternative mental models, and carry out recursive processes, such as counting and arithmetical operations, until they overwhelm its processing capacity.

One way in which to overwhelm the deliberative system is to force it to reason about disjunctions. An inclusive disjunction, such as:

There's a triangle or there's a circle, or both

includes the joint possibility of both the triangle and the circle, and so it refers to three sorts of possibility. It therefore calls for three mental models, which we show on separate rows in this diagram:

In this case, spatial relations play no role in the use of the models. An exclusive disjunction, such as:

Either there's a triangle or there's a circle, but not both

exclude the joint possibility, and so it calls for only two mental models:

Models preoccupy system 2, and so more models mean more work. The theory therefore predicts that deductions from exclusive disjunctions should be easier than those from inclusive disjunctions, as when either of the disjunctions above occurs with the categorical assertion:

There isn't a circle.

This assertion eliminates any model in which there is a circle, and so it follows validly in both cases that:

There is a triangle.

Evidence corroborates the prediction (e.g. Johnson-Laird, Byrne, & Schaeken, 1992), and it also shows that inferences from conjunctions, which have just one model, are easier than those based on disjunctions (García-Madruga, Moreno, Carriedo, Gutiérrez, & Johnson-Laird, 2001).

The sorts of inference that can overwhelm the deliberative system are "double disjunctions" (Johnson-Laird et al., 1992), which are from pairs of disjunctive premises, such as:

June is in Wales, or Charles is in Scotland, but not both.

Charles is in Scotland, or Kate is in Ireland, but not both.

What follows?

The two possibilities compatible with the first premise are relatively easy to envisage, but it is difficult to update them with those from the second premise, although the result is just two possibilities:

June in Wales Kate in Ireland

Charles in Scotland

Of course, real mental models represent these spatial relations, and are not phrases, which we use here for convenience. The two models yield the conclusion:

Either June is in Wales and Kate is in Ireland or Charles is in Scotland.

Inferences become even harder when disjunctions are inclusive. In one experiment, 25% of the participants, who were from the general public, drew valid conclusions from exclusive disjunctions, but this figure fell to below 10% for inclusive disjunctions. The result is hardly surprising, but what was striking was the nature of the modal errors: for all the inferences, the participants drew conclusions corresponding to a model of a single possibility. Just under a third of all the participants' responses were conclusions of this sort. The result suggests that when the task was too much for them, they fell back on their intuitions and envisaged just a single model of the premises. So, their conclusions were consistent with the premises but did not follow from them. The performance of undergraduates showed the same pattern. But, when the disjunctions were presented in equivalent electrical circuit diagrams or analogs of them, they performed better and faster (Bauer & Johnson-Laird, 1993). Their conclusions, however, still bore out a failure to consider all the possibilities, and so most errors were at least consistent with the premises.

Mental models represent possibilities, and so the more models that are necessary to make an inference, the harder that inference is to make. Individuals are in danger of overlooking a model. When the deliberative system is vastly overburdened, reasoners may even fall back on the intuitive system and draw a conclusion that is consistent with only a single model. One side effect of the use of models is that reasoners are most unlikely to draw

conclusions that throw semantic information away by adding disjunctive alternatives.

4. ICONS AND SYMBOLS

Mental models are iconic insofar as possible. What "iconic" means is that their structure corresponds to the structure of what they represent (see Peirce, 1931–1958, Vol. 4, paragraph 447). One example is the mental model of the assertion, *the triangle is on the right of the circle*, which we diagrammed in the previous section. Another example is an electrical circuit diagram with the same structure as the circuit it denotes. The great advantage of an icon, as Peirce realized, is that its inspection yields new information. Given the premises:

The triangle is on the right of the circle.

The square is on the right of the triangle.

The intuitive system can build the model:

It yields a new relation, namely, *the square is on the right of the circle*, and so this transitive inference emerges from scanning the model. To establish its validity, reasoners need to call on the deliberative system to check that no alternative model of the premises refutes the conclusion.

When premises are consistent with more than one spatial layout, inferences are more difficult than the preceding example (e.g. Byrne & Johnson-Laird, 1989; Carreiras & Santamaría, 1997; Vandierendonck, Dierckx, & De Vooght, 2004). Likewise, reasoners try to construct initial models that do not call for a rearrangement of entities (e.g. Jahn, Knauff, & Johnson-Laird, 2007; Knauff & Ragni, 2011), and inferences that call for such rearrangements are more difficult than those that do not (e.g. Krumnack, Bucher, Nejasmic, Nebel, & Knauff, 2011). Analogous results bear out the use of iconic representations in temporal reasoning, whether it depends on relations such as "before" and "after" (Schaeken, Johnson-Laird, & d'Ydewalle, 1996a) or on the tense and aspect of verbs, as in:

John has cleaned the house.

John is taking a shower.

John is going to read the paper.

Mary always does the dishes when John cleans the house.

Mary always drinks her coffee when John reads the paper.

What is the relation between Mary drinking coffee and doing the dishes?

Participants inferred that Mary drinks her coffee after doing the dishes, in an experiment that controlled such factors as order of mention (Schaeken, Johnson-Laird, & d'Ydewalle, 1996b).

The earlier example of a transitive inference is child's play (see, e.g. Bryant & Trabasso, 1971). But, many inferences based on iconicity are more complex, such as those that combine both spatial and temporal relations in kinematic simulations (see Section 10). The intuitive system can also mislead adult reasoners. It constructs only a simple model of a typical situation. Given this sort of problem:

Ann is a blood relative of Beth.
Beth is a blood relative of Cal.
Is Ann a blood relative of Cal?

Many adult reasoners respond, "Yes". The relation holds in their model, which represents lineal descendants or siblings. They fail to search assiduously for an alternative model—it takes work to engage the deliberative system, or a clue to a possible alternative model, such as a reminder that people can be related by marriage. Indeed, Ann and Cal could be Beth's parents, not blood relatives of one another (Goodwin & Johnson-Laird, 2005, 2008).

Visual images are iconic, and so some theorists suppose that they play a key role in reasoning (e.g. Kosslyn, 1994, p. 404). They do play a role: they impede reasoning. To see why, it is crucial to distinguish among relations that elicit visual images, such as "dirtier than", relations that elicit spatial relations, such as "on the right of", and relations that are abstract, such as "better than". Individuals are slowest in reasoning from visual relations (Knauff & Johnson-Laird, 2002), but do not differ reliably in reasoning from the other sorts of relation. As an fMRI study showed, only visual relations elicited additional activity in visual cortex (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003). Knauff (2013) tells the whole story: visual imagery is not necessary for reasoning, which is just as well because some relations, such as those between sets, have iconic representations that may not be visualizable.

Not everything can be represented in an icon. A crucial example is a negation, such as:

The triangle is *not* on the right of the circle.

Reasoners could try to list all the alternative affirmative possibilities—the triangle is on the left of the circle, behind it, and so on—but it would be critical, not only to include all the possibilities but also to make explicit that the list is exhaustive. Alas, neither these conditions nor the meaning of negation itself can be represented in an icon. The model theory accordingly introduces a symbol for negation, which is linked to its meaning: a negative assertion

or clause is true if, and only if, its corresponding affirmative is false. This meaning goes back to Aristotle's *De Interpretatione*, and with some exceptions, it holds for English usage (Khemlani, Orenes, & Johnson-Laird, 2012). The mental model of the preceding assertion is denoted in the following diagram:

¬ [● ▲]

where "¬" denotes the symbol for negation, and the brackets symbolize the scope of the negation, that is, what it applies to. Hence, a comparison of this model with an actual scene would yield the value "true" if, and only if, the relevant circle and triangle were not in the spatial relation represented in the embedded model.

Few people grasp the concept of "negation", and so prudent experimenters ask them about the "denial" of assertions. But, even so, most reasoners err in enumerating the possibilities referred to by compound assertions, such as:

> He denied that John was watching TV and smoking, or else Ann was writing a letter.

Once again, however, number of models is the key variable (Khemlani et al., 2012). It is harder to enumerate the possibilities for the denial of a conjunction, A and B, which has three models:

¬ A ¬ B

¬ A B

A ¬ B

than to enumerate the possibilities for the denial of an inclusive disjunction, A or B, which has one model:

¬ A ¬ B

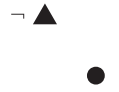
There are plenty of other abstract concepts, such as "possibility", "truth", and "obligation" that transcend iconicity.

5. THE PRINCIPLE OF TRUTH

The model theory postulates a principle of truth: mental models represent what is true, not what is false unless assertions refer to falsity. Here is an example of an exclusive disjunction with a negative clause:

> Either there isn't a triangle or there's a circle.

It has two mental models:



They represent the possibilities in which the disjunction is true, not the possibilities in which it is false. But, the principle of truth applies at a lower level. Both of the preceding models represent clauses in the disjunction only when they are true. In contrast, *fully explicit* models also represent clauses that are false. In the first model above, it is false that there is a circle; and in the second model, it is false that there isn't a triangle, that is, there is a triangle. Hence, the fully explicit models of the exclusive disjunction are:

where we use negation to represent falsity. The fully explicit models show that the disjunction is equivalent to the *biconditional* assertion:

There isn't a triangle if, and only if, there isn't a circle.

Reasoners don't immediately grasp this equivalence—a failure that shows that they rely on mental models, not fully explicit models.

When participants are given a compound assertion, such as a disjunction, and are asked to list what is possible, the principle of truth constrains them, and so they list the possibilities corresponding to mental models (see, e.g. Barres & Johnson-Laird, 2003; Johnson-Laird & Savary, 1995). The advantage of the principle is that it reduces the processing load of reasoning. But, when we implemented the principle in a computer program, we discovered an unexpected downside.

Could both of these disjunctions be true at the same time?

Either the pie is on the table or the cake is on the table, but not both.

Either the pie isn't on the table or the cake is on the table, but not both.

Most people say, "Yes" (Johnson-Laird, Lotstein, & Byrne, 2012). The mental models of what's on the table according to the first disjunction are:

Pie

Cake

And the mental models of what's on the table according to the second disjunction are:

¬ Pie

Cake

The presence of the cake is common to both sets of models, and so it seems that the two assertions can both be true at the same time, that is, when the cake is on the table. In contrast, the fully explicit models of the two disjunctions are as follows:

Pie ¬ Cake

¬ Pie Cake

and:

¬ Pie ¬ Cake

Pie Cake

As readers can see, no model is common to both assertions, and so they cannot both be true at the same time.

The program implementing the model theory predicts these fallacies, and others too. Their occurrence has been corroborated in many sorts of deductions, including those based on:

- Disjunctions and biconditionals of conditionals (Johnson-Laird & Savary, 1999);
- Disjunctions of conjunctions (Walsh & Johnson-Laird, 2004);
- Disjunctions of disjunctions (Khemlani & Johnson-Laird, 2009);
- Disjunctions of quantified assertions (Yang & Johnson-Laird, 2000).

The fallacies tend to be compelling and to elicit judgments of high confidence in their conclusions, and so they have the character of cognitive illusions. Other illusions led to conclusions about what is probable (Johnson-Laird & Savary, 1995), possible (Goldvarg & Johnson-Laird, 2000), and permissible (Bucciarelli & Johnson-Laird, 2005). And still others concerned the evaluation of the consistency of assertions (Legrenzi, Girotto, & Johnson-Laird, 2003). Each study examined several sorts of illusion and matched control problems. Why so many studies of illusions? Because only the model theory predicts them, and so they are a litmus test for the use of mental models.

6. MODELS AS COUNTEREXAMPLES

In reasoning, a counterexample is a possibility that is consistent with a set of premises, but not with a putative conclusion, and so it shows that the conclusion does not follow validly from the premises. The intuitive system can generate at most a single model at a time. To establish the validity of a

conclusion, the deliberative system has to search for alternative models and to show either that no other mental model can be formed from the premises or that the conclusion holds in the alternatives. If the deliberative system creates a model that is a counterexample then it can search for an alternative conclusion that holds in all the models or, if this search fails, declare that no valid conclusion follows from the premises. As we pointed out earlier, a conclusion such as a conjunction of the premises follows validly from any set of premises, and so an alternative conclusion needs to be parsimonious and to establish a new relation not explicitly asserted among the premises. In short, counterexamples are crucial for rationality. Without the ability to create them, individuals can infer conclusions, but they have no ready way to establish their invalidity. So, to what extent do individuals make use of them?

On the one hand, reasoners often fail to use counterexamples when they are drawing conclusions from premises—to the degree that one model-based theory makes no use of them (Polk & Newell, 1995). On the other hand, all the participants in one study spontaneously used them to revise their responses (Bucciarelli & Johnson-Laird, 1999).

There are two sorts of invalid conclusion. One sort contradicts the premises—their respective sets of possibilities are disjoint. The other sort is consistent with the premises, but does not follow from them, that is, there are possibilities to which the premises, but not the conclusion, refer. The model theory predicts that the invalidity of contradictions should be easier to detect than the invalidity of consistent premises: the former don't have a mental model in common with the premises whereas the latter do. The theory also predicts that when individuals are asked to explain why a conclusion does not follow from the premises, they should tend to point out the contradiction in the first case but to exhibit a counterexample in the second case. A study corroborated both of these predictions (Johnson-Laird & Hasson, 2003). The participants were more accurate in identifying invalid inferences in which the conclusion contradicted the premises (92% correct) than those in which the conclusion was consistent with the premises (74% correct). To justify their judgments, they used counterexamples more often for conclusions consistent with the premises (51% of cases) than for conclusions inconsistent with them (21% of cases). Of course, they used other strategies too. One participant, for instance, pointed out that a piece of necessary information was missing from the premises. But, the use of counterexamples correlated with accuracy in the evaluation of the inferences.

An fMRI study contrasted reasoning and mental arithmetic from the same premises (Kroger, Nystrom, Cohen, & Johnson-Laird, 2008).

The participants read a statement of the problem, then three premises, and finally either a conclusion or an arithmetical formula, which they had to evaluate. The experiment included easy inferences that followed immediately from a single premise and difficult inferences that should lead individuals to search for counterexamples, as in this case:

There are five students in a room.
Three or more of these students are joggers.
Three or more of these students are writers.
Three or more of these students are dancers.
Does it follow that at least one of the students in the room is all three: a jogger, a writer, and a dancer?

Most people think first of a possibility in which the conclusion holds. But, those who search for a counterexample may find one, such as this model in which each of the five individuals shown in separate horizontal rows is a student:

Jogger	Writer	
Jogger	Writer	
Jogger		Dancer
	Writer	Dancer
		Dancer

Hence, it doesn't follow that a student is all three. While the participants were reading the premises, the language areas of their brains were active (Broca's and Wernicke's areas), but then other areas carried out the solution to the problems. Right prefrontal cortex and inferior parietal lobe were more active for reasoning than for calculation, whereas regions in left prefrontal cortex and superior parietal lobe were more active for calculation than for reasoning. Right prefrontal cortex—a region known as the right frontal pole—was active only during the difficult inferences calling for a search for counterexamples. Other studies have shown that difficult inferences activate right frontal cortex (Kroger et al., 2002; Waltz et al., 1999). The anterior frontal lobes evolved most recently, they take longest to mature, and their maturation relates to measured intelligence (Shaw et al., 2006). Whether they are activated merely by problems calling for deliberation remains unclear.

7. MODULATION AND THE USE OF KNOWLEDGE

In logic, the interpretation of logical terms is constant, as for its idealized connectives akin to "if", "and", and "or". But, the model theory recognizes that their interpretation in everyday language can be modified by the meanings of the clauses that they connect, the entities referred to in these clauses, and general knowledge. We refer to the process as *modulation*, and we illustrate it with the most notorious case, conditional assertions (see Johnson-Laird & Byrne, 2002).

Conditionals of the grammatical form, *if A then B*, receive a logical interpretation by default. When individuals have to list the fully explicit possibilities to which a conditional refers, they tend to list:

A B

¬ A B

¬ A ¬ B

where *A* and *B* have as values actual propositions (see, e.g. Johnson-Laird & Savary, 1995). Barrouillet and his colleagues have shown that children around the age of 8 years list only one possibility, *A and B*, a conjunctive interpretation; around the age of 11 years, they include another possibility, ¬ *A* and ¬ *B*, a biconditional interpretation; and around the age of 15 years, they list the three possibilities above (Barrouillet & Lecas, 1998). The processing capacity of working memory is a better predictor than chronological age for the number of possibilities that children list (Barrouillet, Grosset, & Lecas, 2000; Barrouillet & Lecas, 1999).

In reasoning, individuals rely on the mental models of conditionals, which consist of an explicit model of the salient case in which both clauses hold, and a content-less placeholder for other possibilities in which the if-clause is false:

A B

. . .

One corollary concerns inferences of the form:

If A then B.

A.

What follows?

Individuals easily infer the conclusion, *B*. It follows at once from the mental models. A contrasting inference is:

If A then B.

Not-B.

What follows?

The second premise eliminates the one explicit mental model, and so it seems that nothing follows—a common response. Only if reasoners flesh out their mental models, or adopt some analogous strategy, can they make the valid inference: *Not-A*. The difference between the two sorts of inference is highly robust. A more striking corroboration of the model theory is that the presentation of the premises in the opposite order improves performance with the difficult inference—it renders unnecessary the need to construct the explicit mental model of the conditional, thus making room for models of the possibilities in which *not-A* holds (see Girotto, Mazzocco, & Tasso, 1997).

Modulation has several effects, and one of them is to block the construction of models of possibilities. A conditional, such as:

If she played a musical instrument then it wasn't the flute

refers to only two possibilities because knowledge that a flute is a musical instrument blocks the construction of the possibility that she didn't play a musical instrument but did play the flute. Hence, the conditional alone yields the conclusion that she didn't play the flute. The principal possibility to which almost all conditionals refer is the one in which both the if-clause and the then-clause hold. Hence, the theory postulates that if a conditional refers to more than one possibility, then this possibility must be one of them. Modulation can accordingly yield the preceding interpretation or the biconditional interpretation. Still other effects of modulation occur with then-clauses that themselves express only a possibility, such as "if Hillary runs then she may win".

Experiments have corroborated that modulation blocks the construction of models (Quelhas, Johnson-Laird, & Juhos, 2010). Consider these two conditionals translated from the Portuguese:

If the dish is lasagne then its basis is pasta.

If the cake is made of eggs then it can be suspiro.

For the first sort of conditional, participants allow that the dish can be pasta but not lasagne. But, for the second sort of conditional, they do not allow that the cake can be suspiro but not made of eggs—all Portuguese know that suspiro is made from eggs. The two sorts of conditionals yield appropriately different patterns of inference.

Consider the inference:

Luisa didn't play soccer.

Therefore, if Luisa played music then she didn't play soccer.

The conditional conclusion refers to three possibilities in which Luisa played, respectively:

Music ¬ Soccer

¬ Music ¬ Soccer

¬ Music Soccer

Hence, the conclusion refers to the possibility to which the premise refers, and so the inference is valid. Yet, most people reject it. Why? One answer is that it is unacceptable because it throws information away, that is, its conclusion also refers to an alternative possibility that conflicts with the premise:

¬ Music Soccer

This conflict, as Orenes and Johnson-Laird (2012) argued, may deter individuals from drawing the inference. If so, then modulation that blocks the conflicting model should yield an acceptable inference, for example:

Luisa didn't play soccer.

Therefore, if Luisa played a game then she didn't play soccer.

The conditional now refers to just two possibilities in which Luisa played:

A game ¬ Soccer

¬ A game ¬ Soccer

The conditional can't refer to the case in which Luisa didn't play a game but played soccer because soccer is a game. So, both the preceding possibilities refer to the same possibility as the premise. In this case, a highly reliable increase occurs in the percentage of participants who accepted the inference. And analogous phenomena occur with inferences to disjunctive conclusions.

Another effect of modulation is to introduce spatial, temporal, or other relations between the if-clause and the then-clause. As a consequence, individuals make different inferences (Quelhas et al., 2010). For example, given these premises:

If Laura got the virus, then she infected Renato.

If she infected Renato, then he went to hospital.

Laura got the virus.

Participants tend to infer that Laura got the virus before Renato went to hospital. But, given these premises:

If Cristina wrote the article, then Marco asked her to write it.

If Marco asked her to write it, then he met her at the meeting.

Cristina wrote the article.

Participants tend to infer that Cristina wrote the article after Marco met her. The temporal inferences depend on the participants' general knowledge about the typical orders of events.

A subtle effect of temporal modulation is illustrated in the following contrasting examples (Juhos, Quelhas, & Johnson-Laird, 2012). The first example is:

If the author writes the book, then the publisher publishes it.

The author writes the book.

What follows?

Individuals tend to infer:

The publisher publishes it.

The second example is:

If the author writes the book, then the publisher publishes it.

The publisher publishes the book.

What follows?

Individuals tend to infer:

The author wrote the book.

The difference is that for the first inference, the participants tended to use the present tense (the experiment was carried out in Portuguese), whereas for the second inference, they tended to use the past tense. As in English, which has no future tense, the present tense in Portuguese can be used to refer to future events. The same phenomenon occurred in inferences from disjunctions. The categorical premise accordingly establishes a reference time, and events prior to it are referred to in the past tense, and events subsequent to it are referred to in the present tense. This sort of modulation is tacit—participants are not usually aware of its effects—but it shows that general knowledge influences the interpretation of conditionals and disjunctions.

A crucial corollary of modulation concerns logical form. A typical formal rule of inference is:

A or B.

Not-B.

Therefore, not-A.

This rule is applicable to any premises that have the corresponding logical forms, which are transparent in logic, because they are defined by its grammar. In language, however, logical forms are far from transparent, and no algorithm exists to determine them because they are not just a matter of grammar. They depend on the possibilities to which assertions refer. So too does validity, and it therefore can be decided only on a case-by-case basis. The model theory makes no use of logical form, but merely the grammatical structure of sentences, and it uses meaning, reference, and knowledge to modulate logical interpretations.

8. INDUCTION AND ABDUCTION

Modulation depends on knowledge, and so it is a bridge from deduction to induction. Inductive inferences yield specific conclusions, generalizations, and, above all, explanations. They depend on knowledge and its availability (Tversky & Kahneman, 1973). An induction may yield a true conclusion; but it may not, even if its premises are true. The engineers in charge of Chernobyl induced that the reactor was intact after the explosion. Their inference was plausible because no nuclear reactor had ever melted down before. But, they were wrong, and their delay in making the correct inference cost lives. Induction is indeed risky.

Logic is "monotonic" in that further premises warrant further conclusions, and no subsequent premise ever calls for a valid conclusion to be withdrawn—not even its direct contradiction. At Chernobyl as in daily life, individuals withdraw conclusions, even valid ones, in the light of subsequent information. Their reasoning is "nonmonotonic". They withdraw some conclusions because they are based on assumptions made by default. They infer, say, that Fido has four legs because Fido is a dog and by default dogs have four legs, but then they discover that poor old Fido has only three legs. This process of withdrawing conclusions based on default assumptions is integral to the model theory (Johnson-Laird & Byrne, 1991). But, retractions also occur in other cases. You believe, say, that if someone pulls the pistol's trigger then it will fire. Someone pulls the trigger. Yet, the pistol does not fire. Hence, there is a conflict between a valid inference from your beliefs—that the pistol fires—and the incontrovertible fact that it doesn't fire. So, you have to withdraw your conclusion and modify at least one of your beliefs. Artificial intelligencers have devised various systems of nonmonotonic reasoning to deal with such cases, but these approaches have grown increasingly remote from psychological plausibility (see Brewka, Dix, & Konolige, 1997). In fact,

at the heart of human performance is the abduction of explanations that resolve inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004). It is these explanations that, as a by-product, yield revisions to beliefs.

An inconsistent set of assertions is a potentially serious matter in daily life. For example, disasters at sea are often a consequence of a conflict between a mariner's mental model and reality (Perrow, 1984). The ability to detect inconsistencies is accordingly one hallmark of rationality. Reasoners can do it by trying to construct a single model of all the relevant information. If they succeed, they evaluate the information as consistent; but if they fail, they evaluate it as inconsistent (see Johnson-Laird et al., 2004, for corroboratory evidence). Once they have detected an inconsistency, they can use their knowledge to try to explain it. The rest of this section focuses on such explanations, that is, abductions.

The basic units of explanations are causes and their effects. In the case of an inconsistency, the effect makes possible the facts of the matter. According to the model theory, causation refers to what is possible and to what is impossible in the co-occurrence and temporal sequence of two events (Frosch & Johnson-Laird, 2011; Goldvarg & Johnson-Laird, 2001). A computer program implementing this account constructs mental models of the premises, as in the pistol example, detects the inconsistency, and uses its models of causal relations to build a chain resolving the inconsistency, for example, *a person emptied the pistol and so there were no bullets in the pistol* (Johnson-Laird et al., 2004). Such an explanation is bound to repudiate at least one previous belief, which reasoners can modify to refer to a situation that was once possible, but that did not occur, as in the counterfactual conditional, *if a person hadn't emptied the pistol and there were bullets in the pistol then the pistol would have fired* (see Byrne, 2005). Experimental evidence showed that individuals are usually able to create such explanations, which tend to refute the conditional premise (Johnson-Laird et al., 2004). Individuals rate the cause and effect as more probable than either the cause alone or the effect alone—a fallacy in which a conjunction is wrongly judged to be more probable than its constituents (Tversky & Kahneman, 1983).

A study of abduction (Johnson-Laird et al., 2004) examined such inferences as:

If a pilot falls from a plane without a parachute then the pilot dies.

This pilot didn't die. Why not?

Some participants made a valid deduction:

The pilot didn't fall from a plane without a parachute.

But, other participants made explanatory abductions, such as:

The pilot fell into a deep snowdrift and so wasn't hurt.
The plane was on the ground and he [sic] didn't fall far.
The pilot was already dead.

An inadvertent demonstration of the imaginative power of human abductions used pairs of sentences chosen at random from pairs of stories, also chosen at random (see Johnson-Laird, 2006, chap. 14). The result was pairs of sentences such as:

Celia made her way to a shop that sold TVs.
Maria had just had her ears pierced.

The participants' task was to describe "what is going on" in such scenarios. To the experimenters' surprise, the participants were usually able to comply. The task was easier in another condition in which the sentences were edited minimally to ensure that they both referred to the same individual:

Celia made her way to a shop that sold TVs.
She had just had her ears pierced.

Typical examples of the participants' responses in this case were:

She's getting reception in her earrings and wanted the shop to investigate.
She wanted to see herself wearing earrings on closed-circuit TV.
She won a bet by having her ears pierced, using money to buy a new TV.

What was striking was how rarely individuals were stumped for an explanation. Human reasoners are adept at abductions—they outperform any existing computer program. Their explanatory ability underlies superstitions (Johnson-Laird, 2006, chap. 14). It also underlies science, but scientists test putative explanations: they search for counterexamples.

A long-standing view of a rational reaction to inconsistency is encapsulated in William James's remark: "[The new fact] preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty" (James, 1907, p. 59). Cognitive scientists have often defended the same view (e.g. deKleer, 1986; Gärdenfors, 1992; Harman, 1986; cf. Elio & Pelletier, 1997, for results to the contrary). Naive individuals, however, are much more concerned to *explain* inconsistencies because explanations can help them to decide what to do. They readily sacrifice minimalism for this goal. For instance, when reasoners are asked what follows from inconsistent premises, they spontaneously offer an explanation that resolves the inconsistency, and they judge that such explanations are more probable than revisions to the premises that restore consistency (Khemlani & Johnson-Laird, 2011). Once they have formulated such an explanation, a striking phenomenon occurs. It becomes harder for them to detect the inconsistency in comparison with cases in

which instead of explaining the inconsistency, they rate which assertion is more surprising. They seem to have explained the inconsistency away, perhaps by reinterpreting the generalization in the premises as holding by default so that it is less vulnerable to contrary facts (Khemlani & Johnson-Laird, 2012b). In sum, reasoners are able to resolve inconsistencies. They tend to do so by using knowledge to abduce causal models that explain the origins of the conflicts. This reasoning usually makes sense of the inconsistency, although on some occasions it fails to yield any explanation whatsoever.

9. PROBABILITIES: EXTENSIONAL AND INTENSIONAL

Induction is often uncertain, and uncertainty implies probability. Individuals who know nothing of the probability calculus happily infer probabilities. How they make such inferences should be part of a unified theory of reasoning. Following Tversky & Kahneman (1983), we distinguish between *extensional* reasoning in which the probability of an event is inferred from the different mutually exclusive ways in which it can occur and *nonextensional* (intensional) reasoning in which the probability of an event is inferred from some relevant evidence or index. In principle, extensional reasoning is deductive, whereas nonextensional reasoning is inductive, and much of it, as Tversky & Kahneman (1973, 1983) showed, depends on heuristics. In daily life, probabilistic reasoning may mix extensional and nonextensional processes. Hence, we explain how the model theory applies to both of them.

Mental models represent possibilities, and so simple extensional inferences can be made on the assumption that each possibility is equiprobable barring evidence to the contrary (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). The probability of an event is accordingly the proportion of models in which it holds. The theory allows that models can also be tagged with numerals denoting their probabilities. Similar principles underlie other theories of probabilistic reasoning (e.g. Falk, 1992; Shimojo & Ichikawa, 1989). The model theory, however, assigns equiprobability, not to events but to models of events. Classical probability theorists, such as de Laplace (1995); (originally published in 1819), advocated an analogous principle of "indifference", but ran into difficulty because events can be partitioned in different conflicting ways (Hacking, 1975). Mental models merely reflect the probabilities that the individual constructing them assigns to events: different individuals can therefore partition events in different ways without self-contradiction.

A simple extensional problem (from Johnson-Laird et al., 1999) is:

In the box, there is a green ball or a blue ball or both.

What is the probability that both the green and the blue balls are there? The mental models of the premise are:

Green

Blue

Green Blue

Hence, the equiprobability principle predicts correctly that individuals will tend to estimate the probability of *green and blue* as 1/3. But, a telltale sign of mental models is that individuals succumb to illusory inferences about probabilities. Here is an example:

> There is a box in which there is at least a red marble or else there is a green marble and there is a blue marble, but not all three marbles. What is the probability that there is a red marble and a blue marble in the box?

The mental models of the premise represent two possibilities:

Red

Green Blue

They imply that the red and blue marbles cannot occur together, and so their probability is zero. And most people make this estimate. However, the disjunction means that when its first clause is true, its second clause is false, and it can be false in three ways:

Red Green ¬ Blue

¬ Green Blue

¬ Green ¬ Blue

When the second clause of the premise is true, the first clause is false:

¬ Red Green Blue

Hence, there are four distinct possibilities for what's in the box, and, on the assumption of equiprobability, the probability of green and blue is, not zero, but 25%. The participants in an experiment performed much better with the control problems than with the illusory problems of this sort (Johnson-Laird et al., 1999).

Turning to nonextensional reasoning, consider this question about a unique event:

What is the probability that Hillary Clinton is elected US President in 2016?

Some psychologists argue that such probabilities are meaningless because probabilities concern only the natural frequencies with which events occur (e.g. Cosmides & Tooby, 1996). However, naive reasoners are happy to make estimates for unique events, and, as we have observed, their estimates correlate reliably over different contents, ranging from politics to climate (Khemlani et al., 2012). The psychological mystery about such estimates is what mental processes underlie them, and, in particular, where do the numbers come from?

We proposed a dual process theory (see Section 3) in which the intuitive system given, say, the question about Hillary, adduces evidence, such as: *Hillary was a very effective Senator; and many effective Senators have become President*. It uses a mental model representing this evidence to construct a primitive non-numerical representation of a degree of belief in the proposition. This iconic representation is akin to the following sort of diagram:

```
| ---^      |
```

in which the left vertical represents impossibility, the right vertical represents certainty, and the pointer at the end of the line corresponds to the strength of the particular belief, such as, *Hillary will be elected President*. The intuitive system can translate this representation into the sorts of description that a non-numerate individual would use, such as: "it's as likely as not".

The deliberative system can map the degrees of belief represented in an icon into a numerical estimate. Because this system has access to working memory, it can carry out proper arithmetical operations. It can also try to keep track of the complete joint probability distribution (the JPD). Given two unique events, such as the election of Clinton in 2016 and the Democrats gaining control of Congress, the JPD consists in the set of probabilities for each possible combination of the affirmations and negations of the relevant propositions:

Hillary is President & Democrats control Congress	35%
Hillary is President & not (Democrats control Congress)	30%
Not(Hillary is President) & Democrats control Congress	15%
Not(Hillary is President) & not(Democrats control Congress)	20%

The JPD provides all the information needed to estimate any probability concerning the domain. There are many different sets of probabilities from

which the values of the JPD can be inferred. For instance, if you know the values of the three probabilities in each of the following triples then, in principle, you can infer that values of the probabilities in the JPD, where, say, *A* denotes "Hillary is President" and *B* denotes "Democrats control Congress":

P(A), P(B), P(A and B)
P(A), P(B), P(A or B, or both)
P(A), P(B), P(A | B)

The last of these triples includes P(A | B), which is the conditional probability of *B* on the assumption that *A* occurs.

Granted the limited ability of the intuitive system to carry out loops of operations (see Section 3), it is capable of only a small number of primitive analogs of arithmetical operations of the sort found in infants (Barth et al., 2006; Dehaene, 1997; Xu & Spelke, 2000) and adults in non-numerate cultures (Gordon, 2004). It can add two pointers, subtract one from another, take their mean, and multiply a proportion signified by one pointer by another—all within the bounds between certainty and impossibility and all in crude error-prone ways. The theory accordingly postulates that to estimate the probability of a conjunction of events, reasoners should tend to split the difference between them, but some may take the proportion of a proportion. The latter is a more complex operation (in terms of Kolmogorov complexity, see Li & Vitányi, 1997), and so it should tend to be used less often. Reasoners should likewise make analogous inferences in estimating conditional and disjunctive probabilities.

We implemented the intuitive and deliberative systems in a computer model and tested its predictions in experiments (Khemlani et al., 2012). The results showed that the participants concurred in the rank orders of their estimates of the probabilities of unique events. For example, they agreed that the US is more likely to make English the official language of the country (the average estimate was a probability of 46%) than to adopt an open border policy (an average estimate was a probability of 15%). Hence, they are to some extent relying on knowledge and mental processes in common. They tended to estimate the probability of a conjunction by taking the mean of their estimates of the probabilities of its conjuncts. This tendency was even evident in the overall means, for example, their mean estimate of the conjunction of the US adopting an open border policy and making English the official language was 26%, a

value falling between their mean estimates of the two conjuncts. It yields a violation of the JPD, that is, the negative probability in the third conjunction shown here:

English	Open borders:	26%
English	¬ Open borders:	20%
¬ English	Open borders:	-11%
¬ English	¬ Open borders:	65%

Violations of the JPD, however, were smaller when the conjunction came last as opposed to first in the sequence of judgments. When it was last, the participants had already made numerical estimates of the probabilities of its conjuncts, and so they could use a deliberative procedure, such as taking a proportion of a proportion. This method is appropriate only for independent events, but a prior study established that they were not independent.

The model theory of probabilities dispels some common misconceptions. Probabilistic reasoning isn't always inductive. Extensional estimates can be deductively valid, but they can also yield illusory values. Likewise, nonextensional estimates of unique events depend on intuitions, and the resulting violations of the JPD suggest that the probability calculus is not native to human cognition. Individuals simulate events, but their restricted repertoire of intuitive methods leads them into error.

10. MENTAL SIMULATIONS AND INFORMAL PROGRAMS

Is there one sort of thinking that depends on mental simulation and that cannot be explained in any other terms? In our view, there is. It is the thinking that underlies the creation of algorithms and computer programs. Expert programming is an intellectual discipline that depends

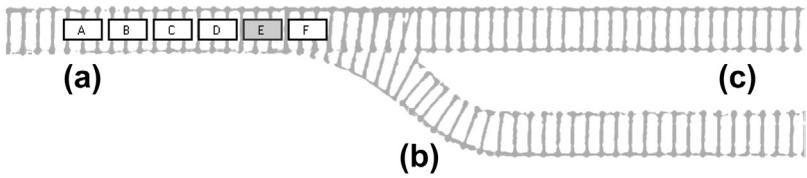


Figure 1.1 The railway domain with an example of an initial configuration in which a set of cars is on the left side (a) of the track, the siding (b) can hold one or more cars while other cars are moved to the right side of the track (c).

on knowledge of programming languages. Hence, our studies focused on how nonprogrammers formulate algorithms in everyday language (Khemlani & Johnson-Laird, 2013). To make the task easy, the programs concerned the railway domain shown in Figure 1.1. The participants had to rearrange the cars in a train on the left track using the siding so that they arrived on the right track in the required new order. In the terminology of automata theory, the siding acts as a "stack" on which to store cars temporarily. Items on the siding move back to the left track, which therefore also functions as a stack. Automata with two stacks are equivalent to Universal Turing machines. Hence, if cars can also be added and removed, the railway serves as a general-purpose computer (Hopcroft & Ullman, 1979).

Most people can solve rearrangement problems in the railway domain. They use a simple variant of "means-ends" analysis in which they work backward from the required goal, invoking operations relevant to reducing the difference between the current state and the goal (e.g. Newell, 1990, Newell & Simon, 1972). For rearrangement problems, they need only envisage each successive car in the goal. Suppose, for instance, they have to rearrange the order ABCD into ACBD. The starting state is:

ABCD[ ]

where the square brackets denote the contents of the siding, which is empty at the start. Their immediate goal is to get D to the far end of the right track:

[ ]…D

So, they move D from left to right track:

ABC[ ]D

The next partial goal is to get B to the right track, and so they need to move C out of the way onto the siding:

AB[C]D

Now, they can move B to the right:

A[C]BD

They move C off the stack:

AC[ ]BD

The next move is intriguing. They should move both A and C together from left to right track. But, if reasoners perseverate, they may move only C to the right track. Their solution won't be minimal because they then have to make a separate move of A to right track. Our initial study investigated

all 24 possible rearrangements of four cars, and the participants easily solved each of them, but they did tend to perseverate: every participant made one or more unnecessary moves.

In the principal experiment, the participants, who were not programmers, had to formulate algorithms for three sorts of rearrangement: *reversals* in which, say, ABCDEFGH on left track becomes HGFEDCBA on right track; *palindromes* in which, say, ABCDDCBA becomes AABBCCDD, and *parity sorts* in which, say, ABCDEFGH becomes ACEGBDFH, that is, cars in odd-numbered positions precede those in even-numbered positions. Each solution calls for recursion, that is, a loop of operations. *Primitive* recursion in the theory of recursive functions corresponds to a loop carried out *for* a given number of times, a so-called "for-loop", whereas *minimization* corresponds to a loop carried out *while* a given condition holds, a "while-loop" (see Rogers, 1967). While-loops are more powerful than for-loops because only they can compute certain functions. Indeed, when a while-loop is entered, there may be no way to determine how many times it will repeat before it yields an output or whether it will ever halt to yield an output. But, how do nonprogrammers formulate informal algorithms? The task isn't deductive: they can deduce the consequences of a program, but they can't create it using deduction alone (Kitzelmann, Schmidt, Mühlpfordt, & Wysotzki, 2002). Likewise, they don't rely on probabilities any more than they do for Sudoku puzzles (Lee et al., 2008). The one viable method is to simulate a solution to a problem, observe what happens in the simulation, and translate these observations into a description. The simulation depends on a kinematic sequence of mental models representing successive states of the world, real or imaginary (Johnson-Laird, 1983, chap. 15).

Let's examine the process in more detail. The first step is to solve two different examples of the relevant rearrangement problem. Without two examples differing in numbers of cars, rearrangements are ambiguous. The solution to reversing a train of four cars is as follows:

ABCD[ ], A[BCD], [BCD]A, B[CD]A, [CD]BA, C[D]BA, [D]CBA, D[ ]CBA, [ ]DCBA

As this protocol illustrates, only three sorts of move are possible, and they occur in these summaries of simulations that solve reversals of trains of four and five cars:

S3 R1 L1 R1 L1 R1 L1 R1

S4 R1 L1 R1 L1 R1 L1 R1 L1 R1

where "S3" means move three cars to the siding from left track, "R1" means move one car to right track from left track, and "L1" means move one car to left track from the siding. The second step uses the two summaries to work out the loop of moves they contain and any moves before or after it (pace Miller, 1974,1981; Pane, Ratanamahatana, & Myers, 2001). The loop in the simulations above is (R1 L1). But, how many times should it be iterated? There are two ways to answer this question, depending on whether reasoners are formulating a while-loop or a for-loop. The simpler way is to observe the conditions in the simulation when the loop halts, which are respectively:

D[ ]CBA
E[ ]DCBA

The condition that halts the loop is that no cars are left on the siding, and so the while-loop should continue as long as the siding isn't empty. The alternative is to compute the number of times that a for-loop should be executed. It calls for the solution of a pair of simultaneous linear equations to obtain the values of *a* and *b* in:

Number of iterations = $a \times$ length of the train + b.

The final step maps the structure of the solution into an informal description. We implemented this entire process in a computer program, which constructs programs for any rearrangement problem based on a single loop. It produces a for-loop and a while-loop in Lisp and translates the while-loop into informal English. Each of these functions solves any instance of the relevant class of rearrangements. Table 1.1 presents its solutions for the three sorts of rearrangement: reversals, palindromes, and parity sorts.

If individuals use simulation to devise algorithms, then they should tend to use while-loops rather than for-loops because it is easier to observe the halting condition of a while-loop than to solve simultaneous equations. The overall difficulty of formulating an algorithm should depend on its Kolmogorov complexity, which is the length of its shortest description in a given language, such as Lisp (Li & Vitányi, 1997). A good proxy is the number of instructions. In Table 1.1, the functions for reversals and palindromes call for four instructions, whereas parity sorts call for five instructions. Within a given level of complexity, another factor should also affect difficulty: the mean number of operands (i.e., cars) per move. This measure distinguishes reversals, which have 1.38 operands per move for eight cars, from palindromes, which have 1.75 operands per move for eight cars. Hence, the three sorts of problem should increase in difficulty from reversals through palindromes to parity sorts.

Table 1.1 Loops for Computing Minimal Solutions to Three Sorts of General Problem: Reversals, Palindromes, and Parity Sorts Using "for"-loops and "while"-loops and Their Informal Description (from the Output of the Computer Program *mReasoner* for Abducing them)

	While-loops	
For-loops	**Lisp**	**Informal English**
a) Reversals (e.g., ABCDEFGH ⇒ HGFEDCBA)		
(setf track (S (+ (* 1 len) -1) track)) (loop for i from 1 to (+ (* 1 len) -1) do (setf track (R 1 track)) (setf track (L 1 track))) (setf track (R 1 track))	(setf track (S (+ (* 1 len) -1) track)) (loop while (> (length (second track)) 0) do (setf track (R 1 track)) (setf track (L 1 track))) (setf track (R 1 track))	Move one less than the cars to siding. While there are more than zero cars on siding. Move one car to right track. Move one car to left track. Move one car to right track.
b) Palindromes (e.g., ABCDDCBA ⇒ AABBCCDD)		
(setf track (S (+ (* ½ len) -1) track)) (loop for i from 1 to (+ (* 1/2 len) -1) do (setf track (R 2 track)) (setf track (L 1 track))) (setf track (R 2 track))	(setf track (S (+ (* ½ len) -1) track)) (loop while (> (length(first track)) 2) do (setf track (R 2 track)) (setf track (L 1 track))) (setf track (R 2 track))	Move one less than half the cars to siding. While there are more than two cars on left track. Move two cars to right track. Move one car to left track. Move two cars to right track.

(Continued)

Table 1.1 Loops for Computing Minimal Solutions to Three Sorts of General Problem: Reversals, Palindromes, and Parity Sorts Using "for"-loops and "while"-loops and Their Informal Description (from the Output of the Computer Program *mReasoner* for Abducing them)—(Cont'd)

	While-loops	
For-loops	**Lisp**	**Informal English**
c) Parity sorts (e.g., ABCDEFGH ⇒ ACEGBDFH)		
(loop for i from 1 to (+ (* ½ len) -1) do (setf track (R 1 track)) (setf track (S 1 track)) (setf track (R 1 track)) (setf track (L (+ (* ½ len) -1) track)) (setf track (R (+ (* ½ len) 0) track))	(loop while (> (length (first track)) 2) do (setf track (R 1 track)) (setf track (S 1 track))) (setf track (R 1 track)) (setf track (L (+ (* ½ len) -1) track)) (setf track (R (+ (* ½ len) 0) track))	While there are more than two cars on left track. Move one car to right track. Move one car to siding. Move one car to right track. Move one less than half the cars to left track. Move half the cars to right track.

Our experiment corroborated the predictions. Individuals were able to create informal algorithms, even though they had no access to the railway domain while they carried out the task. They formulated algorithms for the three sorts of problems, once for trains of eight carriages, and once for trains of any length, in a counterbalanced order. Performance with trains of eight cars was at ceiling, but with trains of any length, it corroborated the predicted trend in accuracy and in time. Likewise, the participants used many more while-loops than for-loops. The use of while-loops correlated with accuracy, whereas the use of for-loops had a negative correlation with accuracy. Differences in ability were striking: the best participant was correct on every problem, whereas the worst participant was correct for only a third of the eight-car problems and for none of problems with trains of any length.

The ability to create algorithms is useful in daily life: loops of operations are ubiquitous in everything from laying a table to shutting down a nuclear reactor. Intelligent individuals are able to carry out this task, and our results corroborated an account that bases their thinking on the ability to make simulations. It is difficult to see how else they could create programs.

11. TOWARD A UNIFIED THEORY

We have now described the main principles of the model theory, illustrated them in various sorts of reasoning, and outlined programs implementing special cases of the theory. But, these programs are a dozen separate pieces, often employing ad hoc notations, and so an urgent task is to integrate them. We have therefore begun to integrate the disparate parts within a single unified theory, implementing it in a large-scale computer program, *mReasoner* (available at http://mentalmodels.princeton.edu/models/mreasoner/). The problem is to bring together various sorts of reasoning (e.g. relational, sentential, modal, causal, quantificational) with various sorts of task (e.g. formulating conclusions, evaluating given conclusions, evaluating consistency) in a way that predicts various sorts of phenomena (e.g. accuracy, latency, effects of modulation). Our aim here was to describe the more important insights that have emerged so far.

A mental model can represent the spatial relations among a triangle, circle, and square (as in Section 4), but the size of the figures, their distance

apart, and so on, may not be intended to represent their real sizes or distances apart. Analogous issues occur with mental models of other sorts of assertion. For instance, a quantified assertion, such as:

Some of the actors are bakers

has the following sort of iconic model shown in this diagram of four individuals:

Actor Baker

Actor Baker

Actor

Baker

The numbers of mental tokens in this case are not intended to represent the actual numbers of actors or bakers. Only the overlap between the two sets is iconic. When reasoners search for an alternative model of a set of premises, they can modify all but the essentials of a model. So, how does the system keep track of the essentials? A single uniform answer is that it relies on the meanings of assertions. Hence, *mReasoner* uses a grammar, a lexicon, and a parser to construct representations of meanings, that is, *intensional* representations (for an account of their construction, see Khemlani, Lotstein, & Johnson-Laird, submitted for publication). They are then used to build *extensional* representations, that is, mental models. Both sorts of representations are crucial in reasoning, and to illustrate this point, we consider the intuitive and deliberative systems in reasoning from quantified assertions.

The model theory treats the intensions of quantified assertions as relations between sets (see Boole, 1854; Cohen & Nagel, 1934). The advantages of this treatment are twofold. First, it dovetails with a long-standing treatment of quantifiers in the model theory (Johnson-Laird, 1983, chap. 15): a set is represented iconically as a set of mental tokens, and a quantified assertion is represented as a relation between such sets. Second, it works for all quantifiers in natural language, including those such as "more than half of the artists", which cannot be defined using the quantifiers that range over entities in logic (Barwise & Cooper, 1981). Here are some illustrative examples of this treatment of quantifiers, which the intensions of assertions capture:

All A are B	$A \subseteq B$	(A is included in B.)
Some A are B	$A \cap B \neq \varnothing$	(Intersection of A and B is not empty.)
No A is a B	$A \cap B = \varnothing$	(Intersection of A and B is empty.)
Some A are not B	$A - B \neq \varnothing$	(Set of A that are not B is not empty.)
Most A are B	$\lvert A \cap B \rvert > \lvert A - B \rvert$	(Cardinality of intersection > cardinality of A that are not B.)
More than half the A are B	$\lvert A \cap B \rvert > \lvert A \rvert / 2$	(Cardinality of intersection > ½ of cardinality of A.)

A corollary is that determiners, such as "most", have parameters specifying such matters as the minimal cardinality of A, the cardinality of the quantifier, "most A", constraints on the relation between the two cardinalities, and so on.

The intuitive system can construct a model of an assertion and update the model according to subsequent assertions. Hence, given the premises of a syllogism:

Some of the actors are bakers.

All the bakers are colleagues.

It updates the model above to:

Actor	Baker	Colleague
Actor	Baker	Colleague
Actor		
	Baker	Colleague

In general, affirmative premises are added so as to minimize the number of distinct sorts of individual, whereas negative premises are added so as to maximize the number of distinct sorts of individual.

Once the intuitive system has an initial model, it can draw a conclusion establishing a new set-theoretic relation, that is, a relation that is not asserted in the premises. In the past, the model theory has eschewed heuristics, but it now embodies them to frame both the quantifier in the conclusion (its mood) and the order of the terms that occur in it: "actors" and "colleagues" (its figure). When two premises differ in mood, one of them dominates the other in determining the mood and figure of the initial conclusion. The order of dominance reflects two principles governing valid inferences:

- A negative premise can yield only valid conclusions that are negative.
- Within both negative and affirmatives, a particular premise—one based, for instance, on "some"—can yield only valid conclusions that are particular.

The resulting order of dominance for syllogisms is accordingly:

Some _ are not _ > No _ are _ > Some _ are _ > All _ are_

In our example, the first premise is dominant: "artists" is its subject, and so "artists" is the subject of the conclusion, and the term in the other premise, "colleagues", is in the predicate of the conclusion, that is, "some of the artists are colleagues". Analogous principles apply to other sorts of premise. They account for the well-known figural effect that occurs in syllogistic reasoning, for example, the tendency to infer the conclusion above rather than the converse, "some colleagues are artists" (see, e.g. Bucciarelli & Johnson-Laird, 1999). The order of dominance is the same as the order invoked in Chater and Oaksford (1999) from probabilistic considerations. But, since our principles derive from valid inferences, and yield only conclusions that hold in initial mental models, they do not depend on probabilities. The heuristics operate without storing any information in working memory, and so they are rapid, but fallible.

The deliberative system makes a recursive search for alternative models falsifying an initial conclusion. When the system finds a counterexample, it formulates a new conclusion if one is possible or else declares that no definite conclusion follows about the relation between the end terms. It searches for counterexamples using the same operations as did participants working with external models in the form of cutout shapes (Bucciarelli & Johnson-Laird, 1999): adding a new individual to a model, breaking an individual into two, and moving a property from one individual to another. As a meta-analysis showed, the resulting theory embodied in *mReasoner* outperforms all current theories of syllogistic reasoning (see Khemlani & Johnson-Laird, 2012c; Khemlani et al., submitted for publication).

The implementation of the unified theory has so far focused on unique probabilities and quantified assertions. It draws its own conclusions from quantified premises, evaluates given conclusions about what is necessary or about what is possible, formulates counterexamples to putative conclusions, and evaluates whether or not a set of quantified assertions is consistent. It carries out these tasks using both the intuitive system, which builds initial models, and the deliberative system, which searches for alternative models. Hence, it is able to predict human inferences. It provides the beginnings of a unified computational account, which we are extending to accommodate sentential and relational reasoning.

12. CONCLUSIONS

The psychology of reasoning would have been simpler if human beings were logicians or probabilists. Logic and the probability calculus are not native mental faculties but cultural discoveries. Some individuals master these technologies; some do not. And, in our culture, most individuals have smatterings of them at best. As the model theory predicts, deductive and probabilistic inferences are difficult and fallible. An awareness of the occurrence of errors led Aristotle and his intellectual descendants to devise logical and probabilistic calculi. These technologies, however, are unlikely foundations for human reasoning. So, what is? We have argued that it is mental simulation. Reasoners build models of premises and base their inferences on them. This view seems undeniable for reasoning that creates informal algorithms. The evidence we have presented shows that it applies also to all the main domains and tasks of reasoning, from deductions based on sentential connectives to inductions about the probabilities of unique events. But, its manifold applications are a source of its main weakness—its potential disintegration into a bunch of separate subtheories. Their unification is viable because each subtheory is constrained by the main principles of the theory. What is much harder is to implement a computer program that predicts the responses that reasoners make to any inferential task, but *mReasoner* is a step toward that goal.

ACKNOWLEDGMENTS

This research was supported by a National Science Foundation Grant No. SES 0844851 to the first author to study deductive and probabilistic reasoning and by a National Research Council Research Associateship to the second author. We are grateful to more colleagues than we can name here, but many of them can be found among the references.

REFERENCES

Bar-Hillel, Y. (1964). *Language and information processing*. Reading, MA: Addison-Wesley.

Barnes, J. (Ed.), (1984). *The complete works of Aristotle*. Princeton, NJ: Princeton University Press.

Barres, P., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking & Reasoning*, *9*, 1–42.

Barrouillet, P., & Lecas, J.-F. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, *67*, 209–253.

Barrouillet, P., & Lecas, J.-F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, *5*, 289–302.

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, *75*, 237–266.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. S. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition*, *98*, 199–222.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*, 372–378.

Boole, G. (1854). *An investigation of the laws of thought*. London: Macmillan.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*, 1–21.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: An overview*. Stanford, CA: CLSI Publications, Stanford University.

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, *232*, 456–458.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naive deontics: a theory of meaning, representation, and reasoning. *Cognitive Psychology*, *50*, 159–193.

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory & Language*, *28*, 564–575.

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT.

Carreiras, M., & Santamaría, C. (1997). Reasoning about relations: spatial and nonspatial problems. *Thinking & Reasoning*, *3*, 191–208.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*, 191–258.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. London: Routledge & Kegan Paul.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–370.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Re-thinking some conclusions of the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *Evolutionary psychology handbook* (pp. 584–627). New York: Wiley.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Dehaene, S. (1997). *The number sense*. Oxford, UK: Oxford University Press.

deKleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, *28*, 127–162.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21*, 419–460.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, J. St. B.T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*, 5–31.
Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition*, *43*, 197–223.
Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*, 280–291.
García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology*, *54A*, 613–632.
Gärdenfors, P. (1992). Belief revision: an introduction. In P. Gärdenfors (Ed.), *Belief revision* (pp. 1–20). Cambridge, UK: Cambridge University Press.
Girotto, V., Mazzocco, A., & Tasso, A. (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, *63*, 1–28.
Goldvarg, Y., & Johnson-Laird, P. N. (2000). Illusions in modal reasoning. *Memory & Cognition*, *28*, 282–294.
Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.
Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468–493.
Goodwin, G. P., & Johnson-Laird, P. N. (2008). Transitive and pseudo-transitive inferences. *Cognition*, *108*, 320–352.
Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science*, *306*, 496–499.
Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.
Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press, Bradford Book.
Henle, M. (1962). On the relation between logic and thinking. *Psychological Review*, *69*, 366–378.
Hopcroft, J. E., & Ullman, J. D. (1979). *Formal languages and their relation to automata*. Reading, MA: Addison-Wesley.
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge, Chapman & Hall.
Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*, 2075–2087.
James, W. (1907). *Pragmatism—a new name for some old ways of thinking*. New York: Longmans, Green.
Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York, NY: McGraw-Hill.
Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, *109*, 646–678.
Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition*, *31*, 1105–1113.
Johnson-Laird, P. N., & Savary, F. (1995). How to make the impossible seem probable. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the seventeenth annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.
Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: a novel class of erroneous deductions. *Cognition*, *71*, 191–229.
Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. S. (1992). Propositional reasoning by model. *Psychological Review*, *99*, 418–439.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J.-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.
Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., Lotstein, M., & Byrne, R. M. J. (2012). The consistency of disjunctive assertions. *Memory & Cognition*, *40*, 769–778.
Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 7–54). Springdale, NJ: Erlbaum.
Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press. (Cambridge, MA: Harvard University Press).
Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 179–212). New York: Psychology Press.
Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
Juhos, C., Quelhas, C., & Johnson-Laird, P. N. (2012). Temporal and spatial relations in sentential reasoning. *Cognition*, *122*, 393–404.
Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Strauss, Giroux.
Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition*, *37*, 615–623.
Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 276–288.
Khemlani, S., & Johnson-Laird, P. N. (2012a). The processes of inference. *Argument and Computation*, 1–17. (iFirst).
Khemlani, S., & Johnson-Laird, P. N. (2012b). Hidden conflicts: explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*, 486–491.
Khemlani, S., & Johnson-Laird, P. N. (2012c). Theories of the syllogism: a meta-analysis. *Psychological Bulletin*, *138*, 427–457.
Khemlani, S., & Johnson-Laird, P.N. (2013). *Mental simulation and the construction of informal algorithms*. In M. Knauff M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2012). The probabilities of unique events. *PLoS ONE*, 7, 1–9. (Online version).
Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. *A unified theory of syllogistic reasoning*, submitted for publication.
Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*, 541–559.
Kitzelmann, E., Schmidt, U., Mühlpfordt, M., & Wysotzki, F. (2002). Inductive synthesis of functional programs. In J. Calmet B. Benhamou, et al. (Eds.), *Artificial intelligence, automated reasoning, and symbolic computation* (pp. 26–37). New York: Springer.
Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*, 363–371.
Knauff, M., & Ragni, M. (2011). Cross-cultural preferences in spatial reasoning. *Journal of Cognition and Culture*, *11*, 1–21.
Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, *15*, 559–573.
Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.
Kosslyn, S. M. (1994). *Image and brain*. Cambridge, MA: MIT Press.
Kroger, J. K., et al. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral Cortex*, *12*, 477–485.
Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, *1243*, 86–103.
Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., & Knauff, M. (2011). A model for relational reasoning as verbal reasoning. *Cognitive Systems Research*, *11*, 377–392.
de Laplace, P.-S. (1995). *Philosophical essay on probabilities*. New York: Springer- Verlag. (Originally published in 1819).

Lee, N.Y. L., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological problem of sudoku. *Thinking & Reasoning*, *14*, 342–364.
Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science*, *14*, 131–137.
Leibniz, G. W. (1685, 1952). The art of discovery. In P. P. Weiner (Ed.), *Selections—Gottfried Wilhelm Leibniz*. New York: Charles Scribners. (Originally published 1685).
Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer-Verlag.
Miller, L. (1974). Programming by non-programmers. *International Journal of Man–Machine Studies*, *6*, 237–260.
Miller, L. (1981). Natural language programming: styles, strategies, and contrasts. *IBM Systems Journal*, *20*, 184–215.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.
Orenes, I., & Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. *Mind & Language*, *27*, 357–377.
Osherson, D. N. (1974). 1976). *Logical abilities in children*. (Vols. 1–4). Hillsdale, NJ: Erlbaum.
Pane, J. F., Ratanamahatana, C. A., & Myers, B. A. (2001). Studying the language and structure in non-programmers' solutions to programming problems. *International Journal of Human-Computer Studies*, *54*, 237–264.
Peirce, C. S. (1931–1958). Collected papers of Charles Sanders Peirce. (Vols. 8). C. Hartshorne, P. Weiss, & A. Burks, (Eds.), Cambridge, MA: Harvard University Press.
Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.
Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.
Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, *63*, 1716–1739.
Reitman, W. R. (1965). *Cognition and thought*. New York: Wiley.
Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
Rogers, H. (1967). *Theory of recursive functions and effective computability*. New York: McGraw-Hill.
Schaeken, W. S., Johnson-Laird, P. N., & d'Ydewalle, G. (1996a). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.
Schaeken, W. S., Johnson-Laird, P. N., & d'Ydewalle, G. (1996b). Tense, aspect, and temporal reasoning. *Thinking & Reasoning*, *2*, 309–327.
Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, *440*, 676–679.
Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: theoretical and experimental analyses of the "problem of three prisoners". *Cognition*, *32*, 1–24.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.
Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.
Störring, G. (1908). Experimentelle Untersuchungen über einfachen Schlussprozesse. *Archiv für die gesamte Psychologie*, *11*, 1–27.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*, 292–315.

Vandierendonck, A., Dierckx, V., & De Vooght, G. (2004). Mental model construction in linear reasoning: evidence for the construction of initial annotated models. *Quarterly Journal of Experimental Psychology, A*, *57*, 1369–1391.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, *11*, 278–293.

Walsh, C., & Johnson-Laird, P. N. (2004). Co-reference and reasoning. *Memory & Cognition*, *32*, 96–106.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74B*, 1–11.

Yang, Y., & Johnson-Laird, P. N. (2000). Illusions in quantified reasoning: how to make the impossible seem possible, and vice versa. *Memory & Cognition*, *28*, 452–465.

CHAPTER TWO

The Self-Organization of Human Interaction

Rick Dale[*,1], Riccardo Fusaroli[*,‡,§], Nicholas D. Duran[*,†], Daniel C. Richardson[†]

[*]Cognitive and Information Sciences, University of California Merced, Merced, CA, USA
[†]Division of Psychology and Language Sciences, University College London, London, UK
[‡]Interacting Minds Center, Aarhus University, Aarhus, Denmark
[§]Center for Semiotics, Aarhus University, Aarhus, Denmark
[1]Corresponding author: E-mail: rdale@ucmerced.edu or cognaction.org/rick

Contents

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00002-2

Abstract

We describe a "centipede's dilemma" that faces the sciences of human interaction. Research on human interaction has been involved in extensive theoretical debate, although the vast majority of research tends to focus on a small set of human behaviors, cognitive processes, and interactive contexts. The problem is that naturalistic human interaction must *integrate* all of these factors simultaneously, and grander theoretical mitigation cannot come *only* from focused experimental or computational agendas. We look to dynamical systems theory as a framework for thinking about how these multiple behaviors, processes, and contexts can be integrated into a broader account of human interaction. By introducing and utilizing basic concepts of self-organization and synergy, we review empirical work that shows how human interaction is flexible and adaptive and structures itself incrementally during unfolding interactive tasks, such as conversation, or more focused goal-based contexts. We end on acknowledging that dynamical systems accounts are very short on concrete models, and we briefly describe ways that theoretical frameworks could be integrated, rather than endlessly disputed, to achieve some success on the centipede's dilemma of human interaction.

1. INTRODUCTION: THE "CENTIPEDE'S DILEMMA" OF INTERACTION RESEARCH

Next time you have a conversation, pay close attention to what you and your partner are doing. This self-consciousness can be a bit jarring. Like the so-called "centipede's dilemma", attempting awareness of your numerous cognitive processes and behaviors, and those of your conversation partner, can quickly disrupt a natural flowing performance. Dialog otherwise seems so easy (Garrod & Pickering, 2004). How do we do it? The famous poem by Katherine Craster has a toad posing to a centipede, "Pray, which leg moves after which?" The centipede goes about some introspection, attempting awareness of this coordination, only to find that she can no longer move.

The same thing seems to happen if we do this in a conversation. The information, and relevant cognitive mechanisms, that can bear on a conversational performance probably outnumbers a centipede's legs, especially if you include the array of processes that are not available to conscious report. How do we coordinate everything?

In this review article, we consider a fundamental and still unsolved puzzle faced by the fields that study human interaction. Much like consciously focusing on ourselves while conversing, the *scientific agenda itself* also suffers from a kind of centipede's dilemma. Ongoing work tends to focus on particular levels of analysis. For example, we know that people use similar vocabulary during interaction, may tend to match in other linguistic styles,

or even take on similar bodily postures and movements. Through exploration of these levels, there are many theoretical proposals of candidate cognitive and social processes. Terminologies that populate these theories are diverse, conceptually overlapping, and still don't enjoy consensus definitions: mirroring, simulating, coupling, entrainment, coordination, imitation, mimicry, alignment, synchrony, joint action, theory of mind, perspective-taking, mutuality, accommodation, empathy, contagion, and more.

What is still lacking is a systematic agenda to uncover how these various processes work together to bring about *multimodal coordination* between two interacting people. The current agenda of isolating processes and developing broad theoretical proposals from relatively circumscribed domains is somewhat like the centipede's analysis of its performance leg by leg by leg. There is now a heterogenous assemblage of experimental techniques and observational analyses and an associated array of diverse theoretical mechanisms that have yet to be integrated. The vibrant debates that ensue focus almost entirely on some subset of experimental paradigms, cognitive processes, or social contexts. The result is an exciting, evolving, but fractured domain.

Admittedly, framing the problem in this way may seem overly grandiose. But if a comprehensive theory of human interaction is our goal, then this is a real puzzle to be solved. We allay any enticement (or skepticism) here: we're not going to solve the puzzle in this paper. We will, however, propose one potential route to a solution. To do this, we look to concepts of adaptive and self-organizing systems, drawing from the tradition of the "dynamical systems framework" as it has come to be known in the cognitive sciences. Importantly, this is a *framework* for thinking about the problem of coordination during linguistic interaction. It does not answer the question of what the precise array of mechanisms is or the processes of their interaction. It also does not, at least by necessity, replace extant theoretical proposals. Our positive thesis is simple: we will argue that the dynamical systems framework may help to integrate existing theories.

In what follows, we begin with more background on the debate about the mechanisms underlying human interaction. We then introduce some fundamental concepts of dynamics. These concepts bring about some generic expectations about how human interaction should be coordinated and structured. We argue that, in fancy terminology, "self-organization into functional synergies" should be (and is) evident in interactive data (Section 2).

We showcase some evidence for this in a review of empirical literature. This background review focuses on two key aspects of linguistic interaction. The first (Section 4) is that basic social variables *can* sharply modulate the

many behaviors involved in interaction, at several levels of analysis. We look to low-level visual attention, and then higher level spatial perspective-taking, during linguistic interaction. The second empirical review is a glance at complementarity between behaviors of two people interacting, extending current theories of alignment in dialog and looking to the usefulness of the concept of "synergy" (Section 5).

Following this, in a concluding discussion, we relate dynamical systems to other theories (Section 6). We speculate on some ways that these dynamic concepts could be pursued in other computational models as a means of becoming more precise about *what exactly* is coordinating during human interaction. As we articulate in this section, a common and important objection to dynamical systems accounts is that they are weak on identifying mechanisms: dynamical theorists argue "interactions dominate cognition," and focus almost exclusively on relatively indirect measurement outcomes of that interaction; however, where there is interaction, there must be things interacting. We address this issue in this final section by arguing that an important way to move forward is to integrate useful concepts from dynamics with some computational frameworks already exploring language and complex cognition.

2. AN EXAMPLE THEORETICAL DEBATE AND THE NEED FOR INTEGRATION

There are some prominent theoretical debates in the realms of discourse and psycholinguistics. One of the best known debates revolves around how, and how much, human beings track information about one another as they interact. Some theories posit a two-stage process, with primacy given to egocentric (self-centered) processes, and more social "other-centric" processes coming online only more slowly and strategically (e.g. Barr, 2008; Keysar, Lin, & Barr, 2003; Lin, Keysar, & Epley, 2010). Other theories posit a fundamental sensitivity to a conversation partner, with a rich layering of common ground that emerges while two people talk (e.g. Clark, 1996; Schober & Brennan, 2003; for recent discussion, see Brennan, Galati, & Kuhlen, 2010; Brennan & Hanna, 2009; Brown-Schmidt & Hanna, 2011; Shintel & Keysar, 2009).

Other theoretical agendas in this debate have aimed to specify key cognitive processes that permit one person to keep track of, or continually adapt to, their conversation partner. Some of these accounts centralize a process of multilevel "alignment," which can be automatic and often nonconscious

and builds common representational states across individuals while they talk (Garrod & Pickering, 2004; Pickering & Garrod, 2004). Some have taken the suggestion of a human "mirror system" to be central, specifying core social processes that must be in place for us to interact successfully (for review, see Gallese, 2008). Recent accounts have articulated the important role of executive function during conversation (Brown-Schmidt, 2009a,b), of memory (Horton, 2005; Horton & Gerrig, 2005), and of the integration of basic contextual parameters of an interaction (Brennan et al., 2010). Still others have identified *kinds* of coordination, such as the emergent versus nonemergent linguistic interaction that see different origins in activities done jointly (Knoblich, Butterfill, & Sebanz, 2011).

We see at least three exciting characteristics to this growing literature. First, researchers in these areas are beginning to tap into the cognitive mechanisms that might underlie social and linguistic interaction (e.g. Brown-Schmidt, 2009a,b; Gambi & Pickering, 2011; Horton, 2005; Mehler, Weiß, Menke, & Lücking, 2010; Pickering & Garrod, 2009; Reitter, Keller, & Moore, 2011). This advances the valuable work on observational and conversation analysis that has shed great light on the structure of interaction (Sacks, Jefferson, & Schegloff, 1995; Schegloff, 2007), but is not capable of identifying the cognitive processes that drive it.[1]

Secondly, and relatedly, social cognitive neuroscience (e.g. Frith & Frith, 2001; Van Overwalle, 2008) and related areas (e.g. imitation: Wang & Hamilton, 2012) have begun to explore these basic mechanisms at the level of the brain. The growth of this subfield of cognitive neuroscience has been very rapid, with many programmatic proposals for studying the circuits underlying social interaction (e.g. Cooper, Catmur, & Heyes, 2012; Dumas, Chavez, Nadel & Martinerie, 2012; Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012; Konvalinka & Roepstorff 2012; Wolpert, Doya, & Kawato, 2003).

Thirdly—and this should sound odd—researchers have come to embrace the inherent social nature of language and to carry out investigations of cognitive processing in more naturalistic circumstances (see Tanenhaus & Brown-Schmidt, 2008 and Fusaroli & Tylén, 2012 for a review). The past century has seen some fundamentally different assumptions for a scientific understanding of language. For example, the classic conception of the ideal speaker–hearer, perhaps useful in some circumscribed domains, is an

[1] Although the brilliant corpus strategies used by researchers like Bard, Aylett, and others can reveal substantial clues about cognitive mechanism through, for example, acoustic properties of what one person says to another (see, e.g. Bard & Aylett, 1999; Bard et al., 2000).

assumption that has outlived any usefulness it may have had in understanding how people actually use language in so wide a circumstance. Conversation analysis and discourse psychology have now been coupled with sophisticated computational and behavioral methods such as natural language processing and computational linguistics (Graesser, Swamer, & Hu, 1997), eye-tracking (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), automated body movement (Paxton & Dale, in press; Schmidt, Morr, Fitzpatrick, & Richardson, 2012) and acoustic analysis (Oller et al., 2010; Wyatt, Bilmes, Choudhury, & Kitts, 2008), dynamical systems methods (Riley & Van Orden, 2005; Shockley, Santana, & Fowler, 2003), and more. Language is a complex and multidimensional activity, and our understanding of it—how it evolved, is learned, and is used—must come from integrating such sophisticated methods in naturalistic circumstances, not *only* from abstract assumptions about linguistic structure that rarely manifest themselves except in preempirical intuitions.

Although these are exciting developments, we would argue that *theoretical integration* has been less emphasized. We can think of a few reasons why this might be. For one, a researcher's theoretical proposals are usually tied to the specific contexts she or he studies. This is a natural feature of any scientific explanation (see Cartwright, 1999), but it limits the generalizability of the processes proposed. The very fact, for example, that language users can be rendered relatively egocentric, or relatively "other-centric," by experimental design means that something more complex is going on cognitively than simply the deployment of fixed architectures (see also Brennan et al., 2010 for discussion). Another reason, in our opinion, is that the multidimensional and "multimechanism" aspect of human interaction means that traditional conceptions of cognitive explanations are fundamentally challenged. In such a complex circumstance, theories seem unlikely to succeed by anchoring to small set of specific mechanisms, but rather to a context-dependent integration of a wide variety of processes acting together. This is what we mean by the "centipede's dilemma": there tends to be much local and circumscribed analysis and much less cross-paradigm and intertheoretical synthesis.

We introduce one way that approaches integration. Specifically, we look to the tools offered by what is often termed the "dynamical systems" approach in cognitive science (Chemero, 2009; Port & Van Gelder, 1995; Richardson, Dale, & Marsh, in press; Spivey, 2007; Thelen & Smith, 1994; Turvey, 1990; more on this below). A dynamical approach to these phenomena affords a variety of theoretical tools that embrace context-dependent

integration, adaptation, and process flexibility. In the study of basic cognitive processing, significant debate has emerged about the usefulness of a so-called "nonlinear interaction-dominant dynamic complex systems" approach (see collection in Van Orden & Stephen, 2012), and whether it really adds much above already present accounts such as constraint–satisfaction mechanisms (e.g. Eliasmith, 2012). These are all very important concerns, and we will address some of them in discussion below. However, this debate seems to have been unfortunately influenced by overly radical and unrealistic theoretical commitments, and perhaps reactionary tendencies in commentators. In many circumstances, a dynamical systems approach can be integrated in telling ways with existing theories. We argue that there is great value in the approach, with specific benefits to be gained from applying it to conversation. We begin by describing the theoretical framework in a highly introductory manner for those who still haven't read much about it or have been too skeptical to get into it.

3. SELF-ORGANIZATION AND HUMAN INTERACTION

3.1. The Need to Integrate Accounts of Cognition in Linguistic Interaction

The empirical literature on conversation and relevant interaction can stymie many theoretical proposals. A cursory glance at this empirical literature shows a more complex story than is typically portrayed in any single theory. The cognitive processes proposed to be centrally involved in social interaction are numerous. In addition, they operate in a highly context-dependent way. Depending on the experimental paradigm chosen by the researcher, one can highlight some capacities over others. As noted above, particular laboratory interaction tasks may produce behavioral patterns indicative of egocentrism (Barr & Keysar, 2002; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; cf. Shintel & Keysar, 2009); at the same time, putting two people who are highly acquainted in an interaction may have a similar effect, of highlighting egocentrism, since each person can make assumptions that they are likely to be understood (e.g. Wu & Keysar, 2007a; whether this is an explicit metacognitive assumption is unknown). However, when establishing pointed moments of conversational disruption, a conversation partner's needs or abilities, or different cultural contexts, these egocentric tendencies can become reversed (see, e.g. Wu & Keysar, 2007b; Brown-Schmidt, 2009a,b; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Galati &

Brennan, 2010; Roche et al., submitted for publication; Tanenhaus & Brown-Schmidt, 2008; see for review Brennan et al., 2010; Schober & Brennan, 2003).

These are exciting avenues of investigation, with extremely clever experiments destined to fuel this debate and discussion. Nevertheless, an inference to the best set of unique capacities is not possible from a small set of experiments or even from a whole literature that highlights specific designs. The range of possible contexts of human interactions is simply too numerous to do so. For these reasons, it is unlikely to be the case that conversational performance and linguistic interaction, in whole, can be accounted for in terms of a small single subset of mechanisms. Of course, not all theories aim to be so comprehensive as they tend to focus on specific aspects of social interaction. We would argue that, to achieve a more comprehensive account of social interaction, an *integration* of these task contexts, and cognitive capacities, is needed (cf. Brennan et al., 2010). But how can we hook up differing accounts into an overall theoretical framework that can achieve this integration? Here, we argue that a dynamical systems framework may serve these questions in valuable ways. In the following section, we describe what we mean by "dynamical systems account" and describe two basic, but important, features: *self-organization* and *synergies*.

3.2. Dynamics, Self-Organization, and All that Jazz

It is widely known that the dynamical systems approach to cognition utilizes some terminology unfamiliar to many cognitive scientists.[2] This concern has been expressed in many critiques. For example, a recent commentary's tongue-in-cheek title uses the comprehensive phrase "[n]onlinearly coupled, dynamical, self-organized critical, synergistic, scale-free, exquisitely context-sensitive, interaction-dominant, multifractal, interdependent brain-body-niche systems" (Wagenmakers, Van der Maas, & Farrell, 2012). These are legitimate concerns because identifying important new theoretical concepts inside an array of unfamiliar terms requires at least some concrete aspects of the agenda (in fact, this is the important point expressed in the above-mentioned commentary). To be fair, however, we could say the same thing about classical information-processing accounts as they emerged. One could construct such a title for virtually any theoretical account as terms that make subtle distinctions or highlight particular nuances are common in all theoretical domains of cognitive science (rich

[2] This is a deliberately mild way of putting it. Others, such as anonymous reviewers of some journal articles, have described the vocabulary as "Star Wars terminology". Readers may have other examples.

vocabularies appear to be a feature of any domain of human expertise; Tanaka & Taylor, 1991). For example, the classical cognitive approach is the "truth-value preserving, hierarchically organized, discretely symbolic, recursive, satisficing, structure-dependent, information-encapsulated modular" approach. Naturally, as Wagenmakers et al. demand, concrete models help anchor such terms, and it is true that the dynamics approach needs more of them (see, e.g. Kello, 2013, for some recent exciting progress; we also discuss this in concluding Section 6).

So before getting lost in a wave of fancy new terms, we expend some energy in this review article discussing why they are used. To do so, we describe a simplified and shortened version of a dynamic self-organized approach to cognition. We do this to present only the most general ideas and avoid some detailed debate that has emerged even in these areas. Readers may be surprised to discover that even among this tribe of cognitive science, there is significant dispute from teeth-gnashing displays that threaten abstract theoretical constructs to pleas to remain open minded about such constructs (for reviews, see Chemero, 2009; Dale, 2008). Still, the core ideas can be laid out readily in a short section, as we attempt here. Our goal was to showcase the specific aspects of this account of cognition that seem helpful to understand conversation (see Richardson et al., in press, for a thorough presentation of both theory and methods).

3.2.1. Complex System

We can take "dynamics" for granted here. The dynamical systems approach takes the position that it is important to study the time-evolving properties of systems.[3] A dynamical system is simply one that is changing in time and can (in some way) be modeled as such (mathematically, computationally, or just conceptually). Instead, let's start with the notion of a "complex system". The phrase itself seems highly relevant to our language abilities. Carruthers (2002) refers to language as an "intersection" system because, among other functions (e.g. complex thought), effective language use requires a wide variety of mechanisms to successfully intersect. Several mechanisms have already been implicated in theories of perspective-taking during dialog and other aspects of conversation. These have included social memory traces (Horton, 2005), memory for shared experiences (Galati & Brennan, 2010;

[3] We will also take "system" for granted. No further words are offered on this. You have to start from somewhere. Readers wishing to have pure operationalized definitions of all things can consult the success of Rudolf Carnap's early-20th-century attempt to do so. One of the authors would wager that readers recognize this attempt in proportion to the success of it.

Wu & Keysar, 2007), social status adaptation (Duran & Dale, 2011), executive control (Brown-Schmidt, 2009a,b), priming processes and alignment (Garrod & Pickering, 2004), the mirror neuron system (Gallese, 2008), forward models of social and linguistic prediction (Pickering & Garrod, in press), rich common ground representations (Clark, 1996), socially guided attention (Kingstone, Smilek, Ristic, Friesen, & Eastwood, 2003), perceptuomotor linkages (Shockley, Richardson, & Dale, 2009), and even processes at various linguistic levels such as perception of accent (Lev-Ari & Keysar, 2010) and lexical (Bortfeld & Brennan, 1997; Brennan & Clark, 1996; Niederhoffer & Pennebaker, 2002) and syntactic choice (Branigan, Pickering, & Cleland, 2000; Branigan, Pickering, Stewart, & McLean, 2000).

Treated as a system of intersecting mechanisms, our language capacity appears quite complicated. In the parlance of researchers who embrace dynamics and complexity science, this *complex system* is unlikely to be controlled by a central "homunculus". No theory of our language capacity has proposed such a central executive that simultaneously integrates all of these mechanisms. For example, emerging models of sentence processing imply that, even at just this processing level, central processing cannot alone account for our success, and an explanation must derive from exploring the dynamic relationship between memory retrieval, working memory, and focal attention (e.g. Lewis, Vasishth, & Van Dyke, 2006; McElree, 2006; Raczaszek-Leonardi, 2010). In the domain of motor control, where dynamical systems have been and continue to be highly influential, this is sometimes referred to as "Bernstein's problem" or the "degrees of freedom problem" (Turvey, 1990). If the components making up our language system are truly modular, there are simply too many ways in which our overall language system can change, with each mechanism flailing about unto itself unless it is somehow anchored to other processes around it. Put simply, there are too many degrees of freedom in this system for it to be managed by a single control process. These many proposed mechanisms must somehow influence each other, directly and continually, in order for language to function in naturalistic circumstances. In any one experiment, we focus on a very deliberately narrowed set of controlling variables and identify their influence on a very specific set of resultant behaviors. Such is the justifiable nature of experimental science.

Naturalistic language performance seems very unlikely to be based on a single control process. Somehow, our system integrates all of these components simultaneously. There are, at present, a limited number of theories for how this is accomplished (although see, for the closest current

approximation of a grand theory, Pickering & Garrod, 2004, 2009, in press). But this is how we wish to pose the problem: If language, in its naturalistic context, is underlain by such a wide array of processes, then these processes must somehow interact, mutually constrain each other, and act together continually to produce coherent performance. Systems that do this—that have a multitude of parts that mutually interact and constrain each other—are often referred to as "complex systems" (see Gallagher & Appenzeller, 1999; articles therein, for discussion). The term is only meant to highlight the problem of interdependency that must be present among the system's components for it to function.[4]

3.2.2. Self-Organization

So if there is not a control process that "calculates the positions" of all mechanisms (working memory, social judgment, visual attention, etc.), then there must be some other means by which we can understand how they function together. A process that contrasts with the presence of a central controller is *self-organization*. Without a central control process, the mechanisms must mutually constrain each other to behave (in whole) as a stable performance. There are plenty of natural examples that are often raised to exemplify this concept (see Kauffman, 1996, for many examples). For example, the behavior of a beehive, termite, or ant's nest is not controlled by a single entity but is a large self-organizing organism unto itself (see Seeley, 2010, and Richardson et al., in press, for more discussion). The same may be true for human interaction.

This is often where things get heated between dynamical systems researchers and other cognitive scientists. Isn't working memory that executive controller? Clearly it cannot be because there is much more being coordinated during conversation than just a handful of manipulated chunks of information (and see note about sentence processing above). What about process threading in working memory extended over time? Recent computational models of complex cognitive control may be relevant here, but even these require articulating the details of interacting components in the system (see, e.g. Salvucci & Taatgen, 2008). Similarly, what we are suggesting is that there may be "chains of influence" between processes of the cognitive system that we tend not to explore. Whatever one's favorite array of

[4] Complex systems are also figured to involve interactions among components, which produce collective higher level behaviors not reducible to properties of the components themselves. We wish to avoid this debate here ("emergence" or "emergentism"), although it seems likely a natural consequence of the perspective we describe here (see also Knoblich et al., 2011).

theoretical constructs or model formalisms, these processes must be working *simultaneously* to bring about stable cognitive performance, which seems especially true for face-to-face human interaction.

So how does self-organization work? Often, when one is trying to convince skeptical colleagues of the *value* of these concepts, already they may point to two issues. First, perhaps this is trivial: "Okay, so you're saying to put it all together, great. That's obvious." Or, the process is far too vague to be even worthy of consideration: "Okay, so you're saying they work together, but that isn't telling us anything new, because we don't know what processes it relies on!" These critiques are entirely legitimate, but the devil is always in the details. The researchers in the dynamics crowd have identified elegant ways of understanding what goes on during the process of self-organization (even in high-level cognition; see, e.g. Dixon, Stephen, Boncoddo, & Anastas, 2010). Many of these dynamical concepts are descriptions of *form* rather than *function*. They are characterizations of what is taking place in the system as it self-organizes. Here, we consider an important one. When a system of many interacting components self-organizes, it undergoes a *reduction in its degrees of freedom*.

3.2.3. Synergies and the Reduction of Degrees of Freedom

As noted above, a key issue raised in motor control decades ago by Nikolai Bernstein was that in order for a human being to perform any coherent action, a massive array of variables must somehow coalesce in order for it to happen. In the 1940s, Bernstein was trying to understand how motor control harnesses the high number and complexity of the components of the human body (Bernstein, 1967; Kelso, 2009; Latash, Scholz, & Schoner, 2007; Turvey, 1990). There is no way we can micromanage each and every joint and muscle at the same time. He introduced the idea of *synergy*: a functionally driven reduction of degrees of freedom, where components do not simply align, but also complement and compensate for each other. Instead of a top–down microcontrol, he hypothesized that the different components get coupled and constrain each other locally.

A classic example is Bernstein's analysis of chisel and hammer. If we want to strike a chisel with a hammer, this gives direction to and constrains the workings of our body. The exact timing and force of contraction and relaxation of all the individual muscles in our hands, fingers, and arm are locally regulated to comply with that overall goal and the unfolding interaction with the environment. This intuition was tested empirically by measuring the precision of movements at all relevant joints in a blacksmith's

hammering of a chisel. The variability of the trajectory of the tip of the hammer across a series of strikes turned out to be smaller than the variability of the trajectories of the individual joints on the hammering arm (Bernstein, 1967; Latash, 2008). The joints are not acting independently but correcting each other's errors at the relevant timescale, to preserve function, thus supporting the idea that the function itself is the coordinating principle. Importantly, when putting the hammer down and, say, grasping a cup of tea, the very same joints and muscles will flexibly combine in very different ways, thus stressing the functional, that is, task-oriented nature of the synergy.[5]

The same idea may be true of human interaction. The array of mechanisms described above do not *merely* interact. They must have interdependencies operating in a coherent fashion that organizes the system into a lower dimensional functional unit, and possibly a much smaller number of stable higher level behaviors, unexpectedly lower than what would be anticipated from the complexity of the system's composition. For example, perhaps at the coarsest level of description in human interaction, one could see stable modes in the form of *arguing* (Paxton & Dale, submitted for publication) or *flirting* (Grammer, Kruck, & Magnusson, 1998) or *joint decision making* (Fusaroli & Tylén, submitted) or *giving directions* (Cassell et al., 2007; cf. the notion of "oral genres," e.g. Busch, 2007). Beneath these coarse-level quasistable characterizations of interaction, we have different levels of coordination taking place. Within one interlocutor, whatever components compose a cognitive system must work together to support coherent individual performance; across two individuals, a similar process of systematic reduction of degrees of freedom may organize interactions into stable modes of functioning (Shockley et al., 2009).

3.3. Summary, Social Modulation, and Multimodal Coordination

So far, we have argued that the study of interaction has faced a kind of "centipede's dilemma" that the field has specialized in specific experimental paradigms, specific behavioral channels and social contexts but that it has not integrated knowledge in a systematic way. Yet, in an important sense, the cognitive system undoubtedly performs this integration during interaction.

[5] Note that this has sparked decades of exciting work and debate in motor control. Although Bernstein's solution has, in many respects, become standard in broad strokes, how it is solved can be the subject of some debate (see, among many, Latash et al., 2007; Newell, Broderick, Deutsch, & Slifkin, 2003; Todorov & Jordan, 2002; Turvey, 2007).

One framework for thinking about this, which we have outlined in brief in the previous section, is to import the concepts of self-organization and synergies into this discussion. Our reasoning is that there are far too many degrees of freedom available to a dyad during conversational performance for the cognitive system to compute their activities all at once. In the following sections, we offer extensive empirical review, looking to two general features of this issue of conversational performance.

3.3.1. Social Modulation of Cognitive Dynamics

The first is the fundamental role of the social in human conversation. Despite the intrinsic social nature of language and conversation, there has been much debate on whether and to what extent social variables, such as the belief states and presence of another person, modulate the dynamics of performance. The self-organization of some cognitive or conversational performance is shaped by social variables in various ways. For example, the mere presence of a person will greatly change behavior of the visual attentional system (as monitored by eye-tracking). In addition, facts that one knows about a potential social partner may guide perspective-taking strategies. We explore this in review of empirical literature. The dynamical systems approach suggests the following basic insight: significant reorganization of interactive behaviors should occur under different social contexts. This sees interaction as a process that is *organizing itself* around key variables, such as a social ones; it also means that basic theoretical accounts emphasizing of "egocentric" or "other-centric" processes may be a dialectical veil over the deeper flexibility and self-organization taking place during human interaction.

3.3.2. Coordination, Complementarity, Synergies

The second is, even if we acknowledge the importance of social variables on the way that cognitive processing unfolds, there must nevertheless be a process of coordination among the channels in conversation. For example, if one learns something new about a conversation partner, it may suddenly shift one's focus both in the content of what is being said and how one says it. Here is where the concept of synergies becomes most important: two people interacting in a joint task come to form their behaviors through compensatory complementary behaviors. These behaviors influence one another locally and incrementally, making the whole conversational performance itself a kind of self-organizing synergy. We review the research suggesting that this is the case and offer theoretical discussion that is meant to supplement, not replace, current discussion of coordination and alignment.

4. COGNITIVE DYNAMICS UNDER SOCIAL CONSTRAINTS

4.1. Social Modulation of the Dynamics of Low-Level Visual Attention

It can be very lonely, being a participant in a cognitive psychology experiment. If the experimenter wants to study memory, language processing, or decision making, a common first step is to exclude as many social factors as possible. The participant sits alone, typically, interacting with a computer, and perhaps a researcher, whom they don't know, follows a rigid script. Like friction in a high school physics problem, social context is discarded by initial simplifying assumptions in cognitive models. Both forces get in the way of more important aspects of phenomena, the argument goes. But both friction and social context are unavoidable in the world outside of the laboratory. In fact, we argue that they are both are essential to understanding many phenomena. The claim is not that what is discovered in a typical cognitive laboratory is invalid because of an absence of social context. Rather, we argue that perhaps there are interesting dynamic interactions between cognitive processes and social context that occur all the time in the real world (Hutchins, 1995a,b, 2010, 2011), but are left at the door of many a cognitive laboratory.

In this section, we review more naturalistic experimental designs that have looked to how social variables constrain the dynamics of low-level visual attention using eye movements. We survey a range of experiments that have been titrated according to the level of social context that they entail. Each measures visual attention, in the form of eye movements, to see how these varying levels of social context influence cognitive and perceptual processing. In the first, the level of social context was high as two people interacted with each other while having a conversation or an argument. The researchers measured how their thoughts about each other drove their eye movements around a scene or an empty screen. In the second, a participant alone was asked to listen to the opinions of one person talking on a screen, but the researchers found that their eye movements were influenced by the presence and the identity of others in the background. In the lowest level of social context, the researchers gave participants an explicitly nonsocial task of looking at a set of pictures while sat alone. The researchers studied the effect of introducing a minimal social context to the task by telling them that another unseen person

was also looking at the same images at the same time. Across all of these levels, the researchers find a pervasive effect of social context on low-level visual processes.

When two people are engaged in a conversation with each other, they can display great sensitivity to each others' thoughts and beliefs. But what happens at a lower level of social context? Here, we look at the case where the participant is merely a spectator, watching a prerecorded video of a group of people and listening to one of them give their opinion. A standard cognitive approach might be to focus on the words of the speaker and how they are processed by the participant. But Crosby, Monin, and Richardson (2008) looked at the relevance of the other people in the video, the silent bystanders who provided a social context.

In their experiment, participants watched a video of four people giving their views on Stanford University's admissions policies. All four members of the "focus group" could be seen on screen at the same time, in a grid arrangement of cubicles, and they all wore headphones so they could hear what each other said. At one point, the speaker in the top right corner complained that, "certain groups who come from less privileged backgrounds… get an unfair advantage." At this point, participants routinely fixated a man on the bottom row of the screen who was black. It appeared that participants were sensitive to the fact that the speakers' words, criticizing policies of affirmative action which would typically benefit black Americans, might be offensive to members of that group.

A parsimonious explanation, however, is that the participants simply noticed the ethnicity of everyone at the start, and the speaker's words simply activated a memory of one person on the screen. This memory trigger launches an eye movement, as memory representations often do (Richardson, Dale, & Tomlinson, 2009; Richardson & Spivey, 2000). But Crosby et al. (2008) were able to rule out this "association hypothesis". In another condition, before the speaker began talking, an offscreen voice said that the headphones of people on the bottom row were being turned off. Then the participants saw exactly the same video. The association hypothesis predicted that since the black member of the focus group was still on screen, he would still attract a fixation. In this case, however, participants barely looked at him when the potentially offensive remarks were made. Participants were supposed to be simply listening to the speaker's words. But these results show that they were also keeping track of the social identity of all the other people on screen, monitoring whether or not they could hear the speaker and, presumably, anticipating how each might respond to the

speaker's words. In other words, for the participants, the cognitive task of processing speech was embedded in a social context.

Other paradigms too have shown an effect of "social tuning" (Shteynberg, 2010; Shteynberg and Galinsky, 2011). In these paradigms, stimuli are explicitly identified as being relevant by other people and as a consequence are processed selectively by individuals. Recently, researchers have adapted the Simon task and inhibition of return paradigms, splitting these cognitive tasks between pairs of participants. The behavior of the pairs is remarkably similar to the individuals', showing the same patterns of response interference (Knoblich et al., 2011). This work suggests that when participant act jointly with each other, in a very simple social context, they immediately represent each others' tasks and goals.

The final set of experiments attempt to reduce social context to it lowest level. Our strategy was to take a simple perceptual task that participants carry out alone or jointly with another person and to make the difference between those conditions as small as possible (Richardson et al., 2012). Pairs of participants were sat in opposite corners of a laboratory room, each looking up at a screen while their gaze was tracked (see Figure 2.1). On each trial of the experiment, they saw four images on screen for 8 s. Beforehand,

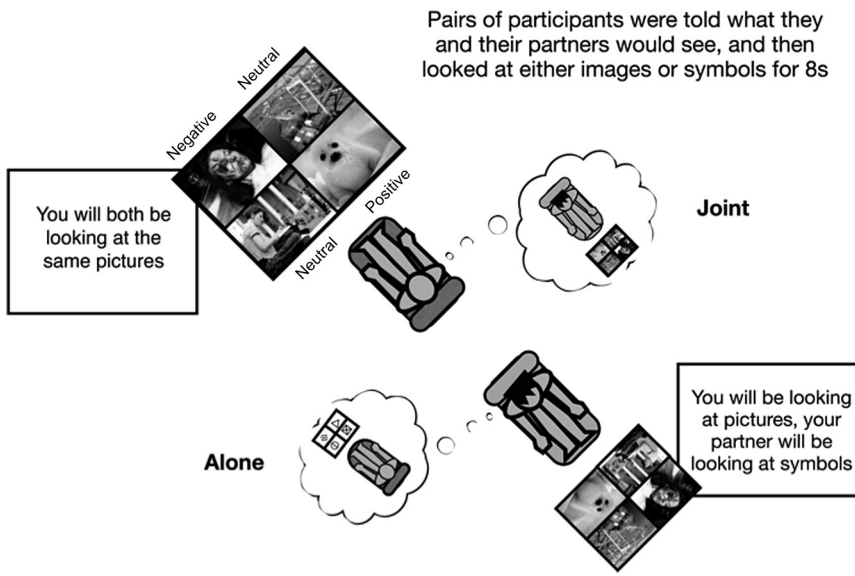


Figure 2.1 Example context in which two people were given different beliefs about what their "socially copresent" partner could see. Imagines of differing emotional valence were presented. *(Adapted with permission from Richardson et al. (2012)).*

they were either told that both they and their partner would be looking at the same images or that they would be looking at images and their partner would be looking at symbols. Participants could not see each other and could not interact at all. Nevertheless, when they thought that their current perceptual experience was being shared with another, their eye movements were systematically changed. They looked more toward pictures with a negative valance than when looking alone. We believe that people are doing so because they each believe that the other person is looking more at the negative images. At least, when participants are told that this is a memory task and that they will score more points if they recall the same pictures as their partners, they too look more at negative images (unpublished data). When the images are replaced by album covers, people will look more at classical albums when they are looking jointly, and their partner (a confederate in this case) walked in carrying a violin (unpublished data). Across all these experiments, we have found that even a minimal social context—the belief that an unseen other was seeing the same stimuli—was enough to manipulate an individuals' visual processing.

From a rich interactive conversation to listening to one person against a backdrop of bystanders, to gazing at images alone, believing that someone else is too, social context can have a pervasive influence on visual attention (Risko & Kingstone, 2011; Laidlaw et al., 2011). Under even subtle circumstances, people take into account each other's knowledge and visual context to coordinate their gaze around an empty display. They anticipate each others' responses to potentially offensive remarks. And even with the thinnest slice of social context, when there is no interaction or contact between people, they will still shift their gaze toward where they think each other is looking. Importantly, this process occurs when the variables are in the right arrangement—for example, when the unintended recipient of the offensive remark *can hear it*.

4.2. Social Modulation of Higher Level Processes, Like Perspective-Taking

Social variables can radically alter the dynamics of visual attention. Here, we review recent research suggesting that these same dynamic changes take place in relatively higher level cognitive processing: perspective-taking. In pragmatic models of language processing, an essential component of how people produce and comprehend language depends greatly on communicative function. The social environment thus has a central role in constraining how language is interpreted and used (Brennan et al., 2010; Brennan & Hanna, 2009; Brown-Schmidt & Hanna, 2011;

Brown-Schmidt & Tanenhaus, 2008; De Jaegher, Di Paolo, & Gallagher, 2010; Hanna & Tanenhaus, 2004; Hanna, Tanenhaus, & Trueswell, 2003). A critical source of constraint is in the common ground that may exist between language users. Common ground corresponds to the shared characteristics derived from the local context of being present with another, such as viewing the same scene, to more global shared histories that arise from being members of the same culture or speaking the same language. This information is brought to bear when interpreting what another says and when choosing words to speak (Clark & Krych, 2004; Clark & Wilkes-Gibbs, 1986; Fussell & Krauss, 1992; Lockridge & Brennan, 2002; Schober & Brennan, 2003).

A central question is *when* common ground information is available in language processing. As described earlier in this review, at one theoretical extreme is a view that people are primarily egocentric and that even when common ground information is available for influencing a particular interpretation, people initially rely on their own frame of reference or act to minimize their own difficulty in processing (Epley, Keysar, Van Boven, & Gilovich, 2004; Keysar, Barr, & Horton, 1998). Otherwise, as the argument goes, to integrate common ground early in processing would result in increased cognitive effort and processing times (Barr, 2008; Keysar et al., 2000). But such conclusions stand in contrast to other studies that show people are quite capable of making rapid social judgments based on briefly presented sources of social information, such as dispositional expressions (Ambady, Bernieri, & Richeson, 2000) or gaze direction (Hanna & Brennan, 2007; Teufel, Fletcher, & Davis, 2010). This social information can also extend to belief attributions about another, such as another's needs, characteristics, or limitations that are initially present in an interaction (Bortfeld & Brennan, 1997), or are emergent factors (Horton & Gerrig, 2002, 2005). Integrating common ground information does not necessarily have to be a cognitively complex process, as simple attributes of another can immediately constrain what and how something is interpreted. Moreover, such integration is commensurate with "incremental models" of language processing. As words are encountered in a sentence, new evidence is provided for the commitment, or abandonment, of a particular interpretation. As a sentence unfolds, multiple interpretations are simultaneously activated and competing for expression, with accruing evidence constraining possible interpretations (Seidenberg & MacDonald, 1999). Based on this account, common ground information is tantamount to just another source of potential constraint.

From a "traditional" dynamical systems perspective, what constitutes a relevant constraint is similar, but is more connected to what can be directly perceived from within an interactive social environment (Marsh, Richardson, Baron, & Schmidt, 2006; Marsh, Richardson, & Schmidt, 2009; Richardson, Marsh, & Schmidt, 2005). Contrary to egocentric accounts of processing, there is no intermediary "representational" stage where another's intentions are first calculated and then acted upon. Rather, the behavior and actions of another hold immediate sway on how people respond to each other. Such direct couplings are possible through the interactive context by which individuals' processing capabilities are reshaped by the presence and actions of social partners (Ramenzoni, Davis, Riley, Shockley, & Baker, 2011; Riley, Richardson, Shockley, & Ramenzoni, 2011). Such connectivity allows for nimble social coordination that cannot be reduced to individual-level contributions, but instead must be evaluated on the basis of the social unit, where the emergence of meaning is reciprocally caused and maintained by social partners during interaction (Marsh et al., 2009).

On the face of it, this traditional dynamical systems approach should be closely aligned with language theories that allow for common ground constraints to have immediate influence on moment-by-moment linguistic processing. Yet, what constitutes common ground often corresponds to simple inferential states based on what another knows or believes. Such inferences are not easily integrated with a dynamical systems approach, where interpersonal coordination is driven by the actions, or possibilities for action, that are expressed and integrated by physically co-situated agents. Rather, what is found in the social environment are opportunities for merging individual-level perceptuomotor systems into a collective system, with evidence taken from joint action tasks where people rapidly converge on patterns of coordinated synchronous movements (Knoblich et al., 2011; Knoblich & Sebanz, 2006). Put simply, common ground invokes the cognitive, and many dynamical systems theorists avoid this.

So if social interaction is grounded exclusively in coupled perceptuomotor systems, there appears to be little room for the role of informationally grounded sources in shaping language processes. To push the boundaries of the dynamical approach, explanations need to go beyond motor behaviors alone to more abstract properties of the social environment (Chemero, 2009). Such attributional properties do not necessarily require elaborate mental operations or representation, but instead can be thought of as spontaneously elicited opportunities for social responding, embodied from past histories of social interaction. As Schmidt (2007) describes, these

experiences elicit tendencies to respond in socially appropriate ways and are sustained by cultural expectations and reinforced by the immediate social context. This behavior is inextricably defined by the relationship between social partners and "affords" opportunities for responding, even during language comprehension.

One of the simplest belief attributions in a language task is whether a communicative partner is an intentional agent (Gallagher, Jack, Roepstorff, & Frith, 2002). Although most interactions provide situated cues to determine veridicality, there are scenarios, such as with computer-mediated interactions, where people may be uncertain whether they are interacting with someone real or simulated. When people are told beforehand the true nature of personhood, their response orientations change in systematic ways, even though the actions of the other remain the same (Nass, Fogg, & Moon, 1996). For example, when a partner is thought to be simulated, people are more likely to use language that is less complex, presumably because of perceived communicative limitations of an artificial intelligence (Branigan, Pickering, Pearson, McLean, & Brown, 2011). Thus, simple belief attributions have the capacity to guide various modes of responding (Wilkes-Gibbs & Clark, 1992). A challenge for dynamical systems is to explain how such behavior arises through an emergent self-organized process in which informational couplings within a social environment produce complex but systematic behaviors (also see Di Paolo & De Jaegher, 2012; Richardson, Marsh, & Schmidt, 2010).

4.3. Perspective-Taking as Self-Organization under Social Constraint

Duran and Dale (in press) present one such attempt by employing a dynamical simulation of response resolution in a task where participants' beliefs were central in disambiguating utterances spoken by another. In this response data, participants and a simulated agent were "connected" within a virtual environment through an elaborate ruse in which the participant was unaware of whether their partner was actually simulated (Duran, Dale, & Kreuz, 2011). This omission allowed participants to form their own impressions about the reality of their partner. For each trial, the task proceeded with the simulated agent instructing the participants to select one of the two objects in the shared environment (see Figure 2.2). The position of the partner shifted from trial to trial, sometimes creating a situation in which the intended referent was ambiguous. When both participants were in the same location, perspectives were ostensibly shared, and an instruction such

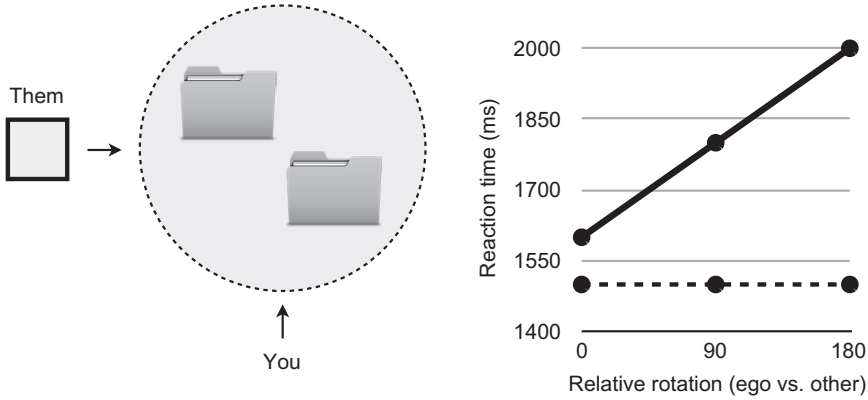


Figure 2.2 When participants are asked to retrieve a folder in an "ostensibly social" computer task as shown on the left, they tend to exhibit different mental rotation functions, predicted by whether they are taking theirs or another person's perspective. When the ostensible partner asks for a folder in ambiguous trials (left), participants taking the other perspective exhibit a mental rotation function (right). Figure based on the designs and typical mental rotation results found in Duran & Dale (in press), Duran et al. (2011), and originally of course the classic Schober (1993).

as, "Grab the object on the right," permitted straightforward identification. However, when participants' positions were across from each other, the same instruction created an ambiguous referent as to whose "right" was the basis for interpretation. That is, instruction receivers could either select the object on their partner's right, thereby taking into consideration the perspective of the other, or they could select the folder on their own right, in what would be an egocentric interpretation.

Despite an informationally situated social context, simple belief attributions, in the form of whether a communicative partner was thought to be real or simulated, the researchers were able to shape perspective-taking strategies across participants. Specifically, when people believed that their partner was simulated, other-centric responding was facilitated in three key ways: (1) the likelihood for other-centrism increased, (2) responses were faster within and across trials, and (3) decreased competition from the egocentric response option. Such behavioral change was also seen in a dynamical systems simulation that treated attributional factors as a control parameter within a low-dimensional attractor landscape of partly stable perspective-taking modes. In this simulation, when a person takes the perspective of another, this response stabilizes through the graded accumulation and competition of both other- and egocentric factors. Thus, contrary to existing models of language processing where these factors are mostly

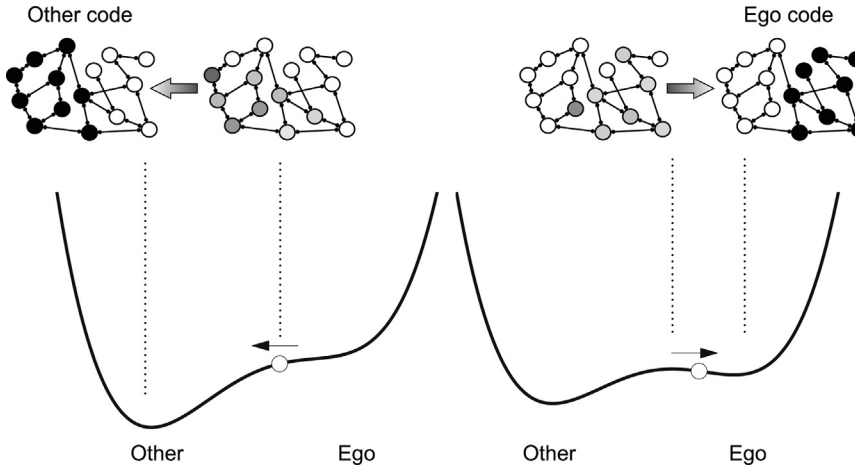


Figure 2.3 Perspective-taking can be modeled as a low-dimensional dynamic process. Simulating this landscape allows qualitative fits to three timescales of human data: (i) decisions, (ii) response times, and (iii) response dynamics (Duran & Dale, in press). See Figure 2.4 for illustration. A full discussion of this lower order characterization of higher dimensional dynamics can be found in Onnis and Spivey (2012).

independent components, here, they are explicitly allowed to interact from the very start of processing to influence behavior over time (e.g. multipotentiality).

To conduct the simulation, Duran and Dale (in press) borrowed from the Haken–Kelso–Bunz model that was originally developed to capture the relative coordination of bimanual motor movements in time and space (Kelso, 1981, 1995; see Figure 2.3). This model has been extended to a variety of domains, showing widespread commonalities between perceptual, cognitive, and motor systems (e.g. Engstrom, Kelso, & Holroyd, 1996; Frank, Richardson, Lopresti-Goodman, & Turvey, 2009; Tuller, Case, Ding, & Kelso, 1994; Van Rooij, Bongers, & Haselager, 2002; see Schmidt & Turvey, 1995; Chemero, 2009, for reviews). It draws from core principles of bistable dynamics to allow complex behaviors to self-organize over time, with responses unfolding within a low-dimensional attractor landscape. Thus, perspective-taking during communication, much like in the previous research, could be described as following coordinative dynamics similar to those observed in perceptuomotor coupling.

There is a precedent for considering high-dimensional processes in low-dimensional forms. Recently, Onnis and Spivey (2012) have advocated for linking such means of modeling and visualizing systems. Assuming, for example, that perspective-taking can be represented as a high-dimensional

neural process, akin to a population code (top row of Figure 2.3), one can derive a direct visualization in lower dimensions of how the system is transitioning between stable states (bottom row of Figure 2.3). Self-organization into "other" or "ego" perspectives can be seen as traversing a low-dimensional landscape.

In the current instantiation of the model, a control parameter in a potential function is set to initiate bistable attractor basins of other-centric or egocentric interpretations. These basins are a reduction of system complexity to a quantifiable and transparent outcome variable of the two perspective types. In other words, perspective-taking is characterized as in a system of substantially reduced degrees of freedom, a "lower dimensional" cognitive space in which choices are made. The particular shape of each basin corresponds to the likelihood and speed in which the system can settle into a particular response, with deeper and steeper basins indicating a stronger pull and therefore more rapid stabilization. During the time course of a single trial, "settling" occurs through nonlinear competition between landscape shape, initial conditions (i.e. starting position in the landscape), and a subtle noise impulse. When a response threshold is met, the control parameter is adjusted, and a new trial is allowed to run. This is analogous to updating a belief about one's partner being real or simulated at the end of each trial, with beliefs becoming stronger and more stable over time. In doing so, the global characteristics of response choice stability are captured, as well as the competition effects that influence the moment-by-moment processes involved in response execution. By capturing the response dynamics also exhibited by human participants, simple social constraints, in the form of belief attributions, are essential pieces of information that bias a system's "perspectival" landscape and thus its eventual behavioral strategy (see examples in Figure 2.3).

Figure 2.4 shows the timing and response characteristics of data seen in Duran and Dale (in press) and Duran et al. (2011). In the top left, a representative *response* histogram of this task is shown. The model can capture two stable strategies, namely as attractors in the lower dimensional landscape. Models of this kind naturally display organization in the form of two stable "modes" or strategies. The bottom left panel shows that, over trials, humans display a drop in reaction time across trials, but more so for ego versus other reaction times. Finally, the panel to the right shows a zooming into the very dynamics of responses. Subjects often show faster more direct ego trajectories; if they are responding using the other-centric perspective, their mouse movement trajectories tend to show more curvature. These are three

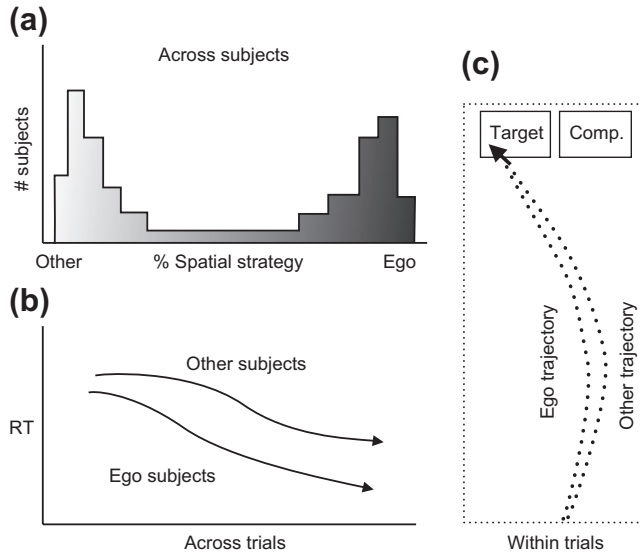


Figure 2.4 A representation of the three timescales of a task that the lower dimensional model of Duran & Dale (in press) can capture.

interconnected timescales from the response distribution of participants down to response times to the fine-grained dynamics of responses. Duran and Dale (in press) show that setting one set of parameters can simultaneously model all three such timescales.

Perspective-taking in communication is fundamental to how language is used and understood. This was certainly evident in the Duran et al. (2011) study. Somewhat counterintuitively, participants were more likely to consider the other's perspective when they believed the other to be simulated. This result makes sense when communication is viewed as a collaborative process between language users. The goal of communication is to maximize mutual understanding, and when one partner is unable or hard-pressed to do so, the other will compensate by putting in increased effort, such as engaging in other-centric perspective-taking behavior (Clark, 1996; Clark & Krych, 2004; Clark & Wilkes-Gibbs, 1986; Goodwin, 2003). This tendency emerges, although past histories of social interaction where people actively attempt to establish mutual understanding. Attributions about others' abilities to cooperate are eventually embodied by language users and brought to bear in responding, even in simple communicative scenarios such as the one described above (also see Schober, 1993, 1995). When participants think they are interacting with a simulation, where cooperation

by the other is not even possible, the "afforded" response is to assume the other's perspective in interpreting their ambiguous instructions.

Such spontaneous perspective-taking occurs despite enacting increased cognitive demands. However, demand is minimized in communication where assessments about a partner can be reduced to simple alternatives, such as whether a conversational partner is very young or not or has a language disorder that changes the goals of mutual understanding (Newman-Norlund et al., 2009; Perkins & Milroy, 1997). Of course, reciprocal and emergent constraints occur during the course of an interaction that subsumes any individual-level sources of difficulty. The dyad operates as a unit that collaboratively minimizes processing load, with success depending on the level of coordination shared between the language users (Fusaroli et al., 2012; Louwerse et al., 2012).

5. COORDINATION, COMPLEMENTARITY, AND INTERACTIVE PERFORMANCE

The foregoing review suggests that the dynamics of both low- and high-level cognition can be seen as responding to subtle social variables present in the environment. Subtle changes to these variables can lead to rapid changes in the organization of those processes (e.g. visual attention radically changing when you learn something new about your partner). Such rapid changes are hallmarks of self-organizing systems—the capacity for rapid nonlinear change.

However, even below these subtle social variables, there are considerably more dimensions that must be managed by a dyad, namely the array of behavioral and cognitive possibilities during an interaction. As noted in the introduction, language is increasingly acknowledged as a *social coordination* device, a way of accomplishing otherwise difficult or impossible coordination of actions and cognitive processes (Clark, 1996; Fowler, Richardson, Marsh, & Shockley, 2008; Fusaroli, Gangopadhyay, & Tylén, submitted for publication; Fusaroli & Tylén, 2012; Galantucci, 2009; Galantucci & Sebanz, 2009; Hasson et al., 2012; Hutchins & Johnson, 2009; Louwerse, Dale, Bard, & Jeuniaux, 2012; Pickering & Garrod, 2004; Tylén, Fusaroli, Bundgaard, & Østergaard, in press; Tylén, Weed, Wallentin, Roepstorff, & Frith, 2010). Through language, we can easily entertain a friend while waiting for her bus to arrive, exchange words with a stranger in an elevator, coordinate in carrying a heavy piano down a flight of stairs, negotiate the price of an apartment, share information, and make joint decisions. However, as earlier

described, coordinating via language is a complex business. A great number of studies have been dedicated to unveiling the crucial subtleties in coordinating not only topics, lexical choices, and syntax but also gestures, gaze management, head movements, and postural sways, which underlie conversations (Goodwin, 2000, 2011; Louwerse et al., 2012). Such a multimodal richness seems to imply serious uncertainty (Jaeger, 2010) and cognitive load (Garrod & Pickering, 2009) for the participants: how does a conversant choose between all the possible linguistic and nonlinguistic behaviors on so many different channels at once? How does a person focus his or her attention and interact in a meaningful way? In other words, how can interlocutors seemingly effortlessly orchestrate all these dimensions (each level, presumably, with its own numerous degrees of freedom) in tight intra- and interpersonal coordination?

As a reaction to computationally heavy models of conversations, requiring theory of mind and full accommodation of models of the other, there has been a strong focus on low-demanding bottom-up models of linguistic coordination, such as the model of interactive linguistic alignment (Pickering & Garrod, 2004). After a brief presentation of this model, we will argue that it should be integrated and complemented in the larger model of interpersonal synergies, presenting evidence supporting this and a few studies testing the prediction of the synergy model.

5.1. Behavioral Synchrony and Interactive Alignment

One intuitive way of reducing the complexity in interpersonal interactions is to diminish the range of possible behaviors via a progressive adaptation to each other. By becoming increasingly similar, the interlocutors greatly simplify the cognitive load needed to interact with the other. Indeed, there is strong evidence for behavioral mimicry (Chartrand & Van Baaren, 2009) and interactive linguistic alignment (Pickering & Ferreira, 2008; Pickering & Garrod, 2004). Interacting human beings have been observed to mimic each other's posture, gestures, and other behaviors (Chartrand & Van Baaren, 2009). A prototypical example of an experimental investigation of this kind of human unconscious mimicry is the "Chameleon effect" (Chartrand & Bargh, 1999; Dijksterhuis & Bargh, 2001). In this experiment, participants interacted with an unknown confederate in two consecutive picture-describing sessions. In one session, the confederate either rubbed her face or shook her foot while describing the pictures with the participants, while the second confederate performed the behavior that the first confederate did not. The behavior of the participants, "secretly" recorded on

videotape, showed that participants shook their foot more in the presence of the foot-shaking confederate and rubbed their faces more in the presence of the face-rubbing confederate. Debriefing indicated that participants were unaware of their mimicry. Analogously, facial expressions, gestures, and yawns have been observed to spread across interlocutors and around a room (Louwerse et al., 2012; Platek, 2010).

Pickering and Garrod have argued that mimicry is commonly co-opted in linguistic interactions through what is called "interactive linguistic alignment": interlocutors tend to imitate each other's choice of linguistic forms. Participants primed with a specific syntactic structure are more likely to produce new sentences employing the same syntactic structure under circumstances in which alternative nonsyntactic explanations could be excluded (Bock, 1986; Branigan et al., 2000; Gries, 2005; Hartsuiker & Westenberg, 2000; Levelt & Kelter, 1982; Pickering & Branigan, 1999; Smith & Wheeldon, 2001; Szmrecsanyi, 2005, 2006). Analogously, topics (Angus, Smith, & Wiles, 2012; Angus, Watson, Smith, Gallois, & Wiles, 2012) and lexical choices (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987; Garrod & Clark, 1993; Garrod & Doherty, 1994; Orsucci, Giuliani, & Webber, 2006; Orsucci, Giuliani, & Zbilut, 2004; Orsucci, Walter, Giuliani, Webber, & Zbilut, 1997; Wilkes-Gibbs & Clark, 1992) tend to be imitated across interlocutors. Linguistic alignment can also be found at more subtle levels of linguistic coordination: interlocutors align accent and speech rate (Giles, Coupland, & Coupland, 1991). More recently, a lot of effort has also been put in showing that the organization of pauses in and between interlocutors' speech and their average pitch, intensity, and voice quality tend to become similar over time (De Looze & Rauzy, 2012; Kousidis & Dorran, 2009; Lee et al., 2010; Lelong & Bailly, 2011; Levitan & Hirschberg, 2011; Nishimura, Kitaoka, & Nakagawa, 2008; Pardo, Gibbons, Suppes, & Krauss, 2011; Truong & Heylen, 2012; Vaughan, 2011). In a single conversation, many of these channels will be aligned, as recently shown in a massive study by Louwerse et al. (2012). These channels have been argued not to be independent. On the contrary, aligning on one channel in many cases seems to facilitate alignment on others. For instance, syntactic priming is enhanced when the same lexical items or even just semantically related ones are also repeated (Branigan, Pickering, & Cleland, 2000; Branigan, Pickering, Stewart, et al., 2000; Cleland & Pickering, 2003).

Several mechanisms have been proposed to underlie these phenomena: most researchers seem to agree on an unconscious priming mechanisms, a "perception–action link" (Chartrand & Bargh, 1999; Dijksterhuis &

Bargh, 2001) or "structural priming" (Pickering & Ferreira, 2008), in other words the overlapping between mechanisms involved in perceiving a behavior and producing it, which implies that by perceiving a behavior the participant preactivates the production of the same. Other authors prefer to focus on the conscious aspects of alignment, where interlocutors try to take each other into account, developing conceptual pacts on which words to use (Brennan et al., 2010; Clark & Brennan, 1991).

Whatever the mechanisms at work, alignment within and across modalities is a very effective way to reduce degrees of freedom—namely the possible behaviors from which to choose. Not only the other's behavior can be used as guide in how to behave but also the repertoire of possible behaviors is reduced over time. The mechanism of alignment might be differently motivated, but it is sensible to argue that once it is established, it plays an important role, making linguistic interactions more manageable. However, a few problems arise when we take it at face value as the fundamental motor of linguistic coordination. A few studies are pointing out that not all conversations contain the same amount of linguistic alignment (Healey, Howes, & Purver, 2010; Reitter, Moore, & Keller, 2006) and that coordination might not rely on alignment across neighboring speech turns, but on the contrary across many speech turns, thus escaping the tight temporal constraints of automatic priming (Reitter & Moore, 2007). At a more intuitive level, a conversation constituted exclusively of reciprocal repetitions should not strike anybody as a very productive one.

5.2. An Alternative Model: Interpersonal Synergies

As described earlier in this review, the reduction of degrees of freedom is not a new problem. Bernstein (1967) proposed that functional units of motor control are established through mutual constraint among the parts of the body and motor control system, effectively reducing the degrees of freedom of the system into "synergies". Recently, Ramenzoni and colleagues (Ramenzoni et al., 2011; Ramenzoni, Riley, Shockley, & Baker, 2012; Riley et al., 2011) have been exploring interpersonal motor synergies. They showed that in joint actions participants increasingly coordinate hands, forearms, and torsos, forming reciprocally compensating synergies spanning across individuals. While studies of interpersonal motor coordination are not rare (Marsh et al., 2009; Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007; Schmidt, Bienvenu, Fitzpatrick, & Amazeen, 1998; Schmidt & Richardson, 2008), only very recently has this approach been applied to linguistic coordination (Fusaroli, Raczaszek-Leonardi & Tylén, in press).

The rest of this section will argue that linguistic coordination is achieved through interpersonal synergies, that is, through functionally driven reduction of the degrees of freedom involved in the interaction. This approach does not dispense with alignment but introduces additional mechanisms—complementarity and interactional patterns—and integrates alignment with this dynamical inspired perspective.

5.2.1. Complementarity

Several studies have pointed out that interlocutors strive to complement each other's behavior to develop a structured conversation. For example, turn-taking seems one of the most elementary examples of complementarity: a remarkable—and seemingly universal (Sidnell & Enfield, 2012; Stivers et al., 2009)—ability of humans to *not* do the same thing at the same time, that is, stay quiet when the other speaks. Simultaneous starts are reported to be surprisingly rare in dyadic conversations (Jefferson, 1988), even if more than 50% of the pauses between interlocutors are below the usual threshold for reactions (300 ms). Wilson and Wilson (2005) have been developing a model of turn-taking that explains this fine-tuned complementarity: the beginning of an interaction sets up an oscillator in each of the interlocutors' cognitive systems establishing a shared frequency of speech rate (see also Buder & Eriksson, 1999). This cyclic pattern governs the potential for initiating speech at any given instant for both interlocutors. The interlocutors, in other words, have to keep the same pace (alignment). However, if the oscillators were simply entrained in phase, simultaneous starts would be frequent. Therefore, the oscillators must be entrained in *antiphase*, giving the participants both a common rhythm, constituted by speech rate and length of comfortable pauses, and complementarity—readiness to take the floor must be opposite at any given moment for speaker and hearer. This ability seems to appear at a very early developmental stage (Gratier & Devouche, 2011; Murray & Trevarthen, 1985; Nadel, Carchon, Kervella, Marcelli, & Réserbat-Plantey, 1999; Spurrett & Cowley, 2004; Warlaumont, 2012).

Recent work on conversations involving patients with speech impairment further shows the importance of complementarity. Expert interlocutors—for example, family members—tend to engage compensatory procedures to keep the conversation fluent despite the impairment (Dressler, Buder, & Cannito, 2009; Goodwin, 2003, 2011; Wilkinson, Beeke, & Maxim, 2003). For example, Goodwin reports on Chil, who, after having suffered a severe stroke, can only speak three words: "yes," "no," and "and". Despite this clear impairment, Chil is able to engage in complex conversations by coordinating other people's

utterances. Chil thus relies on different types of reciprocal compensatory moves to restore the dialog: on the one hand, interlocutors have to actively produce utterances completing and supporting Chil's conversational moves. On the other hand, Chil's three words are relational ones: they do not communicate much on their own, but make sense only in a conversational situation. Together with a host of nonverbal means such as facial expressions and gesture, Chil employs his minimal vocabulary to couple with the other interlocutors' communicative activity. Relying on three words, he is able to coordinate, support, supplement, and sometimes reject his interlocutors' utterances (Goodwin, 2011). Similarly, Dressler and colleagues (Dressler et al., 2009) have explored prosodic patterns in conversations with aphasic patients. They report that conversation with familiar interlocutors displays overall prosodic rhythms, which are much more fluent and regular than conversations with unfamiliar interlocutors.

5.2.2. Interactional Patterns

Beyond this basic rhythm of interaction, conversation analysis has persuasively shown how speech turns are often organized in functionally structured sequences of turns, such as adjacency pairs: questions are ordinarily responded to with an answer, not with another question; offers and invitations are ordinarily followed by acceptances or declinations, and so on (Schegloff, 1986). Turns and adjacency pairs are themselves not free-floating entities, but often fulfill a role in larger interactional patterns, locally unfolding routines that scaffold and constrain the possibilities of actions and interpretation in joint activities (Clark, 1996; Levinson, 1983). Interactional patterns are typically conceived of as normative static phenomena already shared—or assumed to be shared—by interlocutors (Sacks, Schegloff, & Jefferson, 1974; Schank & Abelson, 1977). The synergy approach, however, implies that these elements are part of a dynamic context-sensitive interaction. Interactional patterns vary in formality and flexibility from free and relatively unconstrained conversation over the morning coffee to tightly structured and sometimes even explicitly codified task-oriented conversations (Hutchins, 1995a, 1995b; Perry, 2010). Interactional patterns work to reduce the overall degrees of freedom of the system in a functionally driven way and enable a smoother flow of the interaction.

A number of recent studies indirectly show that ad hoc interactional patterns emerge and are maintained in task-related interactions. In a version of "the maze game" (Healey & Mills, 2006; Mills & Gregoromichelaki, 2010), it was observed that, over the course of 12 games, participants radically structured and shortened their linguistic exchanges from more than 150

turns to brief and efficient exchanges. Through a shared history of interaction, the structure of their interaction is stabilized. This enabled participants to smoothly produce and interpret highly elliptical and fragmentary utterances without much negotiation or clarification. Extending this work, Mills (2011) systematically investigated how these interaction patterns emerge and spread in a small speech community. Each participant played a number of games with shifting partners within a "community". Then, in a critical test trial, half of the participants were paired with a member from another community. This perturbation seriously disrupted the interaction in the affected groups. Participants were found to edit their utterances to a much higher degree, were observed to explicitly acknowledge each other's utterances more often, and overall performed less accurately. The findings suggest that interactional patterns emerge from a shared history of interaction and come to implicitly constrain the degrees of freedom of the interlocutors, diminishing ambiguity and supporting a smoother and more effective flow of the coordination (for a more comprehensive discussion of these issues, cf. Mills, in press, and Fusaroli, Raczaszek-Leonardi, & Tylén, in press).

5.3. Testing Models of Linguistic Coordination: Alignment and Synergy

The review so far suggests that complementarity, in the form of systematized patterns of interaction between two people, is a crucial component of human interaction. Three recent studies based on the same experimental design (cf. Figure 2.5) have tried to test implications of the model (Bahrami et al., 2010; Fusaroli et al., 2012; Fusaroli & Tylén, submitted; Fusaroli, Abney, Bahrami, Kello, & Tylén, submitted). In the experiment, pairs of participants were instructed to individually indicate in which of two brief visual displays they had just been shown a contrast oddball. If their individual decisions diverged, they were prompted to discuss and reach a joint decision. In order for a pair to achieve a cooperative benefit, that is, to perform better than the best of the individuals, they had to find ways of assessing and comparing their individual levels of confidence so as to choose, on a trial-by-trial basis, the decision of the more confident participant. In other words, they had to develop an interactional pattern for accurately expressing confidence and smoothly taking joint decisions relying on that.

This paradigm generated a corpus of task-oriented conversations—which emphasizes the development over time of interactional patterns to quickly solve the repeated tasks—as well as an accurate measure of cooperative performance—to assess the efficacy of linguistic coordination. Different

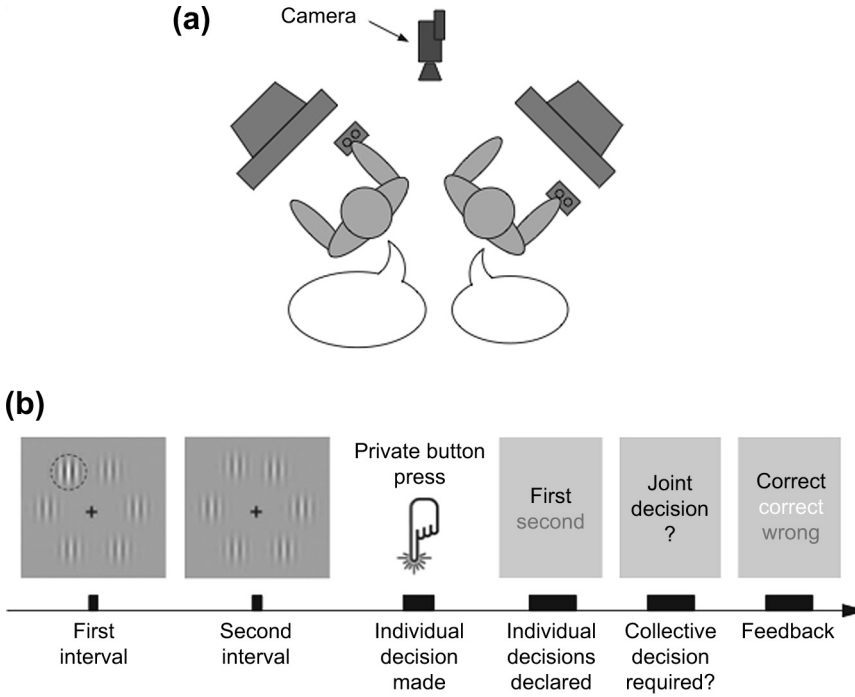


Figure 2.5 Interactive perceptual detection task. (a) Participants both view noisy stimuli and can communicate regarding the presence of a target. (b) The sequence of events in the task from stimulus presentation to the presentation of feedback. Trials began with two stimulus intervals, which contained Gabor patches, with one of them appearing quickly during one of the intervals, and participants had to guess which one. If their decisions did not match, they were required to negotiate about it and come to an agreement. *(Image adapted with permission from Bahrami et al. (2010)).*

aspects of linguistic coordination could be assessed: lexical, prosodic, and acoustic production behaviors.

The first study investigated lexical alignment. As described earlier, this notion of alignment predicts that the more people use the same words, the better they will perform ("indiscriminate" lexical alignment). By contrast, a model of coordination as *synergy* would predict that the alignment of confidence expressions only—serving the interaction's goals of sharing confidence to make a joint decision—would correlate with performance. The analysis did show prominent "indiscriminate" alignment in all pairs: interlocutors displayed a high probability of picking up and employing words used by the other in the previous interaction. However, the more a dyad indiscriminately repeated each other's words, the *lower* the collective benefit they gained from cooperation. Automatic linguistic alignment seemed to be

deleterious to coordination on the task. In contrast, the participants' reciprocal, selective adaptation to vocabularies of expressing confidence (task-motivated selective alignment), turned out to correlate positively with the collective benefit gained from linguistic coordination (see Figure 2.6).

The second study (Fusaroli & Tylén, submitted) more systematically compared linguistic repetitions at three levels: first repetition of triplets of phonemes, second repetition of patterns of pitch, and finally repetition of patterns of speech pause sequences. A model of coordination as alignment would have the structure of repetitions across subjects predicting performance, while a model of coordination as synergy would predict the structure of repetitions at the interaction level—that is, not discriminating between interlocutors—to be correlated with performance. In other words, a synergy model would predict that the relevant coordination happens in interactional patterns where it does not matter which interlocutor shares confidence and which makes a decision, as long as somebody fills those roles in each joint decision. Employing a combination of information theory and recurrence plots (Marwan, Carmen Romano, Thiel, & Kurths, 2007), the authors quantified these repetitions both across interlocutors and in the overall interaction. The results show that the relevant coordination happens at the level of interactional patterns, but not simply across interlocutors: The more the interlocutors develop a regular pattern of lexical choices, pitch and speech pause sequences, which repeats across joint decisions, no matter who is producing its different parts, the better they perform. On the contrary, indices of repetitions across interlocutors did not correlate with performance.

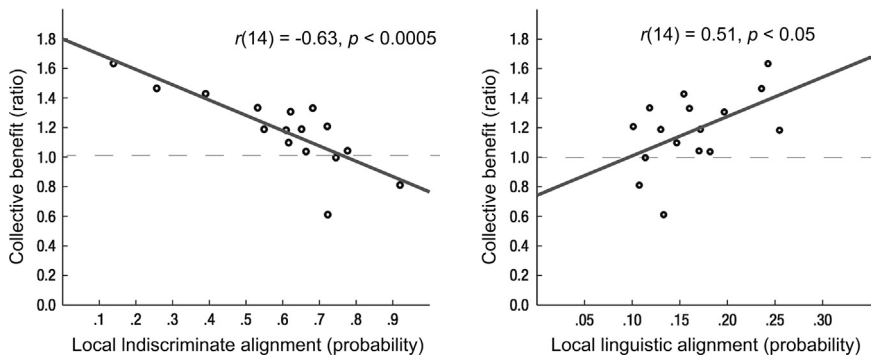


Figure 2.6 Results from the alignment of word usage during the task. The Collective Benefit (*y*-axis) is a measure of how much dyads benefited from their interaction; local linguistic alignment was a benefit (right plot); however, rampant widespread "indiscriminate" alignment predicted a drop in joint performance (left plot). *(Adapted with permission from Fusaroli et al. (2012)).*

A third study further supported these findings and investigated the temporal dimension of coordination, in other words, how synergies self-organize over time (Fusaroli et al., submitted). The researchers showed that individuals' speech production displays scaling laws (lognormal distribution), which are a signature of behaviors constrained by the emerging dynamics of the interaction. Indexes of behavioral alignment (mutual information) were shown to decrease over time, while indexes of more complex multiscale coordination (complexity matching; West, Geneston, & Grigolini, 2008) were shown to increase over time. Finally, the increase in multiscale coordination—and not its initial value—was shown to significantly correlate with performance. In other words, the findings further support the importance of the self-organization of the linguistic interaction in shaping the behavior of its components. Crucially, it is also shown that self-organization happens over time, increasing in strength and efficacy while interlocutors adapt to each other developing cooperative routines.

5.4. Interpersonal Synergies: A Summary

The empirical evidence reviewed strongly suggests that the current focus on interactive alignment as the main engine of coordination has to be integrated in a more complex model of interpersonal synergies, which encompasses complementary dynamics and the development of interactional patterns (coordinative routines). This model makes predictions, which are already supported in a handful of studies: only task-oriented alignment is effective in fostering coordination; important aspects of the linguistic interactions can be described only if we focus on the development of stable patterns of interaction whose role can be indifferently filled in by one or the other of the interlocutors; the self-organization of the interaction takes time, growing in strength and efficacy as the interlocutors adapt to each other, by both aligning and complementing each other.

6. CONCLUSION: TIME FOR MORE MODELS

6.1. Summary

We have offered some discussion and review of how interaction can be understood as a process of self-organization. First, we showed that social variables when perceived, and when taking particular forms, can fundamentally change the cognitive processes and behaviors of a conversation partner. These emerging patterns can be described in the form of "phase transition," where lower level systems become organized differently in a

manner that is shaped by these social variables. But how do these lower level processes constrain each other and act together? Akin to the centipede's dilemma, rather than understanding the interaction "leg by leg by leg," we entertained the notion of a synergy between interacting human beings: the behaviors—turn-taking and rhythms, use of particular words, emergence of adjacency pairs, and so on—can be seen as an array of levels that are mutually constraining, and dynamically evolving, as two people come to form in an important way, a "unit of analysis," and the interaction itself a stable, if temporary, synergy itself. Perspective-taking might be seen as part of this synergetic process, shifting from allo- to egocentric or vice versa, as the interlocutors enact or develop coordinative routines.

6.2. Moving Forward: Models of These Processes

The two key features we have articulated mostly describe the *form* of interaction, rather than the underlying mechanisms that give way to it. This is an issue raised often in discussion and critique of dynamical systems approaches to cognition (e.g. Bechtel, 1997; Dale, 2010; Eliasmith, 1996, 2012; Wagenmakers et al., 2012). In fact, we described that one exciting aspect of growing approaches to social interaction is that these approaches factor in mechanism. One critique of the current review could be that we have simply advocated for a wholesale integration of as much as can be gleaned about mechanism—and this doesn't really tell us much about mechanism. We have advocated instead for conceptualizing human interaction as a system that self-organizes and adapts to particular contexts, such as social variables, and organizes itself through evolving local interactions, such as in incremental contributions to a dialog, including even nonverbal channels, like winks and nods. These are important critiques, and they should be addressed directly. So, we end this paper with a brief review of some modeling endeavors that will help to guide integration of many channels, helping to solve the centipede's dilemma.

6.3. Surface Network Analysis, and Mechanistic Models

One way to get at the synergies directly is to carry out integrative analysis of "multimodal" (multiperson, multibehavior, multilevel) corpora. The past decade has seen a growing agenda to build large-scale corpora of human interactions, capturing a variety of interpersonal behaviors, linguistic contributions, contextual variables, and so on (e.g. Fusaroli et al., 2012; Louwerse et al., 2012). After such collection of data, researchers often go about identifying the relationship among

particular variables, such as gestures and group collaboration or prosodic contours in particular discourse situations. These agendas are important for understanding interactions at particular levels of analysis—the manner in which gestures are deployed, and in what context, and how prosody may index particular modes of interaction.

The argument we have made is that it is a nontrivial mission, both methodologically and theoretically, to discover the manner in which these multiple behaviors, and cognitive processes, are integrated during ongoing interaction. One way to do this "at the surface" is to translate corpora into a form that allows the analysis of the temporal relationship between behaviors. In other words, different behaviors such as nodding, gestures, use of particular words, and so on, can be rendered into analyzable *time series.* This was done by Louwerse et al. (2012), who, at a rate of 250 ms, tracked patterns of synchrony between interlocutors in a direction-giving dialog (e.g. participants tended to laugh and smile together, nod one after the other at a particular timescale). Extraction of time series would permit an exploration of the dynamic interaction between different channels, and between people, and exploring how these change over the course of an interaction. One way to do this is to project the channels into a network structure, with nodes representing the behaviors and edges representing their relationship (e.g. strength of connection).

Consider the following hypothetical research scenario: investigating bouts of human interaction along a set of four behaviors (A, B, C, and D) and measuring these behavioral channels at 250 ms intervals. Such a hypothetical data set is presented in Figure 2.7. Various circumstances may arise during interaction. The channels may exhibit only weak coincidental structure, with each "degree of freedom" of this system being one of these channels. However, if systems exhibited pure "synchrony," then behavioral channels across individuals serve to constrain each other. So, instead of $2 \times 4 = 8$ degrees of freedom in the interaction, we have only 4 since each channel serves to constrain that in the other person.[6] If a process of alignment were to cascade across levels, as predicted in Pickering and Garrod (2004), for example, we would have a continued shrinking of the degrees of freedom. As displayed beneath the middle panel in Figure 2.8, that saturation would

[6] Here, we are using "degree of freedom" in a very informal way, simply to specify whether a channel, or set of channels, is "free to vary" or whether they constrain each other *in some fashion*. Of course, network analysis can involve gradient aspects of these couplings, but we ignore this for simplicity here.

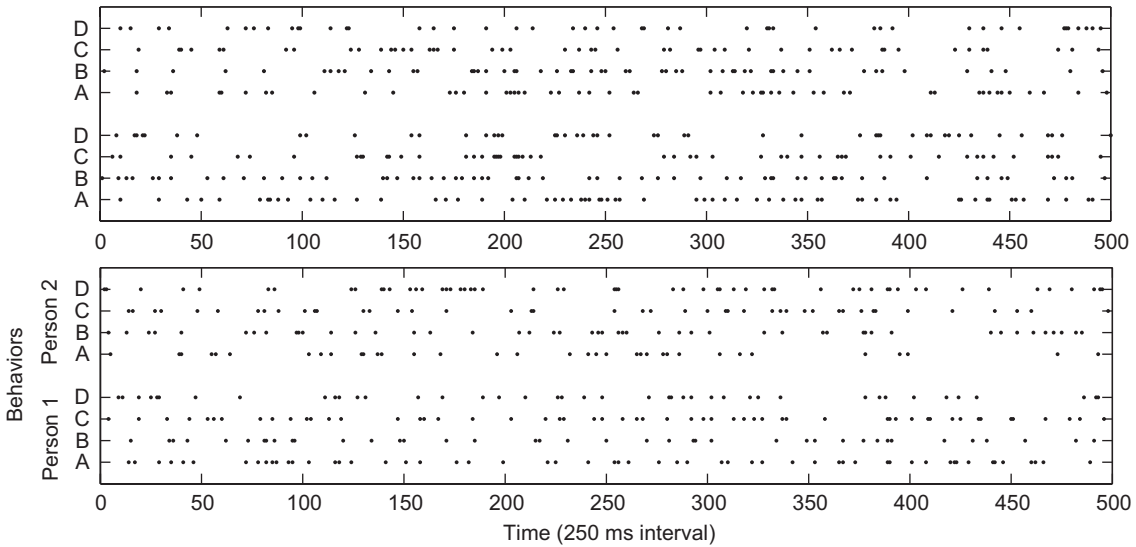


Figure 2.7 *Simulated* point processes of behavioral events. "Person 1" and "Person 2" have four behavioral dimensions (A–D). In human data, these would correspond to delimitable actions such as nods, laughter, or gesture (Louwerse et al., 2009, 2012). Across time, these events occur and may serve to coordinate behavior both *within* and *across* modeled processes. In the top plot, all processes are random; in the bottom, there is a greater probability of "alignment" (e.g. Person 1's A occurring with Person 2's A). This may not be evident by mere visualization, but by inducing a network through (for example) temporal correlation, we can extract the interactive structure of the model (see Figure 2.8).

result in coupling across behaviors. We would effectively have only a single degree of freedom as behaviors fluctuate now all together as one unit.

We know through extensive explorations described earlier in this paper that speech, gesture, and other features of interaction will exhibit coordination (e.g. see Louwerse et al., 2009). This can be identified as clusters in the network that become tightly entrained over time. These are portrayed in the rightmost panel in Figure 2.8. The degrees of freedom relevant to this interaction are now constrained by the number of unconnected subgraphs (in Figure 2.8, top right). As an interaction changes across time, the network structure may change, but the degrees of freedom may stay the same (see lower right figure, "Transition"). We could imagine this sort of thing occurring during face-to-face interaction. Imagine two students discussing lecture briefly, which one of them missed. In this bout of interaction, nodding and gesturing and speaking may have a characteristic temporal interaction. However, if this part of the conversation ended, and one asks the other for directions, suddenly their gaze and gesture may take on that "clamped degrees of freedom" property, while others may change.

This network analysis approach may serve as a powerful means of visualizing and quantifying the "surface configurations" of an interaction and providing clues to underlying mechanisms. The authors are engaged in some early work exploring this possibility (Dale & Louwerse, 2012; Fusaroli et al., submitted; Paxton & Dale, in press; see also related work in Bergmann & Kopp, 2009; Kopp, 2010). There are considerable details in need of investigation if this agenda were to be carried out in the naturalistic context (here we have only sketched this hypothetically using random point processes). For example, what temporal functions best characterize the linking between channels? Gesture and nodding (for example) have a different timescale from, one would suspect, explanation or querying, referred to as "dialog moves". Another issue is what the appropriate measures are to determine that these channels are indeed coupled. Methods such as vector autogression (Dixon & Stephen, 2012), Bayes nets (Bergmann & Kopp, 2009), and related techniques (Shalizi, Camperi, & Klinkner, 2007) may do a better job at capturing the cross-covariation among so many channels.

This "surface network" analysis may be a useful way of proceeding to extract the "hidden" degrees of freedom that are guiding the behavioral structure of a conversation. Still, it is important to note that, in some ways, this research agenda is already unfolding in some prominent projects.

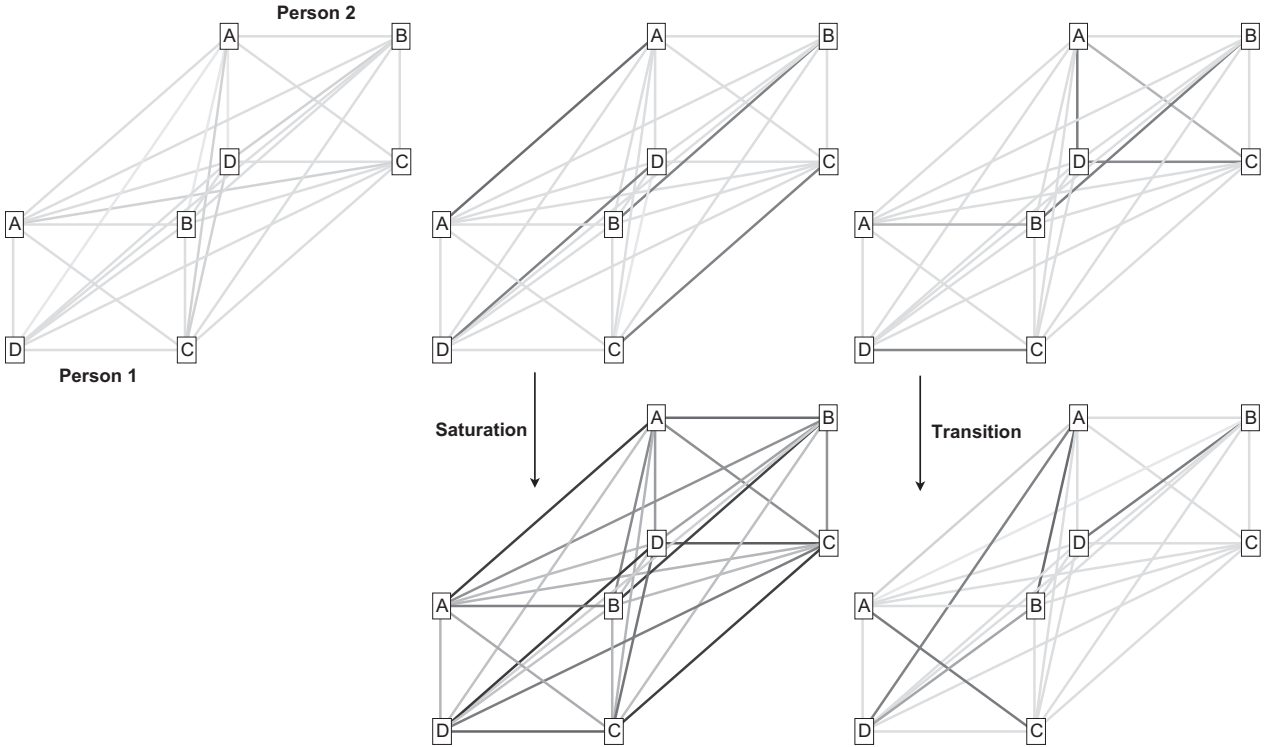


Figure 2.8 An illustration of different graph (network) structures induced from different interrelationships among *simulated* point processes (Person 1 and Person 2). On the top left, only very light gray edges reflect a weak connection across all pairs of nodes (no behaviors are coupled). In the middle column, an illustrate of alignment (A's, B's, and so on, go correlate), with saturation in the behavior (cascading such that behavioral events all occur together in a kind of synchronous multilevel alignment). In the rightmost column, a synergistic structure, where there is occasional alignment, but amidst a variety of other interconnections that may fluctuate from moment to moment (e.g. top panel: $\{A_{Person\ 2}, D_{Person\ 2}, C_{Person\ 2}\}$, $\{A_{Person\ 1}, B_{Person\ 1}, B_{Person\ 2}\}$, $\{D_{Person\ 1}, C_{Person\ 1}\}$, transitions to bottom panel: $\{D_{Person\ 1}, B_{Person\ 1}, A_{Person\ 2}\}$, $\{A_{Person\ 1}, C_{Person\ 1}\}$, $\{D_{Person\ 2}, B_{Person\ 2}\}$).

For example, consider the case of the Augmented Multiparty Interaction corpus (AMI), which tracked several simulated meetings at many behavioral levels (not unlike the above simulation). Computer vision and speech automation techniques permitted the extraction of a wide range of behaviors among several people while discussing topics such as designing artifact prototypes. From extraction of multiple channels, researchers have been developing automated techniques for capturing argumentation (Hakkani-Tur, 2009), the structure of the meeting (Murray, Renals, Carletta, & Moore, 2006), the emergence of particular emotions (Reidsma, Heylen, & Ordelman, 2006), and so on. This work is beginning to leverage, in essence, the probabilistic relationship among multiple channels during interaction. Perhaps, these probabilistic models will help solve the centipede's dilemma.

Once we have this surface structure, and potentially even an estimate of the number of "freely moving parts" of an interaction, there is still the open question of the specific *cognitive processes* that underlie the control of these degrees of freedom. Lack of space precludes a detailed review, but there are many exciting possibilities that may be pursued. In models of reading and sentence processing, Miyake and colleagues have used latent variable modeling to relate cognitive processing tendencies with individual differences measures like the Wisconsin card sort and dual-task batteries (Miyake et al., 2000). This individual differences approach, through statistical modeling, may be useful in the interactive context to identify the cognitive constraints on specific forms of interaction. This agenda has begun in the work of Brown-Schmidt (e.g. 2009a,b) who has identified the role of executive control in predicting the extent to which one person is likely to integrate knowledge of another during interaction. These are statistical models, but there are also some computational possibilities. For example, rational models and adaptive control theory have allowed some researchers to tap into the dynamic relationship among hypothesized cognitive processes and constraints to capture, for example, reading and other language comprehension (Lewis et al., 2006; Bicknell & Levy, 2010; Smith & Levy, 2008). It may be possible to induce something akin to a Hidden Markov process, beneath the "surface structure" we articulated above, and specifying in greater detail the cognitive interactions taking place that control that surface behavior. As noted earlier, mathematical development of these models in the motor control literature has reached a very sophisticated level (e.g. Todorov & Jordan, 2002).

6.4. Conclusion

We have not advocated for an approach that supplants existing theoretical accounts of interaction. The role of memory (Horton, 2005), executive control (Brown-Schmidt, 2009a,b), alignment and priming processes (Pickering & Garrod, 2004), coordination and adaptation (Brennan et al., 2010; Schober & Brennan, 2003), perceptuomotor coupling (Richardson et al., 2009; Shockley et al., 2003), accessibility accounts (Barr & Keysar, 2002), and so on, are all crucial for accounting for interacting persons. Although the authors of this article may wrestle with each other on this grander point, it seems instructive to proclaim that *all of these theories* have central contributions to play in accounting for interaction. One major, and very simple, reason for this could be advanced in the following way: These theories have been usefully deployed in specific experimental contexts investigated by the researchers who have advocated for them. This means that they have strong empirical backing in some subset of human interactive situations; a corollary of this is that they are *predictive* of human interaction in similar situations.

We have argued that it is time to integrate, to go beyond the centipede's dilemma, and gain an understanding of the manner in which processes coordinate and act together. Motivated by basic concepts of self-organization and synergy, we described a series of experiments that show the flexibility of human interaction. Under different social situations, low- and high-level cognitive processes flexibly adapt. Under different task conditions, the dyad self-organizes through local exchanges, incrementally emerging, that develop whole new "synergies". By exploring the structure and underlying control mechanisms, perhaps an integration of these theories will be possible. We haven't done this here, but we have provided some clues that seem useful to us. We hope some readers feel the same way.

ACKNOWLEDGMENTS

This research was supported in part by NSF grants BCS-0826825 and a Minority Postdoctoral Research Fellowship (to the third author), and in part by the Danish Council for Independent Research - Humanities (FKK) project "Joint Diagrammatical Reasoning in Language", and the EUROCORES grant EuroUnderstanding "Digging for the Roots of Understanding".

REFERENCES

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. In Mark P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201–271). Academic Press.

Angus, D., Smith, A., & Wiles, J. (2012). Human communication as coupled time series: quantifying multi-participant recurrence. *IEEE Transactions on Audio, Speech and Language Processing*, *20*, 1795–1807.

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: insights into effective doctor-patient consultations. *PLoS ONE*, 7, e38014. http://dx.doi.org/10.1371/journal.pone.0038014.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*, 1081–1085. http://dx.doi.org/10.1126/science.1185718.

Bard, E. G., & Aylett, M. P. (1999). The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In Proceedings of the 1999 International Conference on Spoken Language Processing (pp. 1753–1756).

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*(1), 1–22.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: listeners anticipate but do not integrate common ground. *Cognition*, 18–40.

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, *46*(2), 391–418.

Bechtel, W. (1997). Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*(3), 295–318.

Bergmann, K., & Kopp, S. (2009). GNetIc–Using bayesian decision networks for iconic gesture generation. In *Intelligent virtual agents* (pp. 76–89).

Bernstein, N. A. (1967). *Coordination and regulation of movement*. New York: Pergamon Press.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1168–1178).

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non–native speakers. *Discourse Processes*, *23*(2), 119–147. http://dx.doi.org/10.1080/01638537709544986.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, 13–25.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. http://dx.doi.org/10.1016/j.cognition.2011.05.011.

Branigan, H. P., Pickering, M. J., Stewart, A. J., & McLean, J. F. (2000). Syntactic priming in spoken production: linguistic and temporal interference. *Memory and Cognition*, *28*, 1297–1302.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*, 1482–1493.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190.

Brown-Schmidt, S. (2009b). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cognitive Science*, *32*(4), 643–684. http://dx.doi.org/10.1080/03640210802066816.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue & Discourse*, *2*(1), 11–33.

Buder, E. H., & Eriksson, A. (1999). Time-series analysis of conversational prosody for the identification of rhythmic units. *Proceedings of the 14th international congress of phonetic sciences*, (Vol. 2, pp. 1071–1074). Retrieved from: http://www.ling.gu.se/~anders/papers/1071.pdf.

Busch, M. W. (2007). Task-based pedagogical activities as oral genres: a systemic functional linguistic analysis. *ProQuest*.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*(06), 657–674.

Cartwright, N. (1999). *The dappled world*. Cambridge, UK: Cambridge University Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893–910.

Chartrand, T. L., & Van Baaren, R. (2009). Human mimicry. *Advances in Experimental Social Psychology*, *41*, 219–274.

Chemero, A. (2009). *Radical embodied cognitive science*. MIT Press.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick (Ed.), *Perspectives on socially shared cognition* (pp. 127–149). Washington: American Psychological Association.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: evidence from the priming of noun- phrase structure. *Journal of Memory and Language*, *49*, 214–230.

Cooper, R. P., Catmur, C., & Heyes, C. (2012). Are automatic imitation and spatial compatibility mediated by different processes? *Cognitive Science*.

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, *19*(3), 226–228.

Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 155–179.

Dale, R. (2010). Critique of radical embodied cognitive science. *Journal of Mind and Behavior*, 31, 127-140.

Dale, R., & Louwerse, M. M. (2012). Human interaction as a multimodal network structure. *Presented at the Conceptual Structures, Discourse, and Language*.

Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*, 21, 1153-1161.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*(10), 441–447.

De Looze, C., & Rauzy, S. (2012). Measuring speakers' similarity in speech by means of prosodic cues: methods and potential (pp. 1393–1396). *Presented at the interspeech 2011*.

Di Paolo, E., & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience*, *6*. http://dx.doi.org/10.3389/fnhum.2012.00163.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, *33*, 1–40.

Dixon, J. A., & Stephen, D. G. (2012). Multi-scale interactions in dictyostelium discoideum aggregation. *Physica A: Statistical Mechanics and Its Applications*. Retrieved from: http://www.sciencedirect.com/science/article/pii/S0378437112006280.

Dixon, J. A., Stephen, D. G., Boncoddo, R., & Anastas, J. (2010). The self-organization of cognitive structure. *Psychology of Learning and Motivation*, *52*, 343–384.

Dressler, R. A., Buder, E. H., & Cannito, M. P. (2009). Rhythmic patterns during conversational repairs in speakers with aphasia. *Aphasiology*, *23*, 731–748.

Dumas, G., Chavez, M., Nadel, J., & Martinerie, J. (2012). Anatomical connectivity influences both intra- and inter-brain synchronizations. *PLoS One*, 7(5), e36414.
Duran, N. D., & Dale, R. (in press). Perspective-taking in dialogue as self-organization under social constraints. *New Ideas in Psychology*.
Duran, N. D., Dale, R., & Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*, *121*, 22–40.
Eliasmith, C. (1996). The third contender: a critical examination of the dynamicist theory of cognition. *Philosophical Psychology*, *9*(4), 441–463.
Eliasmith, C. (2012). The complex systems approach: rhetoric or revolution. *Topics in Cognitive Science*, *4*(1), 72–77. http://dx.doi.org/10.1111/j.1756-8765.2011.01169.x.
Engstrom, D. A., Kelso, J. A., & Holroyd, T. (1996). Reaction-anticipation transitions in human perception-action patterns. *Human Movement Science*, *15*(6), 809–832.
Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327.
Fowler, C. A., Richardson, M. J., Marsh, K. L., & Shockley, K. (2008). Language use, coordination, and the emergence of cooperative action. In A. Fuchs & V. Jirsa (Ed.), *Coordination: Neural, behavioral and social dynamics*. Springer.
Frank, T. D., Richardson, M. J., Lopresti-Goodman, S. M., & Turvey, M. T. (2009). Order parameter dynamics of body-scaled hysteresis and mode transitions in grasping behavior. *Journal of Biological Physics*, *35*(2), 127–147.
Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science*, *10*(5), 151–155.
Fusaroli, R., Abney, D. H., Bahrami, B., Kello, C., & Tylén, K. (submitted). Conversation, coupling and complexity: Matching Scaling Laws Predict Performance in a Joint Decision Task.
Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological science*, *23*(8), 931–939.
Fusaroli, R., Raczaszek-Leonardi, J., & Tylén, K. (in press). Dialogue as interpersonal synergy. *New Ideas in Psychology*.
Fusaroli, R., & Tylén, K. (2012). Carving language for social interaction: a dynamic approach. *Interaction Studies*, *13*, 103–123.
Fusaroli, R., & Tylén, K. (submitted). Individual behavior, interactive alignment or interpersonal synergy? A model-comparison study on linguistic dialog.
Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*(3), 378.
Galantucci, B. (2009). Experimental Semiotics: a new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, *1*, 393–410.
Galantucci, B., & Sebanz, N. (2009). Joint action: current perspectives. *Topics in Cognitive Science*, *1*, 255–259.
Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51.
Gallagher, R., & Appenzeller, T. (1999). Beyond reductionism. *Science*, *284*(5411), 79–79. http://dx.doi.org/10.1126/science.284.5411.79.
Gallagher, H., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, *16*(3 pt. 1). Retrieved from: https://pure.au.dk/ws/files/48455072/gallagher2002.pdf.
Gallese, V. (2008). Mirror neurons and the social nature of language: the neural exploitation hypothesis. *Social Neuroscience*, *3*(3–4), 317–333.
Gambi, C., & Pickering, M. J. (2011). A cognitive architecture for the coordination of utterances. *Frontiers in Psychology*, *2*.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.
Garrod, S., & Clark, A. (1993). The development of dialogue co-ordination skills in schoolchildren. *Language and Cognitive Processes*, *8*, 101–126.
Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181–215.
Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. http://dx.doi.org/10.1016/j.tics.2003.10.016.
Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*, 292–304.
Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accomodation*. New York: Cambridge University Press.
Goodwin, C. (2000). Action and embodiment within human situated interaction. *Journal of Pragmatics*, *32*, 1489–1522.
Goodwin, C. (2003). *Conversation and brain damage*. Oxford; New York: Oxford University Press.
Goodwin, C. (2011). Building action in public environments with diverse semiotic resources. *Versus*, 112–113.
Graesser, A. C., Swamer, S. S., & Hu, X. (1997). Quantitative discourse psychology. *Discourse Processes*, *23*(3), 229–263. http://dx.doi.org/10.1080/01638539709544993.
Grammer, K., Kruck, K. B., & Magnusson, M. S. (1998). The courtship dance: patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Nonverbal Behavior*, *22*(1), 3–29.
Gratier, M., & Devouche, E. (2011). Imitation and repetition of prosodic contour in vocal interaction at 3 months. *Developmental Psychology*, *47*, 67–76. http://dx.doi.org/10.1037/a0020722.
Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research*, *34*, 365–399. http://dx.doi.org/10.1007/s10936-005-6139-3.
Hakkani-Tur, D. (2009). Towards automatic argument diagramming of multiparity meetings. In *IEEE international conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009* (pp. 4753–4756). Retrieved from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4960693.
Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.
Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.
Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61. http://dx.doi.org/10.1016/S0749-596X(03)00022-6.
Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*, 27–39.
Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121. http://dx.doi.org/10.1016/j.tics.2011.12.007.
Healey, P. G. T., Howes, C., & Purver, M. (2010). Does structural priming occur in ordinary conversation? *Presented at the linguistic evidence*.
Healey, P. G. T., & Mills, G. (2006). Participation, precedence and co-ordination in dialogue. *Presented at the 28th annual conference of the science society*.
Horton, W. S. (2005). Conversational common ground and memory processes in language production. *Discourse Processes*, *40*(1), 1–35.
Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: knowing *when* and knowing *how* to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589–606.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. http://dx.doi.org/10.1016/j.cognition.2004.07.001.
Hutchins, E. (1995a). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265–288.
Hutchins, E. (1995b). *Cognition in the Wild*. Cambridge, Mass: MIT Press.
Hutchins, E., & Johnson, C. M. (2009). Modeling the emergence of language as an embodied collective cognitive activity. *Topics in Cognitive Science*, *1*, 523–546. http://dx.doi.org/10.1111/J.1756-8765.2009.01033.X.
Hutchins, E. (2010). Cognitive Ecology. *Topics in Cognitive Science*, *2*(4), 705–715.
Hutchins, E. (2011). Enculturating the Supersized Mind. *Philosophical Studies*, *152*(3), 437–446.
Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62. http://dx.doi.org/10.1016/j.cogpsych.2010.02.002.
Jefferson, G. (1988). On the Sequential Organization of Troubles-Talk in Ordinary Conversation. *Social Problems, 35*(4), 418–441.
Kauffman, S. (1996). *At home in the universe: the search for the laws of self-organization and complexity: the search for the laws of self-organization and complexity*. Oxford University Press.
Kello, C. T. (2013). Critical branching neural networks. *Psychological Review*, *120*, 230–254.
Kelso, J. A.S. (1981). On the oscillatory basis of movement. *Bulletin of the Psychonomic Society*, *18*(63), 9.
Kelso, J. A. (1995). *Dynamic patterns: The self-organization of brain and behavior*. The MIT Press.
Kelso, J. A.S. (2009). Synergies: atoms of brain and behavior. *Advances in Experimental Medicine and Biology*, *629*, 83–91. http://dx.doi.org/10.1007/978-0-387-77064-2_5.
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.
Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, 7(2), 46–50.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.
Kingstone, A., Smilek, D., Ristic, J., Friesen, C. K., & Eastwood, J. D. (2003). Attention, researchers! It is time to take a look at the real world. *Current Directions in Psychological Science*, *12*(5), 176–180. http://dx.doi.org/10.1111/1467-8721.01255.
Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological research on joint action: theory and data. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 59.
Knoblich, G., & Sebanz, N. (2006). The social nature of perception and action. *Current Directions in Psychological Science*, *15*(3), 99–104.
Konvalinka, I., & Roepstorff, A. (2012). The two-brain approach: how can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience. 6.*
Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, *52*(6), 587–597.
Kousidis, S., & Dorran, D. (2009). Monitoring convergence of temporal features in spontaneous dialogue speech. In *Conference papers* (p. 1). Retrieved from: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1003&context=dmccon.
Laidlaw, K. E., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, *108*(14), 5548–5553.
Latash, M. L. (2008). *Synergy*. Oxford; New York: Oxford University Press.
Latash, M. L., Scholz, J. P., & Schoner, G. (2007). Toward a new theory of motor synergies. *Motor Control*, *11*, 276–308.
Lee, C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., et al. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples (Vols. 793–796). *Presented at the interspeech*.

Lelong, A., & Bailly, G. (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes. In Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues (pp. 273–286). Springer Berlin Heidelberg.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*, 78–106.

Levinson, S. C. (1983). *Pragmatics*. Cambridge Cambridgeshire; New York: Cambridge University Press.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions (pp. 3081–3084). *Presented at the interspeech 2011*.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*(3), 550–557.

Louwerse, M. M., Benesh, N., Watanabe, S., Zhang, B., Jeuniaux, P., & Vargheese, D. (2009). The multimodal nature of embodied conversational agents. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1459–1463).

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1420.

Marsh, K. L., Richardson, M. J., Baron, R. M., & Schmidt, R. C. (2006). Contrasting approaches to perceiving and acting with others. *Ecological Psychology*, *18*(1), 1–38.

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, *1*, 320–339.

Marwan, N., Carmen Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*, 237–329.

McElree, B. (2006). Accessing recent events. *Psychology of Learning and Motivation*, *46*, 155–200.

Mehler, A., Weiß, P., Menke, P., & Lücking, A. (2010). Towards a simulation model of dialogical alignment. In *Proceedings of the 8th international conference on the evolution of language (Evolang8), 14–17 April 2010, Utrecht* (pp. 238–245).

Mills, G. (2011). The emergence of procedural conventions in dialogue. *Presented at the 33rd annual conference of the cognitive science society*.

Mills, G. (in press). Dialogue in joint activity: complementarity, convergence and conventionalization. *New Ideas in Psychology*.

Mills, G., & Gregoromichelaki, E. (2010). Establishing coherence in dialogue: sequentiality, intentions and negotiation. *Presented at the SemDial (PozDial)*.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Murray, G., Renals, S., Carletta, J., & Moore, J. (2006). Incorporating speaker and discourse features into speech summarization. In *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics* (pp. 367–374). : Retrieved from: http://dl.acm.org/citation.cfm?id=1220882.

Murray, L., & Trevarthen, C. (1985). Emotional regulation of interactions between two-month-olds and their mothers. *Social Perception in Infants*, 177–197.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Réserbat-Plantey, D. (1999). Expectancies for social contingency in 2 month olds. *Developmental Science*, *2*, 164–173.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678.

Newell, K. M., Broderick, M. P., Deutsch, K. M., & Slifkin, A. B. (2003). Task goals and change in dynamical degrees of freedom with motor learning. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 379–387.

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., Volman, I. A. C., Ruiter, J. P., Hagoort, P., et al. (2009). Recipient design in tacit communication. *Cognition*, *111*(1), 46–54.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Nishimura, R., Kitaoka, N., & Nakagawa, S. (2008). Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling (pp. 534–537). *Presented at the interspeech 2008*.

Oller, D., Niyogi, P., Gray, S., Richards, J., Gilkerson, J., Xu, D., et al. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(30), 13354.

Onnis, L., & Spivey, M. J. (2012). Toward a new scientific visualization for the language sciences. *Information*, *3*(1), 124–150.

Orsucci, F., Giuliani, A., & Webber, C. (2006). Combinatorics and synchronization in natural semiotics. *Physica A: Statistical Mechanics and Its Applications*, *361*, 665–676.

Orsucci, F., Giuliani, A., & Zbilut, J. (2004). Structure & coupling of semiotic sets. *AIP conference proceedings* (Vol. 742, pp. 83). : Retrieved from: http://pdfserv.aip.org/APCPCS/vol_742/iss_1/83_1.pdf.

Orsucci, F., Walter, K., Giuliani, A., Webber, C. L., & Zbilut, J. P. (1997). *Orthographic structuring of human speech and texts: Linguistic application of recurrence quantification analysis*. Arxiv Preprint cmp-lg/9712010.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2011). Phonetic convergence in college roommates. *Journal of Phonetics*.

Paxton, A., & Dale, R. Frame-differencing methods for measuring bodily synchrony in conversation. Behavior Research Methods, in press.

Paxton, A., & Dale, R. *Argument disrupts interpersonal synchrony*, submitted for publication.

Perkins, L., & Milroy, L. (1997). Sharing the communicative burden: a conversation-analytic account of aphasic/non-aphasic interaction. *Multilingua*, *16*(2–3), 199–215.

Perry, M. (2010). Socially distributed cognition in loosely coupled systems. *AI & Society*, 1–14.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, *3*, 136–141.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological Bulletin*, *134*, 427–459. http://dx.doi.org/10.1037/0033-2909.134.3.427.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

Pickering, M. J., & Garrod, S. (2009). Prediction and embodiment in dialogue. *European Journal of Social Psychology*, *39*(7), 1162–1168.

Pickering, M. J., & Garrod, S. An integrated theory of language production and comprehension. Behavioral and Brain Sciences, in press.

Platek, S. M. (2010). Yawn, yawn, yawn, yawn; yawn, yawn, yawn! The social, evolutionary and neuroscientific facets of contagious yawning. *Frontiers of Neurology and Neuroscience*, *28*, 107–112. http://dx.doi.org/10.1159/000307086.

Port, R. F., & Van Gelder, T. (1995). *Mind as motion: explorations in the dynamics of cognition*. Cambridge, Mass: MIT Press.

Ramenzoni, V. C., Davis, T. J., Riley, M. A., Shockley, K., & Baker, A. A. (2011). Joint action in a cooperative precision task: nested processes of intrapersonal and interpersonal coordination. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *211*, 447–457. http://dx.doi.org/10.1007/s00221-011-2653-8.

Ramenzoni, V. C., Riley, M. A., Shockley, K., & Baker, A. A. (2012). Interpersonal and intrapersonal coordinative modes for joint and single task performance. *Human Movement Science*. http://dx.doi.org/10.1016/j.humov.2011.12.004.

Reidsma, D., Heylen, D. K. J., & Ordelman, R. J.F. (2006). *Annotating emotions in meetings. LREC 2006*. Retrieved from: http://eprints.eemcs.utwente.nl/8381/.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637.

Reitter, D., & Moore, J. D. (2007). Predicting success in dialogue. *Presented at the proceedings of the 45th annual meeting of the association of computational linguistics* (Vol. 45, pp. 808–815).

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation (pp. 685–690). *Presented at the proceedings of the 28th annual conference of the cognitive science society*.

Richardson, M. J., Dale, R., & Marsh, K. L. Complex dynamical systems in social and personality psychology: theory, modeling and analysis. In handbook of research methods in social and personality psychology, in press.

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*, *33*(8), 1468–1482. http://dx.doi.org/10.1111/j.1551-6709.2009.01057.x.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, *26*, 867–891. http://dx.doi.org/10.1016/j.humov.2007.07.002.

Richardson, M. J., Marsh, K. L., & Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 62.

Richardson, M. J., Marsh, K. L., Schmidt, R. C., & Richardson, M. J. (2010). Challenging the egocentric view of coordinated perceiving, acting and knowing. *The Mind in Context*, 307–333.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295.

Richardson, D. C., Street, C. N. H., Tan, J. Y. M., Kirkham, N. Z., Hoover, M. A., & Cavanaugh, A. G. (2012). Joint perception: gaze and social context. *Frontiers in Human Neuroscience*, *6*. Retrieved from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3388371/.

Riley, M. A., Richardson, M. J., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in Psychology*, *2*, 38. http://dx.doi.org/10.3389/fpsyg.2011.00038.

Riley, M. A., & Van Orden, G. C. (2005). *Tutorials in contemporary nonlinear methods for the behavioral sciences*. : National Science Foundation.

Risko, E. F., & Kingstone, A. (2011). Eyes wide shut: implied social presence, eye tracking and attention. *Attention, Perception, & Psychophysics*, *73*(2), 291–296.

Roche, J., Jaeger, T. F., Dale, R., & Kreuz, R. J. Don't rush the navigator: disambiguating strategies are hard to establish but easier to maintain, submitted for publication.

Sacks, H., Jefferson, G., & Schegloff, E. A. (1995). *Lectures on conversation*. (Vol. 1). :Wiley. Online Library. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/9781444328301.fmatter/summary.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696–735.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychological Review*, *115*(1), 101.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. Hillsdale, N.J. New York: L. Erlbaum Associates. (distributed by the Halsted Press Division of John Wiley and Sons).

Schegloff, E. A. (1986). The routine as achievement. *Human Studies*, *9*, 111–151.

Schegloff, E. A. (2007). Sequence organization in interaction. *A primer in conversation analysis*. (Vol. 1). : Cambridge University Press.

Schmidt, R. C. (2007). Scaffolds for social meaning. *Ecological Psychology*, *19*(2), 137–151.

Schmidt, R. C., Bienvenu, M., Fitzpatrick, P. A., & Amazeen, P. G. (1998). A comparison of intra- and interpersonal interlimb coordination: coordination breakdowns and coupling strength. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 884–900.

Schmidt, R. C., Morr, S., Fitzpatrick, P., & Richardson, M. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*, *36*(4), 263–279. http://dx.doi.org/10.1007/s10919-012-0138-5.

Schmidt, R. C., & Richardson, M. J. (2008). Dynamics of interpersonal coordination. In *Coordination: Neural, behavioral and social dynamics* (pp. 281–308).

Schmidt, R. C., & Turvey, M. T. (1995). Models of interlimb coordination—equilibria, local analyses, and spectral patterning: comment on Fuchs and Kelso (1994). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 432–443. http://dx.doi.org/10.1037/0096-1523.21.2.432.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1–24.

Schober, M. F. (1995). Speakers, addressees, and frames of reference: whose effort is minimized in conversations about locations? *Discourse Processes*, *20*(2), 219–247.

Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: the role of the partner. In A. C. Graesser, M. A. Gernsbacher & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 211–232). Lawrence Erlbaum.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, *23*(4), 569–588.

Shalizi, C., Camperi, M., & Klinkner, K. (2007). Discovering functional communities in dynamical networks. *Statistical Network Analysis: Models, Issues, and New Directions*, 140–157.

Shintel, H., & Keysar, B. (2009). Less is more: a minimalist account of joint action in communication. *Topics in Cognitive Science*, *35*(2), 281–322.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Shockley, K., Santana, M.V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326–332.

Shteynberg, G. (2010). A silent emergence of culture: the social tuning effect. *Journal of Personality and Social Psychology*, *99*(4), 683.

Shteynberg, G., & Galinsky, A. D. (2011). Implicit coordination: sharing goals with similar others intensifies goal pursuit. *Journal of Experimental Social Psychology*, *47*(6), 1291–1294.

Sidnell, J., & Enfield, N. (2012). Language diversity and social action. *Current Anthropology*, *53*, 302–333.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 595–600). Retrieved from: http://idiom.ucsd.edu/~rlevy/papers/smith-levy-2008-cogsci.pdf.

Smith, M., & Wheeldon, L. (2001). Syntactic priming in spoken sentence production—an online study. *Cognition*, *78*, 123–164.

Spivey, M. J. (2007). *The continuity of mind*. Oxford, UK: Oxford University Press.

Spurrett, D., & Cowley, S. J. (2004). How to do things without words: infants, utterance-activity and distributed cognition. *Language Sciences*, *26*, 443–466.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10587–10592. http://dx.doi.org/10.1073/pnas.0903616106.

Szmrecsanyi, B. (2005). Language users as creatures of habit: a corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*, 113–149.

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken english: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin/New York: Mouton de Gruyter.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. http://dx.doi.org/10.1016/0010-0285(91)90016-H.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1105.

Tanenhaus, M., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Teufel, C., Fletcher, P. C., & Davis, G. (2010). Seeing other minds: attributed mental states influence perception. *Trends in Cognitive Sciences*, *14*(8), 376–382.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: The MIT Press.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235.

Truong, K. P., & Heylen, D. K. J. (2012). Measuring prosodic alignment in cooperative task-based conversations. *Presented at the interspeech 2012*.

Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(1), 3.

Turvey, M. T. (1990). Coordination. *The American Psychologist*, *45*, 938–953.

Turvey, M. T. (2007). Action and perception at the level of synergies. *Human Movement Science*, *26*(4), 657–697.

Tylén, K., Fusaroli, R., Bundgaard, P., & Østergaard, S. (2013). Making sense together: a dynamical account of linguistic meaning making. *Semiotica*, *194*, 39–62.

Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. *Mind & Language*, *25*, 3–29.

Van Orden, G., & Stephen, D. G. (2012). Is cognitive science usefully cast as complexity science? *Topics in Cognitive Science*, *4*, 3–6.

Van Overwalle, F. (2008). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, *30*(3), 829–858.

Van Rooij, I., Bongers, R. M., & Haselager, W. (2002). A non-representational approach to imagined action. *Cognitive Science*, 345–375.

Vaughan, B. (2011). Prosodic synchrony in co-operative task-based dialogues: a measure of agreement and disagreement. In *12th annual conference of the international speech communication association*.

Wagenmakers, E. J., Van der Maas, H. L. J., & Farrell, S. (2012). Abstract concepts require concrete models: why cognitive scientists have not yet embraced nonlinearly coupled, dynamical, self-organized critical, synergistic, scale-free, exquisitely context-sensitive, interaction-dominant, multifractal, interdependent brain-body-niche systems. *Topics in Cognitive Science*, *4*, 87–93.

Wang, Y., & Hamilton, A. F. C. (2012). Social top-down response modulation (STORM): a model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience*, *6*.

Warlaumont, A. S. (2012). A spiking neural network model of canonical babbling development. *Presented at the 2012 IEEE International Conference on Development and Learning (ICDL)*.

West, B. J., Geneston, E. L., & Grigolini, P. (2008). Maximizing information exchange between complex networks. *Physics Reports*, *468*, 1–99.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183–194.

Wilkinson, R., Beeke, S., & Maxim, J. (2003). Adapting to conversation. In C. Goodwin (Ed.), *Conversation and brain damage* (pp. 59–89). : Oxford University Press.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*, 957–968.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 593–602.

Wu, S., & Keysar, B. (2007a). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*(1), 169–181.

Wu, S., & Keysar, B. (2007b). The effect of culture on perspective taking. *Psychological science*, *18*(7), 600–606.

Wyatt, D., Bilmes, J., Choudhury, T., & Kitts, J. A. (2008). Towards the automated social analysis of situated speech data. In *Proceedings of the 10th international conference on ubiquitous computing* (pp. 168–171). Retrieved from: http://dl.acm.org/citation.cfm?id=1409658.

This page intentionally left blank

CHAPTER THREE

Conceptual Composition: The Role of Relational Competition in the Comprehension of Modifier-Noun Phrases and Noun–Noun Compounds

Christina L. Gagné[1], Thomas L. Spalding
Department of Psychology, University of Alberta, Edmonton, AB, Canada
[1]Corresponding author: E-mail: cgagne@ualberta.ca

Contents

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00003-4

Abstract

Compositionality and productivity, which are the abilities to combining existing concepts and words to create new concepts and phrases, words, and sentences, are hallmarks of the human conceptual and language systems. Combined concepts are formed within the conceptual system and can be expressed via modifier-noun phrases (e.g. *purple beans*) and compound words (e.g. *snowball*), which are the simplest forms of productivity. Modifier-noun phrases and compound words are often paraphrased using a relation to connect the constituents (e.g. *beans that are purple, ball made of snow*). The phrase or compound does not explicitly contain the underlying relation, but the existence of the relation can be shown by manipulating the availability of the relation and observing the effect on the interpretation of the phrase or compound. This chapter describes how novel modifier-noun phrases and established compounds are interpreted. We present a theoretical account of relational interpretation of combined concepts and present the empirical evidence for the use of relational structures. We then present the empirical evidence supporting our theoretical account's specific predictions about how relational interpretations are selected and evaluated and how the relational interpretation is elaborated to create a fully specified new concept.

1. INTRODUCTION

Concepts are mental entities that allow a person's experience with the world to be organized into meaningful units. The ability to store and access conceptual information is vital to human cognition because many tasks that people perform on a daily basis (such as classification, prediction, reasoning, and communication) require the application of existing knowledge to new situations. Indeed, concepts have been called the "building blocks of cognition" (Solomon, Medin, & Lynch, 1999). Concepts are not represented in isolation but rather are part of a complex and highly interrelated conceptual system. Relational information plays a vital role in the organization of the conceptual system (for an overview, see Medin, 1989) as well as a central role in nominal compounding (e.g. Allen, 1978; Gleitman & Gleitman, 1970; Kay & Zimmer, 1976; Levi, 1978). In particular, knowledge about how concepts interrelate underlies conceptual composition, which is a process through which new concepts are created by combining existing concepts.

In English, combined concepts are often expressed as modifier-noun phrases (e.g. *chocolate recipe*) or, as they become more commonly used, as compound words (e.g. *handbag*). Teall (1892, p. 5) made the observation that "in the English language it is very common to name a thing, or express an attribute, or assert an action or manner of action by omitting minor or connecting parts of a full expression, and using only the principal elements

in more or less arbitrary association and frequently in inverted order". For example, *box for hats* can be expressed as *hat box*. In other words, linguistic expressions of combined concepts typically omit the implicit underlying relation. However, there is a long tradition in linguistic work on compounding that points to the involvement of these underlying relations, and combined concepts can often be paraphrased by expressing the underlying implicit relation (e.g. *flu virus is a virus that causes flu*; Allen, 1978; Bisetto & Scalise, 2005; Downing, 1977; Lees, 1960; Levi, 1978; Marchand, 1969; Scalise & Bisetto, 2009; Warren, 1978). As Kay and Zimmer (1976, p. 29) point out, "… there is a relation between the two nouns that the hearer must supply" to comprehend the phrase. These relation-based modifier-noun constructions are in contrast to copulative constructions (e.g. *singer–songwriter*), which some linguists argue have a dual head structure in which the two constituents are linked via the conjunctive *and* (Bauer, 2008; Bisetto & Scalise, 2005; Booij, 2005; Olsen, 2004; Olsen & Van der Meer, 2001; Scalise & Bisetto, 2009).

Psychological research examining the processing of modifier-noun phrases and compounds can provide useful insight into the nature and use of relational information within the conceptual system. Indeed, researchers interested in conceptual combination have investigated the role of relational information, although they vary in terms of the relative importance placed on the use of this information. Some have suggested that relation processing is one of two processes that can be used (Estes, 2003; Wisniewski, 1996, 1997). Wisniewski (1996), for example, argues that combined concepts are interpreted by one of two fundamentally different processes: relation linking or property mapping. Others (e.g. Gagné & Shoben, 1997; Spalding, Gagné, Mullaly, & Ji, 2010) have argued that relation processing is the initial process and that the derivation of properties follows the selection of a relation. That is, relation linking and property derivation are not mutually exclusive. Instead, the relation plays a key role in determining which properties can be inferred (Gagné, 2000; Gagné & Murphy, 1996; Spalding et al., 2010). However, regardless of one's theoretical perspective on conceptual combination, it has been acknowledged that the issue of relational information is highly relevant for developing a deeper understanding of the processing of combined concepts.

This chapter reviews the literature on relational interpretation of combined concepts and describes how people interpret and understand modifier-noun phrases and compound words. In particular, we focus on how relational interpretations arise and compete with each other during

comprehension. We begin by describing our theory of relational interpretation of combined concepts in general terms, in particular focusing on the somewhat different roles of the modifier and head constituents. We then describe the evidence for relational interpretations and for the nature of relations and relational competition during the interpretation process. As it is important to understand the role of relational competition within the larger context of the full interpretation process, we end by discussing how the cognitive system evaluates and then elaborates possible relational interpretations.

2. MODIFIER-NOUN PHRASES AND COMPOUNDS AS EXPRESSIONS OF COMBINED CONCEPTS

Modifier-noun phrases and compound words can be used to express combined concepts. Both types of linguistic expressions represent, in some sense, the least complex way of combining words of human languages, in that they are the midpoints between single words and sentences. In short, they constitute the beginning of linguistic productivity.

Because both modifier-noun phrases and compound words are linguistic representations of combined concepts, a commonality of structure and processing between the two kinds of linguistic representation suggests itself, namely that interpretation and access to the underlying meaning might involve an attempt at computing a meaning involving the constituents. On the one hand, this might seem reasonable as compound words generally enter the language as novel modifier-noun phrases and only later take on the status of a known compound word (at least in English; see Downing, 1977; Libben & Jarema, 2006). Thus, given that novel modifier-noun phrases require computed meanings, it might not be surprising that something about that meaning computation is maintained even after the phrase enters the language as a compound word. On the other hand, from a linguistic perspective, it might be thought quite surprising that such processing would survive for compounds that are well known as it is generally believed that it is easier to remember a meaning than to compute it (see Bertram, Laine, & Karvinen, 1999, for an example of this claim with respect to morphologically complex words). Furthermore, it might be thought strange from a conceptual perspective: the meaning of the known compound has to have been computed previously, probably many times; so it could be stored as a separate concept that just happens to be linguistically identified by a compound word.

However, within the compound word literature, there is growing evidence of meaning construction during the access and interpretation of compound words (Fiorentino & Poeppel, 2007; Gagné & Spalding, 2009; Gagné, Spalding, Figueredo, & Mullaly, 2009; Inhoff, Radach, & Heller, 2000; Ji, Gagné, & Spalding, 2011; Koester, Gunter, & Wagner, 2007). That is, despite the fact that compound words are known, and at least potentially stored in some form of mental lexicon, meanings based on the constituent words are constructed during processing. In short, the processing of a compound word is highly sensitive to its constituent structure. In fact, as we will discuss in Section 6, the meaning computation process seems to be initiated even for opaque compounds (e.g. *hogwash*), those for which the actual meaning cannot be computed from the constituents. Given the commonality of underlying structure, and the evidence that compound words undergo a meaning construction process during comprehension, we believe that theories of conceptual combination must apply to both novel and familiar combinations.

3. THEORETICAL FRAMEWORK: A THREE-STAGE THEORY OF CONCEPTUAL COMBINATION

The competition among relations in nominals (CARIN) theory of conceptual combination claims that relational information plays a critical role in the processing of novel modifier-noun phrases (Gagné & Shoben, 1997; Spalding & Gagné, 2008) and was the first primarily relational psychological theory of conceptual combination. In general, information about how objects, people, and so on, interact in the world is used to select a relation that links the constituent concepts during the formation of a new concept. In particular, according to this theory, the availability of relational information varies from constituent to constituent, and the availability affects the ease with which combined concepts can be processed. For example, people know that when *chocolate* is used as a modifier, the compound can usually be paraphrased using a MADE OF relation but that other relations such as FOR and ABOUT are also possible. Consequently, some relations are more readily available than others, and this difference in availability influences the time required to interpret a compound. Another key claim of this theory is that relations compete with each other for selection. Consequently, it should be easier to interpret a compound that requires a relation that is highly available (i.e. a relation that is a strong competitor) than to interpret a compound that requires a relation that is less available

(i.e. a relation that is a weak competitor). Moreover, changing the availability of a relation (by manipulating prior context, for example) should affect the ease of comprehension. According to the CARIN theory, relation selection is most heavily influenced by relational information pertaining to the modifier concept.

The relational interpretation competitive evaluation (RICE) theory (Spalding et al., 2010) is an extension and refinement of the CARIN theory. According to RICE, the interpretation of a compound is obtained through a "suggest evaluate elaborate" process (Figure 3.1).

Like CARIN, the RICE theory posits that multiple relational structures are constructed and evaluated as possible interpretations. These structures compete with one another, and ease of interpretation depends on how quickly a relation structure can be identified as the most likely candidate. The RICE theory distinguishes between the assignment of a constituent to a particular morphosyntactic role (e.g. to either the modifier or head noun role) and the selection of a relational structure. According to this theory, relational information about the constituents is accessed in the context of the constituents' morphosyntactic roles. Furthermore, the search for suitable relations is triggered after the constituents have been assigned to their respective morphosyntactic roles, and relation availability is associated with the concepts in their particular role (i.e. the relational availability for snow in *snow hill* is based on snow's previous use as a modifier and is unaffected by snow's history of use as a head). As in the CARIN theory, the relations are initially suggested with a strength proportional to the relations' availability for the modifier, and then the appropriateness of the suggested

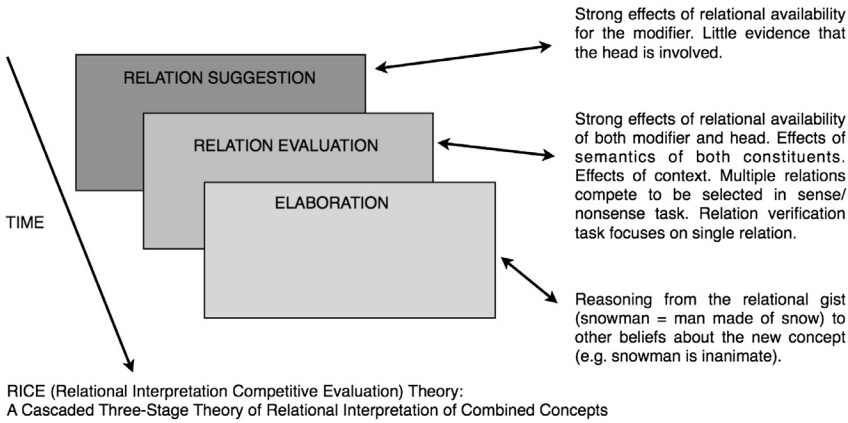


Figure 3.1 Schematic of the RICE (Relational Interpretation Competitive Evaluation) theory.

relations is evaluated. However, the RICE theory is more comprehensive than the earlier theory, in that it more clearly identifies the roles played by the modifier and head noun and suggests that the constituents are differentially involved in relation suggestion and evaluation. In particular, the modifier is more heavily involved in relation suggestion, while information associated with both constituents is heavily involved in the relation evaluation phase of interpretation.

In addition, the RICE theory removes CARIN's claim that relation availability is associated only with the modifier concept. Unlike CARIN, RICE proposes that relational interpretations are evaluated using semantic and pragmatic information, but also relational information, associated with both the modifier and the head concepts. Thus, as with CARIN, an important aspect of the evaluation process is the ability of the constituents to function as arguments for a given relation structure. In particular, a constituent must fit the entailments required to fulfill a particular function within a particular relation. Thus, the constituent *snow* can function as a modifier in the MADE OF relation because it is a material. Likewise, the constituent *planet* can function as a head in a LOCATIVE relation because planets can be in a physical location. However, although each constituent must fit the restrictions of its respective role, the restrictions are codetermined. Thus, the MADE OF relation, when actually paired with a particular head noun, requires that the modifier be a material, but not just any material. It must be a material that is appropriate for the head noun. *Snow sculpture* satisfies these restrictions, but *snow hospital* does not. This aspect of the evaluation process relies on world knowledge. Gagné (2002) and Spalding and Gagné (2008) use the example of *mountain planet* to illustrate the use of this type of knowledge. The interpretation "planet-LOCATED mountain" is rejected because, to name just one pragmatic restriction, planets are too large to be located in the mountains. Levi (1978) refers to this type of knowledge as extralinguistic knowledge, and she outlines several semantic and pragmatic considerations that are used to determine the contextually most plausible reference for a given compound (for further discussion concerning the use of extralinguistic information, see Downing, 1977; Finin, 1980; Levi, 1978; Meyer, 1993; Štekauer, 2005, 2006, 2009). As mentioned above, unlike the CARIN theory, however, RICE also assumes that relational information about the head plays a role during this evaluation stage. In particular, in addition to the semantic and pragmatic factors listed above, a relation should be more easily (and positively) evaluated when it is commonly used with the head constituent.

Finally, once a suggested relation has been selected and evaluated as appropriate, there is an elaboration stage in which the full meaning of the compound is developed. The relational gist interpretation (e.g. man MADE OF snow for *snowman*) must then be elaborated so that a snowman is known to be cold rather than warm. This information is not explicitly in the gist but is derived based on that gist. Importantly, the nature of the elaboration depends on the relational gist and is not derived just from the contents of the modifier and head concepts: a man made of snow is inanimate and white, but a man for snow is animate and likely has a shovel. Clearly, the gist makes a difference in how the elaboration proceeds. One interestingly open point concerns the extent to which the relational gist is elaborated and whether elaboration proceeds automatically when the combined concept is interpreted or only when needed (see Spalding et al., 2010 and Gagné & Spalding, 2011, for related discussions).

4. EVIDENCE OF THE MODIFIER'S ROLE IN RELATION SUGGESTION

Both the CARIN and the RICE theories predict that relation availability for the modifier influences conceptual combination. The investigations of the role of relation availability have generally focused on two aspects of experience with relations that were expected to affect relation availability. First, a person's experience with the ways in which a given constituent tends to be used in combinations, and in particular what relations tend to be used with that constituent, should affect how easily they can understand new combinations. We refer to this factor as general usage. Second, the ease of interpretation should be affected by what relations have most recently been used with the constituent. We refer to this factor as recent usage. Both streams of investigation point to the modifier constituent playing a role in relation suggestion.

4.1. General Usage

One factor that influences the competitiveness of a particular relation is knowledge about how likely it is to be used with the modifier constituent. For example, because *paper* is often used in a compositional sense (MADE OF) when it is in the modifier position, this relation will be a stronger competitor than will other relations. Thus, a combination *paper X* will be interpreted more quickly when the phrase requires the MADE OF relation than when it requires a relation that is a weaker competitor. This prediction

was initially tested using information about which relations are most likely to be used with constituents of novel combinations. To estimate the competitiveness of various relations for particular constituents, Gagné and Shoben (1997) created a set of potential novel compounds by crossing 91 modifiers and 91 head nouns and each pairing was evaluated in terms of whether it had a sensible literal interpretation. Each of the 3239 sensible phrases were classified in terms of relational categories that were based on Levi's (1978) set of thematic relations. Levi's categories are general and are based on her hypothesis that all complex nominals are derived from underlying semantic structures from which a predicate (e.g. CAUSE, HAVE, MAKE, FOR) is deleted in the linguistic expression.

Next, the frequency with which each modifier and head noun appeared with the various relations was determined. For example, of the 38 phrases in which plastic was used as a modifier, 28 used the noun MADE OF modifier relation, seven used the noun ABOUT modifier relation, two used the noun DERIVED FROM modifier relation, and one used the modifier CAUSES noun relation. These distributions were used to identify which relations were highly competitive and which were less competitive for an individual modifier or head noun. Relations were considered a good competitor for a particular constituent if that relation was among the set of relations that was used for the majority (60% or more) of all phrases using the constituent in question. All other relations were considered weaker competitors.

To evaluate whether ease of comprehension was affected by relation competitiveness, three experimental conditions were created. HH (High-High) items had an underlying relation that was among the set of highly competitive relations for both the modifier and the head, HL (High-Low) items were used a relation that was among the set of highly competitive relations for the modifier only, and LH (Low-High) items used an underlying relation that was among the set of highly competitive relations for the head noun only. In Experiment 1, the individual words were controlled in that the identical words were used in all three experimental conditions. In Experiment 2, the relation was controlled such that there was an equal number of each relation type in each condition. Nonsense filler items (e.g. *plastic rain* and *scarf soda*) were included. Each item was presented on a computer screen, and participants indicated, by pressing a key, whether the item had a sensible literal interpretation.

Both experiments demonstrated that it is easier to indicate that a phrase has a sensible interpretation when the underlying relation is highly competitive for the modifier than when it is a weak competitor; the HH and HL conditions yielded faster response times than did the LH condition.

The results also indicated that the competitiveness of the relation with respect to the head noun constituent did not influence response time; the HH and HL conditions did not differ from each other. These two findings were observed both in the analysis of variance (ANOVA), which used a dichotomous measure (H vs L) of a relation's competitiveness, as well as in a regression analysis, which used a continuous measure of a relation's competitiveness.

Although there has been concern that the results reported in Gagné and Shoben (1997) were because of the familiarity and plausibility of the phrase rather than because of relation availability (Wisniewski & Murphy, 2005), regression analyses show this not to be the case (Gagné & Spalding, 2006). Objective familiarity (i.e. frequency) did not differ across the three conditions. Also, even after subjective familiarity and plausibility were statistically controlled, items with highly competitive relations took less time to interpret than did items with less competitive relations. In addition, Gagné and Spalding (2006) found that subjective familiarity ratings were affected by relation availability; noun–noun phrases were more likely to be viewed as familiar when preceded by an item with the same relation than by an item with a different relation. If subjective familiarity were the causal factor, it would not have been influenced by the manipulation of relation availability. Similarly, sense/nonsense judgments (in which the person is instructed to respond "sense" if they can think of a plausible meaning for the phrase) are also affected by manipulations of relational availability (e.g. Gagné, 2001). Hence, neither familiarity nor plausibility are reasonable explanations for the results of Gagné and Shoben (1997).

Another concern has been that relation frequencies based on Gagné and Shoben's (1997) relational distributions might not be consistent with relation frequencies calculated using an actual corpus (Maguire, Devereux, Costello, & Cater, 2007). Contrary to this claim, Spalding and Gagné (2008) found that there is a close relationship ($r = 0.87$) between the relation frequencies derived by Gagné and Shoben and those based on the British National Corpus.

A third concern might be that relational effects such as those demonstrated by Gagné and Shoben (1997) would only occur when the combinations are interpreted out of context. Gerrig and Bortfeld (1999), for example, have argued that interpreting combined concepts in isolation could lead to a distorted view of the process of interpretation. Gagné and Spalding (2004a), however, showed that the modifier-based relation frequency effects are maintained, even when the combined concepts are embedded in a supportive context. Novel modifier-noun phrases were

placed in either a neutral context (i.e. one that did not provide the interpretation for the combination) or a supportive context (i.e. a context that implied the interpretation). Phrases were read more quickly in a supportive context than a neutral context, but the supportive context did not eliminate the effect of modifier relation frequency; sentences containing a combination using a high-frequency relation were read more quickly than were sentences containing a combination using a lower frequency relation. These findings suggest that contextual information might be used in the verification of potential interpretations but that it does not override the effect of competing relations. That is, the context might suggest possible referents for the phrase and that information might be used to evaluate which of the suggested relations is most feasible in the current context.

In sum, the strength of a relation for the modifier constituent of a novel phrase affects the ease of processing that phrase, in or out of context, and even when other factors such as familiarity are controlled.

4.2. Recent Usage

Another factor that influences a relation's competitiveness is recent experience with the constituent concepts. Gagné (2001) and Gagné and Shoben (2002) found a robust relational priming effect: it takes less time to make a sense/nonsense judgment to a novel combination (e.g. *student vote*) when it is preceded by a combination with the same modifier and same relation (e.g. *student accusation*) than when preceded by a combination using a different relation (e.g. *student car*). Gagné (2002) also demonstrated relation priming when the modifier was semantically similar in the prime and target combinations (e.g. *scholar accusation* as a prime for *student vote*). These findings suggest that processing the prime phrase alters the competitiveness of the relation used to interpret that phrase, and this change in the relative availability of the relation influences the ease of processing a subsequent phrase using the same modifier. However, a relation-priming effect was not observed when the head noun was in common between the prime and the target (Gagné, 2001, 2002; Gagné & Shoben, 2002), except when the target combination had two equally plausible relations (Gagné & Shoben, 2002), so that the evaluation was primarily a choice between those two relations (this aspect of the evaluation phase is discussed further in Section 7).

The ease of processing lexicalized compounds is similarly affected by recent exposure to a prime containing the same first constituent. In Gagné and Spalding (2004b), the target items were established compounds that had been selected from the Brown corpus (Francis & Kučera, 1982).

Prime items were constructed for each target compound (e.g. *snowball*); the same-relation prime (e.g. *snowfort*) used the same relation (MADE OF) as the target and the different-relation prime (e.g. *snowshovel*) used a different relation. As was the case for novel compounds, sense/nonsense judgments and lexical decision latencies were faster when the target had been preceded by the same-relation compound than when it was preceded by the different-relation compound. This relation-priming effect in established compounds has been replicated several times (e.g. Gagné et al., 2009; Spalding & Gagné, 2011). Relational priming was also observed when the primes were restricted to existing compounds (Gagné & Spalding, 2009); so the relational priming effect for compounds is not limited to experiments in which a large number of novel combinations are presented. In sum, recent relational usage affects the interpretation of both novel and familiar combinations, strongly suggesting a continuity of processing between novel and familiar combinations.

5. THE NATURE OF RELATIONS AND THE NATURE OF RELATIONAL COMPETITION

The research reviewed in the preceding sections provides strong evidence that relational information affects the interpretation of combined concepts, whether novel or familiar, whether modifier-noun phrases or compound words. In this section, we discuss in more depth the nature of the relations used for relational interpretation and the nature of relational competition.

5.1. Nature of Relations

The RICE theory makes three main assumptions about the relations.

5.1.1. Relational Information is Accessed via Conceptual Representations

First, RICE assumes that the relations used in relational interpretations are in some way accessed through or connected with the conceptual representations, rather than the lexical representations, of the constituents. Because relational information is associated with the conceptual representations, similar relational effects should be likely to occur across (at least some) languages. Indeed, the influence of relation availability has been demonstrated with Mandarin (Ji & Gagné, 2007) and Indonesian (Storms & Wisniewski, 2005) compounds and phrases, in addition to English. Furthermore, Gagné (2002)

showed relation priming when the modifier constituent of the prime combination was highly semantically similar to the modifier constituent of the target, in addition to when the modifier constituents were identical. Hence, identical lexical items are not necessary to demonstrate relational priming. However, when the prime modifier constituent is semantically unrelated to the target, Gagné (2001) found no relational priming (see also Gagné, Spalding, & Ji, 2005, for several failures to find relational priming when the prime and target modifier constituents are unrelated). In addition, Gagné et al. (2009) investigated relation priming when the repeated constituent was either in the same position between prime and target or in a different position (e.g. a given constituent served as the modifier for both prime and target or moved from the modifier position in the prime to the head position in the target). Relational priming was observed only when the repeated constituent was in the same position in the prime and target. Thus, relational access appears to depend on both the concept underlying the constituent and the role (i.e. modifier or head) that the concept is playing in the combination.

5.1.2. Representation of the Relations

Second, although RICE assumes that access is in some way dependent on the constituents, it makes no claim with respect to the representation of the relations themselves. Because relations are inherently parts of other structures, it is not clear whether the relations that are used in relational interpretations require separate representations or whether they are recovered for use from their existence within existing relational interpretations. Clearly, people have representations of semantic information that is similar to these relations. For example, people must have a concept of causality, but the question is whether this separate concept of causality is necessarily implicated in the activation of the relation CAUSE during conceptual combination. Although there have been claims that the relations are separate representations, and not tied to any constituent, the demonstrations of general relational effects have not been convincing. If the relations were represented separately, and not accessed via constituents, one would expect general differences in ease of processing, based on general characteristics of the relations, such as their overall frequency or semantic complexity. Shoben and Medin (as reported by Shoben, 1991) were among the first to investigate relational effects in conceptual combination. They proposed that relations differ in complexity and that this variation in complexity should influence ease of processing. For example, causal relations (noun CAUSES

modifier; modifier CAUSES noun) appear to have more primitives than the MADE OF relation. However, Shoben and Medin found no support for this complexity hypothesis: there were no systematic differences in processing time that could be attributed to relational complexity.

Estes (2003) proposed that the relations have separate representations and, consequently, they could be primed when the constituents of the primes and targets were completely unrelated. Estes (2003) claimed to demonstrate relation priming without access via the constituents. However, Gagné et al. (2005) showed that the relational effect observed in Estes (2003) was confounded with an extremely strong semantic priming effect. For example, *swimming flippers* primed *rugby shoes* more than did *road construction*. However, it is clear that *swimming* and *rugby* are both sports, and *flippers* and *shoes* are also obviously semantically related. Thus, the same-relation primes were semantically similar to the targets, in addition to being relationally similar. This is problematic for Estes' conclusion because Gagné (2002) had previously demonstrated relation priming following the presentation of a prime containing a semantically similar modifier; so it was already clear that lexically identical constituents were not required for relation priming. Indeed, as Gagné et al. showed, this semantic similarity between primes and targets was common in Estes' materials, leading to a highly significant difference in semantic similarity between the same-relation pairs and the different-relation pairs. Gagné et al. failed to find relational priming in four experiments when the prime-target semantic similarity was controlled. This outcome is consistent with the results of Experiment 6 of Gagné (2001) in which relation priming was not observed when only the relation, but neither constituent, was in common between the prime and the target, as well as the finding that repeating only the head constituent was not sufficient to result in relation priming (Gagné, 2001, 2002; Gagné & Shoben, 2002).

In subsequent work, Estes and Jones (2006) claimed to have demonstrated relation priming without access via the constituents, while controlling for semantic similarity between the primes and the targets. However, they used only one relation in all their experimental target items; thus, an unusually large proportion of the "sense" target items were a MADE OF relation, while the filler items were of many different relations. Hence, a problem arises with this data. If the prime is a MADE OF, then the target item was more likely to require a "sense" response, than if the prime has some other relation. Being able to predict the correct response to the target from the prime is, of course, a critical problem in a priming paradigm

(Estes and Jones did have sense–sense filler items, but they varied in relation; so they did not provide a complete control for this particular problem). Thus, again, the purported general relation-priming effect is likely not because of relation priming, per se. We should perhaps note here that some degree of facilitation may occur with only a repeated relation under some circumstances because the relations do carry semantic information and semantic information can affect relational interpretation (e.g. Spalding & Gagné, 2007). However, it appears that relation priming without at least strong similarity of the constituent concepts is unlikely in the standard relational priming paradigm as a number of attempts to demonstrate such priming have failed, as discussed above.

5.1.3. Level of Relational Abstraction

Third, the RICE theory is not committed to a particular set of relations, nor to relations at a particular level of abstraction. For the purposes of making empirical predictions, the set of relations used in experiments meant to test the theory (e.g. in constructing the experimental materials) have been based on Levi (1978). The theory, however, is not reliant on relations at any particular level of generality; the experimental results indicate that this set of relations at this level of abstraction is sufficiently specific to give rise to reliable and consistent effects, which is the most important consideration in terms of testing the theory. In general, we suggest that relations are hierarchically organized (as are concepts themselves) in level of generality. Thus, chihuahuas are not identical to beagles, but both are still usefully characterized as dogs as they share much of their meaning both with each other and with other dogs. Similarly, there might be subrelations of a particular relation such that, for example, HAS-PART and HAS-POSSESSION are both examples of a HAS relation. It may be that relations, like concepts, have a preferred level of abstraction. Another possibility is that the initial interpretation may be at a more abstract level, and the relation may be refined and made more specific as the combined concept is elaborated. These speculations await empirical test.

5.2. Nature of Relational Competition

5.2.1. Is Competition Because of Number of Competitors?

The evidence discussed thus far indicates that relations compete for selection. But there are many forms of competition. The CARIN theory (Gagné & Shoben, 1997) mathematically instantiated competition by using a ratio that has the frequency (represented as a proportion) of the selected relation in

the numerator and the sum of the frequency of the three strongest competitors and the selected relation in the denominator (Gagné & Shoben, 1997). An exponential function is applied to each proportion. Although Maguire et al. (2007) have suggested that this ratio does not embody competition, Spalding and Gagné (2008) have demonstrated that the ratio makes sharp distinctions between items on the basis of the number of competing relations that are stronger than the selected relation and that these distinctions become sharper with increasing numbers of competitors. The original analysis (in Gagné & Shoben, 1997) found that including more than the three strongest relations in the denominator of the strength ratio did not improve the fit of the model; therefore, the number of stronger competitors was based on the three strongest relations in a given modifier's relational distribution. Spalding and Gagné showed that three mathematical models that instantiate sensitivity to number of stronger competitors fit the data better than a model that includes only frequency of the selected relation. In addition, a model that simply includes the number (0–3) of stronger competitor relations fits the Gagné and Shoben data as well as the original Gagné and Shoben mathematical model. The model that simply has the number of stronger competitors also fits the Gagné and Shoben (1997) data as well as (or better than) the rank of the selected relation (Experiment 1, $r = 0.47$ for rank and $r = 0.45$ for number of competitors, and Experiment 3, $r = 0.25$ for rank and $r = 0.36$ for number of competitors). Thus, as in Gagné and Shoben's analyses and models, including more stronger competitors (i.e. ranks beyond 3) did not improve the fit of the model.

If the process of suggesting, evaluating, and eliminating relations is serial, then using more competitors (i.e. rank) should have been a much better predictor than just the number of stronger competitors as it should have been able to pick up any variance associated with the competitors beyond the third. However, the analysis shows that competitors beyond the third have little impact on ease of processing. Overall, then, the data are consistent with a parallel process in which the number of strong competitors matters greatly, but the number of weak competitors (even if still stronger than the required relation) does not matter as much. This might be because such weak competitors require very little negative evidence from the head to rule them out, while strong competitors require substantial negative evidence from the head. As shown in Spalding et al. (2010) and discussed below, the head plays a large role in the evaluation of suggested relations, and perhaps the relations beyond the third competitor have little impact because the information from the head has been brought online by this

point in processing, supporting the required relation. Thus, perhaps low-frequency relations (for either constituent) are hard to accept and easy to reject. This pattern would also make sense on a parallel processing account in which the amount of activation to a given relation is a function of its frequency for the modifier and in which the difficulty of ruling out an activated relation is related to its level of activation.

5.2.2. Is Competition due to Inhibition or Facilitation?

The relational priming results described in Section 4.2 give rise to an interesting question about the nature of the priming effect and thus about the nature of relational competition effects in general. The differences in response times following the same-relation prime conditions vs the different-relation prime conditions could have two sources. First, relation priming could be because of the availability of the required relation being increased because of recent exposure to a combination using that relation. Second, it could be because of the different-relation prime becoming more available and thus becoming a more successful competitor, which then slows the selection of the required relation. The fact that the effect of relational availability seems to be driven by strong competitors to the required relational interpretation (Spalding & Gagné, 2008) suggests that one might expect that the relation-priming effect would be primarily because of the activation of a competing relation by the prime rather than because of facilitation of the target relation.

To examine this issue, Spalding and Gagné (2011) replicated the relation-priming effects, but included a baseline modifier-only condition along with the same-relation and different-relation conditions. For example, for the target *snowball* (ball MADE OF snow), the baseline prime was *snow*, the same-relation prime was *snowfort* (fort MADE OF snow) and the different-relation prime was *snowshovel* (shovel FOR snow). Presenting only the modifier should activate the modifier's relational distribution, but will not activate any particular relation above its normal baseline within the distribution. In all three conditions, the modifier is seen and thus this design controls for priming because of lexical repetition from the prime to the target.

If the effect is pure facilitation, then the same-relation condition should be fast and the different-relation condition and the baseline should both be slow. If the effect is due purely to inhibition, then the different-relation condition should be slow and the same-relation condition and baseline both should be fast. If there is both facilitation and inhibition, then the same-relation condition should be fast, the different-relation condition should be

slow, and the baseline should be in the middle. The results were consistent with inhibition. In both experiments, recent presentation of a compound using a different relation made it more difficult to interpret the target compound, whereas responses in the same-relation and modifier-only prime conditions did not differ.

5.2.3. Is There Competition among Alternative Interpretations?

In the previous sections, our operationalizations of relation competition have been based on the relations' use with individual constituents. However, competition can also be defined in terms of alternative relational interpretations at the level of the whole phrase/compound. According to the RICE theory, language users must identify a particular interpretation when a phrase or compound is encountered, but in doing so, multiple interpretations are considered. These alternative interpretations influence ease of processing. Gerrig (1989) measured the interpretation time of familiar compounds (e.g. *foot race*) that were presented in a discourse context that required either the conventional interpretation (e.g. "a race on foot") or an innovative interpretation (e.g. "a race the length of King Louis' foot"). The results indicated that the conventional meaning interfered with the derivation of an innovative meaning, in that phrases that were more readily interpreted with their conventional reading were more difficult to interpret with the innovative reading.

While it may be unsurprising that an established meaning can interfere with an innovative interpretation, by our view, an innovative interpretation should also be able to interfere with an established meaning. Libben (1998) presents a case in which such interference might be occurring. Libben provides examples where a person with mixed aphasia produced the literal meaning of opaque compounds, rather than the established meaning. For example, the person paraphrased *blueprint* as "a print that is blue" and *belly-button* as "a button in your stomach". Perhaps, this only occurs with aphasics; perhaps, it only occurs because the established meaning is "lost" and therefore unavailable to the person.

However, it appears that non-impaired individuals also demonstrate interference from innovative meanings: recent exposure to an innovative meaning for a familiar compound increases the difficulty of selecting the conventional meaning Gagné, Spalding, & Gorrie (2005). A familiar compound (e.g. bug spray) was presented as part of a sentence that was consistent with either the conventional reading (e.g. "Because it was a bad season for mosquitoes, Debbie made sure that every time she went outside, she

wore plenty of bug spray") or the innovative reading (e.g. "As a defense mechanism against predators, the Alaskan beetle can release a deadly bug spray"). Immediately after viewing the sentence, the participants saw the phrase (e.g. bug spray) with a definition that was based on either the conventional interpretation (e.g. "a spray for bugs") or the innovative interpretation (e.g. "spray produced by bugs") and indicated whether the definition was a plausible interpretation for the phrase. When the sentence used the established meaning, the conventional definition was judged plausible 89% of the time. However, when the sentence used the innovative meaning, the conventional definition was judged plausible only 64% of the time. In terms of response time, participants took longer to indicate that the conventional reading was plausible when the sentence supported the innovative reading than when it supported the conventional reading. We should note that these results occurred even though the participants were told explicitly that the definition did not have to be the best definition but that they should indicate "yes" if the definition was plausible. Thus, the competition between the relations was not restricted to a case in which the interpretations were explicitly pitted against each other, as would be the case if the participants were told to select the best definition. These findings indicate that competition from the innovative interpretation made it more difficult to accept the conventional interpretations. This finding is highly consistent with the fact that relation priming for established compounds is primarily a result of increasing the availability of a relation other than the one required for the target item (Spalding & Gagné, 2011), as previously discussed.

6. THE ROLE OF RELATION COMPETITION IN THE PROCESSING OF COMPOUNDS THAT LACK AN UNDERLYING RELATION

The investigations of familiar compounds described in Section 4.2 had, as their primary purpose, the demonstration that manipulations of relation availability affect interpretation even of well-known compounds. Beyond this specific purpose, the results strongly suggest that understanding compound processing in general requires the application of what we know about conceptual combination. In particular, the results above suggest considering all compounds as being subjected to a meaning construction process as a relatively obligatory aspect of comprehension.

Although the relation-priming results described above provide good evidence for the meaning construction approach (and relational interpretation,

in particular) in compound word processing, opaque compounds provide a particularly stringent test for the idea that meaning construction is relatively obligatory for comprehension. Opaque compounds are those for which the meaning cannot be computed from the meanings of the constituents, as in, for example, the compound *hogwash*. Similarly, constituents that contribute to the meaning of the compound (e.g. *snow* or *ball* in *snowball*) are semantically transparent, while those that do not (e.g. *hog* or *wash* in *hogwash*) are semantically opaque. Critically, a person simply cannot determine the meaning of *hogwash* from *hog* and *wash*. Thus, such compounds provide an extremely strong test of the idea that meaning construction is attempted for all compounds. Such compounds also provide a strong theoretical contrast between meaning construction approaches such as the RICE theory and the more common conjunctive activation approach to compound processing.

The meaning construction approach is currently a minority position in the literature on compound processing as most current psycholinguistic theories assume that compound processing proceeds primarily by activating the constituents' stored representations, which are used to activate the compound representation (e.g. Libben, 1998; Schreuder & Baayen, 1995, 1997; Taft, 2003, 2004; Taft & Kougious, 2004; Zwitserlood, 1994; Zwitserlood, Bolwiender, & Drews, 2005). That is, theories that rely on conjunctive activation posit that the meaning of a compound is retrieved via activation from constituent representations, rather than actively computed via a meaning construction process.

In general, activation approaches to compound processing account for effects of semantic transparency by assuming that semantically opaque constituents are unable to pass any activation to the semantic representation of the constituent, while semantically transparent constituents are able to do so. Thus, constituents may (or may not) contribute to the semantic access of the compound to the extent that they pass activation to the compound's semantic representation. For example, in Libben's (1998) theory, activation of lexical representations of compounds results in either the activation (in the case of transparent constituents) or the inhibition (in the case of opaque constituents) of the corresponding semantic representations (see also Libben, Gibson, Yoon, & Sandra, 2003). Zwitserlood (1994) proposed that there are no links between the lexical representation of opaque constituent(s) and the corresponding semantic representation, and, consequently, at the semantic level, opaque compounds behave like monomorphemic words. In sum, the conjunctive activation does not involve the active construction

of a compound's meaning based on the constituents. Instead, the already-existing semantic representation of the compound is activated via the constituent representations. Thus, effects of constituents' semantic transparency on compound processing are explained by the spread of activation among the stored representations of the constituents and the compound.

The RICE theory, on the other hand, suggests that the semantic representation corresponding to an activated lexical representation is activated regardless of whether the constituent is transparent and that meaning construction is always attempted. This approach, then, faces the burden of explaining the existence of semantic transparency effects in compound processing. We believe that the influence of the activated semantic representations of the constituents can be obscured because of conflict resulting from the computed meanings based on these representations. As with novel combinations and transparent compounds, if the established meaning and the computed meaning (or meanings) become available within the same time frame, then the system must evaluate which meaning is most likely, and the meanings are likely to compete with each other.

The research discussed above has demonstrated such competition effects in the case of transparent compounds and phrases. In the case of opaque compounds, by definition, the constructed meaning conflicts with the conventional meaning, which should introduce processing costs because this conflict must be resolved as the system attempts to settle on one meaning. On average, there should be less competition between the lexicalized meaning and the computed meaning for transparent compounds than for opaque compounds because there is more consistency among these meanings for transparent compounds (again, by the definition of opaque and transparent compounds). Ji et al. (2011) and Ji (2008) found that opaque compounds were slower than matched transparent compounds, consistent with the claim that opaque compounds face some form of penalty in processing. Furthermore, Ji et al. found that experimental manipulations that aided morphological decomposition (and, thereby, aided semantic composition) slowed the processing of opaque compounds, but did not slow the processing of transparent compounds. Thus, the penalty faced by the opaque compounds appears to be one that is specifically related to meaning construction. Additionally, the frequency of the first constituent differentially influenced response time for transparent and opaque compounds; the processing of transparent compounds was helped by having a high-frequency first constituent, but opaque compounds were hindered by having a high-frequency first constituent.

That is, the more available the constituent representation, the more difficult it was to process an opaque compound, as would be expected if the processing of opaque compounds entailed an attempt at meaning construction that would interfere with the established meaning. In general, these results are highly consistent with the idea that meaning construction occurs for both opaque and transparent compounds. However, this process produces a meaning that strongly conflicts with the conventional meaning of opaque compounds.

Gagné and Spalding (submitted for publication) investigated whether transparent and opaque compounds are sensitive to potentially competing relational interpretations, as suggested by Ji et al. (2011). To answer this question, Gagné and Spalding used the lexical decision data for transparent and opaque compounds from Ji et al. and attempted to predict response time using the diversity of relational interpretations that are possible for those compounds. In this case, the potentially competing relations were assessed at the level of the whole phrase using a possible relations task and using a completely different group of research participants. In the possible relations task, participants are presented with a compound consisting of two words and are asked to pretend that they are learning English and know each of the two words, but have never seen the words used together. Then, they are asked to select the most likely interpretation of the words when used together. The participants select one relation from a list of relations derived from Levi (1978). Relational diversity was computed for each compound simply by counting the number of relations that were chosen by any participant for a given compound. This relational diversity measure was a significant predictor of lexical decision response time for both transparent and opaque compounds. Thus, although there was no obvious reason for participants in Ji et al.'s lexical decision task to be considering the possible relational interpretations of the compounds, the ease of making the lexical decision was affected by how many possible alternative interpretations exist for the compound. In particular, the more possible relations there were, the slower the lexical decision response time was, as predicted by the view that competing relational interpretations slow processing. That such an effect occurs even for opaque compounds (i.e. those compounds where the established meaning cannot be computed from the constituents' meanings) is strong evidence that the meaning construction process is likely to be automatically initiated when compound words are encountered. Furthermore, this meaning construction process appears to be generally relational in nature so that there seems to be a very strong continuity between the processing of novel

modifier-noun phrases and well-established compound words, as suggested by the RICE theory.

7. EVALUATION OF RELATIONAL INTERPRETATIONS

Although the main focus of this chapter is the role of relational competition in conceptual combination, it is also important to understand how relational competition fits into the larger interpretive process. In this section, we consider the evaluation stage of relational interpretation, and in the next we consider elaboration. Most of the research reviewed to this point would suggest that the relation availability for the head constituent has little impact on relation selection. Indeed, this is what Gagné and Shoben (1997) concluded. However, the RICE theory proposes that both semantic and relational information associated with the head noun is used to evaluate the relational interpretations suggested by the modifier (Spalding et al., 2010).

It is clear that semantic (and pragmatic) information associated with the head affects relation selection. First, as initially suggested by Gagné and Shoben (1997), semantic information associated with the head noun has a strong influence on what relational interpretation is chosen for a combined concept. Thus, the phrase *mountain planet* is not interpreted as a planet in the mountains (despite the very strong likelihood that a combined concept with mountain as the modifier will be interpreted using a locative relation) because planets are too big to be located in mountains. As Spalding et al. (2010) pointed out (following Gagné & Shoben, 1997), nearly any interpretation in which the relation is infrequent for the modifier is likely to involve information (either semantic or relational) associated with the head affecting the choice of relation. Otherwise, it is hard to see why one would ever interpret a combined concept in a way that is low in frequency for the modifier. Research specifically focused on the semantics of the head (Spalding & Gagné, 2007) has shown that activating semantic properties of the head in isolation can affect the relational interpretation chosen for a combined concept. For example, noting that some machines are fragile biases people to interpret *clay machine* as "a machine made of clay," while noting that some machines are sturdy biases people to interpret *clay machine* as a "machine for clay". In other words, activating semantic information about the head that is consistent (or inconsistent) with a particular relational interpretation makes that interpretation more (or less) likely compared with activating semantic information about the head that is neutral with respect to the relational interpretation (e.g. noting that some machines are

expensive). In some sense, these results should not be terribly surprising as the relations that are used in relational interpretation are themselves largely semantic in nature. Hence, prior activation of semantic information should and does affect relation selection.

According to RICE, the lack of relational effects associated with the head constituent in most previous research is because of the fact that tasks such as the sense/nonsense judgment are largely driven by the modifier suggesting multiple relations as possible interpretations. In fact, relational information associated with the head has also been shown to affect the relational interpretation of combined concepts. This has been shown, for example, using relational priming of ambiguous combined concepts (e.g. *brick factory* could be factory that makes bricks or a factory made of brick) by Gagné and Shoben (2002), where the interpretation is a choice between two relatively strong relational interpretations. However, unambiguous combined concepts tend to show no effect of the relational distribution of the head when the task is a sense/nonsense judgment (e.g. Gagné & Shoben, 1997). Indeed, this asymmetry between the modifier and the head is one of the most striking aspects of the early research on relational interpretation of combined concepts (Spalding et al., 2010). Nevertheless, the head-based priming effect with ambiguous combinations shows that head-based relational information does exist and can be used. The modifier-head asymmetry appears to be largely and perhaps entirely because of the fact that the sense/nonsense task is particularly sensitive to the modifier's role as the suggester of relations, probably because the sense/nonsense task requires the complete process of interpretation of the phrase. Thus, the sense/nonsense task results tend to strongly reflect what happens at the initial relation suggestion phase of interpretation when the suggested relations are competing with each other on the basis of their strength for the modifier. If this explanation is correct, then a task that isolates, or at least emphasizes, the evaluation phase of relational interpretation should show relational effects associated with the head.

Spalding et al. (2010) showed that the head's relational distribution strongly affects relational interpretation for unambiguous combined concepts when the task is relation verification. In the relation verification task, a combined concept is presented with its relational interpretation (e.g. *snowball = ball made of snow*) and the participants are asked whether this is a good interpretation. Thus, when the task does not require the person to suggest relational interpretations, but simply to verify that the presented relation is appropriate, then the head's relational distribution shows robust effects

(as does the modifier's relational distribution). Spalding et al. also demonstrated a robust head-based relational priming effect for unambiguous combined concepts with the relation verification task. Finally, Spalding et al. showed that the particular form of relational competition found in the sense/nonsense task (i.e. the particular sensitivity to stronger competitor relations, as discussed above) did not play a strong role in the relation verification task. In other words, in the evaluation of an individual relation, it is only the required relation's strength (for both the modifier and the head) that seems to play a role in relation verification response time. The strength of other relations appears to play little role in the verification task. We should note that this complex pattern of asymmetries between the modifier and the head, and the differences among the different phases of the relational interpretation process, are not consistent with any other existing theory of conceptual combination (see the discussion in Spalding et al., 2010, pp. 304–307).

8. ELABORATION OF COMBINED CONCEPTS FOLLOWING RELATION SELECTION

One major weakness in the literature on relational interpretation of combined concepts has been the lack of detailed understanding of the processing during interpretation. The RICE theory and the associated research went some way toward remedying this situation with respect to the relation suggestion and verification phases, as well as elucidating the relationship between novel conceptual combinations and familiar compounds. The elaboration phase of RICE, which describes how we move beyond the relation-based gist, however, is underdeveloped. Understanding this elaboration process is critical to creating a complete theory of conceptual combination and will also contribute to the understanding of noun phrase comprehension in general.

If combined concepts are interpreted via the relational process described in the RICE theory, then the elaboration must proceed from the gist. For example, consider the combination *snowman*. If the gist is "*man* MADE OF *snow*," then the elaboration proceeds on that basis. But, if the gist is "*man* FOR *snow*," then the elaboration proceeds quite differently. A man made of snow would be inanimate and white, while a man for snow would be animate and have a shovel. On the other hand, if combined concepts are elaborated via a feature transference process (such as is posited in schema-based theories of conceptual combination; e.g. Wisniewski, 1997), then the

relational gist would seem to arise largely as an interpretation of the newly created concept with its new features because otherwise there would be no explanation for why the features are consistent with the relational gist. In short, given that the features are consistent with the relational gist, it seems very likely either that the features are derived from the gist or that the gist is an interpretation of the newly created set of features, although it is possible that some other process gives rise to both features and gist. In terms of the time-course of processing, data reported in Gagné (2000) indicate that relation-based interpretations (e.g. *mountain bird*, which refers to a bird that lives in the mountains) take less time than do property-based interpretations (e.g. *robin canary*, which refers to a canary that has a red breast). This result strongly suggests that the inclusion of the features does not precede the creation of the relational gist, which in turn suggests that elaboration follows relation selection. The data also indicated that people show a preference for using paraphrases that reflect the underlying relation, rather than the properties of the newly formed concept. In both cases, it appears that the elaboration of a combined concept proceeds from the relational gist interpretation.

Two important points arise from this characterization of the elaboration of combined concepts. First, the elaboration appears to involve a process of reasoning from the interpretation, rather than a more mechanistic merging or blending of features from the constituent concepts' representations. This view of elaboration fits well with recent work on the "modification effect" (Connolly, Fodor, Gleitman, & Gleitman, 2007; Gagné & Spalding, 2011; Jönsson & Hampton, 2008). In the modification effect paradigm, participants judge the likely truth of a feature for a modified noun (e.g. *baby ducks have webbed feet*) or for the noun itself (e.g. *ducks have webbed feet*). The result is that the judgments are lower for the modified noun than for the noun (Connolly et al., 2007; Gagné & Spalding, 2011; Jönsson & Hampton, 2008). While the empirical result is robust, three different theoretical explanations have arisen. Connolly et al. claimed that the ratings are lower because the features of modified nouns have to be learned from the world, rather than inherited from the head noun. Jönsson and Hampton claimed that the feature was directly inherited from the head and then adjusted because it is unlikely for the modifier. Gagné and Spalding showed that the effect occurs for combinations with non-word modifiers (e.g. *flag ducks*), which does not fit well with either explanation: the feature would never be learned for a non-word combination, and the inheritance view predicts either perfect inheritance or no inheritance at all. Gagné and Spalding propose that the

modification effect is driven by inferential processing at the point of decision about the feature and that the effect itself is because of participants' expectation that the modified noun should be similar to, yet contrast with, the noun: if there was no contrast, why modify the noun?

Second, one possible implication of such an understanding of elaboration would be that judgments about features (even for unmodified concepts) are also driven by processes other than simply inspecting the contents of the concept. For example, we have shown (Gagné & Spalding, 2007) that creating a relational interpretation of a combined concept can change the likelihood of a feature for the head concept itself; interpreting the phrase *peeled apple* makes it more difficult to affirm that apples are red. Note here, that it is not, in general, a property of peeled things to remove redness—instead, it is the interpretation (i.e. an apple that has been peeled), along with the color of the apple beneath the peel, that gives the color of the peeled apple. Even this, though, by most theories of conceptual combination, should not make it more difficult to verify that apples are red. After all, the head noun's conceptual representation should be unaffected and its color feature intact. So, what is it that makes this verification more difficult? Gagné and Spalding suggested that the effect is driven by the fact that the interpretation of the phrase peeled apples causes the activation of representations of particular apples (namely, apples that have been peeled) and that these highly activated apples, because they are not red, inhibit the verification that apples in general are red.

These results on elaboration, in turn, might suggest that the relation between concepts and features is potentially rather different than that in most theories of concepts. In particular, it might support a view of features closer to that of classical philosophical approaches to the relationship (in particular, a relation of predication; Spalding & Gagné, in press) rather than the kind of "containment" metaphor that is common in modern psychology (see, e.g. the discussion in Laurence & Margolis, 1999). The point here is that, in classical philosophical approaches, properties were predicated of a term via reasoning and that while such predications could become well remembered and habitual, this did not entail that the feature predicated became part of the concept. The Aristotelian–Thomistic view of features, for example, is much more like the type of theory identified by Laurence and Margolis (1999) as connecting features to the concept via inference (Spalding & Gagné, in press). In the Aristotelian–Thomistic view, features (i.e. properties and accidents) are not part of the concept, per se. The fact that most individual dogs have legs as parts does not mean that the concept

DOG has LEGS as a part, nor does the fact that people know that dogs have legs mean that their concept of DOG must include LEGS as a part. Instead, the features are themselves generally concepts linked to the (target) concept in the act of predication. Aristotle often refers to this as combining and dividing. Putting together two concepts (combining; e.g. dogs have legs) or denying their combination (dividing; e.g. dogs do not have hard drives) is a separate cognitive act from those involved in creating the concept or in using the concept to represent the thing to the mind. The predication can be remembered and can become habitual, but it is not part of the concept. Critically, in such views, the properties predicated of concepts need not combine in any mechanistic way, but are highly amenable to reasoning about whether the property would continue to be true of a modified concept.

One point that has not yet been investigated much is the extent to which elaboration is immediate and relatively automatic or neither immediate nor automatic. The results of Gagné and Spalding (2007) that it becomes more difficult to say that apples are red following a presentation of peeled apples suggest at least some relatively immediate and automatic elaboration. On the other hand, the results of Gagné and Spalding (2011) suggest that some elaboration is not automatic and might only occur as needed (e.g. to answer a particular question or to integrate with a new context). Furthermore, if elaboration proceeds from the relational gist, then the elaboration of combined concepts will presumably have strong similarity to other areas of text or discourse processing. In particular, elaboration of combined concepts would have to be highly similar to the processing of full noun phrases. As Wisniewski (1996) points out, elaboration will also often involve selecting different senses of modifiers (see also Mullaly, Gagné, Spalding, & Marchak, 2010) and property terms, which will raise further complexities of processing. Clearly, the elaboration of combined concepts is highly complex and is much less well understood than relation selection. However, investigating elaboration may provide important insight into a number of cognitive processes, including reasoning, text or discourse processing and the relationship between concepts and features. In short, this is a promising avenue for further research.

9. SUMMARY

This chapter reviews the recent literature on the relational interpretation of modifier-noun phrases and compound words. Section 1 argues for the importance of conceptual composition processes in understanding how

the conceptual system as a whole operates. Section 2 presents evidence that novel modifier-noun phrases and known compound words may be comprehended via a shared meaning construction process and thus that theories of conceptual composition should apply to both novel modifier-noun phrases and known compounds.

In Section 3, we present a description of the RICE theory (summarized in Figure 3.1), which is a theory of both modifier-noun phrase and compound word processing, based on an underlying common process of conceptual combination. RICE provides a theoretical framework for understanding relational interpretation of modifier-noun phrases and compounds as a three stage process in which (1) relational interpretations are suggested with a strength proportional to the relation's availability for the modifier, (2) these relational interpretations compete to be selected and are evaluated with respect to both semantic information and relational availability for both the modifier and the head noun, and (3) the selected relation is elaborated beyond the relational gist interpretation.

Section 4 reviews the literature showing that the ease of comprehending modifier-noun phrases and compound words is strongly affected by relational availability of the modifier. Such effects arise whether that availability is based on long-term prior usage or recent usage, whether the phrases/compounds are novel or well known, and whether they are presented in or out of context.

Section 5 shows that the competition among relational interpretations takes on a specific character such that the primary determinant of ease of processing is the number of competitors that are more available (for the modifier) than the required relation. Consistent with this finding, modifier-based relation-priming effects (i.e. repeated relations lead to faster processing than unrepeated relations) are shown to result primarily from competition from the different relation in the unrepeated condition, rather than facilitation in the repeated relation condition. Finally, relational competition effects are shown to occur at the level of the whole interpretation, as well as at the level of the individual constituents.

Section 6 shows that relational competition effects can be seen even for opaque compounds—those that lack an underlying relational interpretation entirely. For example, lexical decision RT's for opaque compounds can be (partially) predicted by the number of possible relational interpretations that the opaque compound could be construed as having. The more such possible interpretations, the slower the lexical decision judgment, suggesting competition among the possible interpretations and the established

meaning of the opaque compound. Such effects strongly suggest that the relational interpretation process is more or less obligatory when modifier-noun phrase or compound word structures are encountered.

In Section 7, we describe research that used a relation verification task to isolate the evaluation process and focus it on a single relational interpretation. This research demonstrates effects of the relation's availability for both the modifier and the head and contrasts these results with the results of experiments using sense/nonsense judgments. This section shows two critical points. One, although the effects of relational availability for the head can be hidden in the sense/nonsense task, those effects are real. Second, this result strongly suggests that the effect of multiple competitor relations arises in the evaluation stage, even though the effects are related to the relational availability for the modifier.

Section 8 provides a discussion of how the elaboration of the relational interpretations might proceed. In particular, the idea that the elaboration proceeds from the relational gist via reasoning processes is explored and related to some other aspects of the literature on concepts.

10. CONCLUDING REMARKS

Understanding how people comprehend modifier-noun phrases and compound words, and the combined concepts that underlie them, is an important step to understanding the productivity that is a hallmark of human cognition. It would seem that combining two words with a relatively fixed syntactic structure should be straightforward. Indeed, it is relatively simple from the perspective of the person doing the task. Within about 1 s, the person has comprehended the phrase or compound, even if it is completely new to them. Yet, this simplicity is also deceptively complex. We have seen that the processing itself makes use of information about the distribution of relations with each constituent, with the semantics and pragmatics associated with each constituent, and with an even broader range of information that can be brought to bear via reasoning processes in (at least beginning) the elaboration beyond the relational gist interpretation. Other reasoning processes are likely to arise in elaboration well beyond the 1 s that it takes for comprehending the combination and creating the relational gist. This seemingly simple use of the conceptual and language systems provides an outstanding opportunity to learn about human cognition.

REFERENCES

Allen, M. R. (1978). *Morphological investigations*. University of Connecticut. (Ph.D. Dissertation).

Bauer, L. (2008). Dvandva. *Word Structure*, *1*, 1–20.

Bertram, R., Laine, M., & Karvinen, K. (1999). The interplay of word formation type, affixal homonymy, and productivity in lexical processing: evidence from a morphologically rich language. *Journal of Psycholinguistic Research*, *28*, 213–226.

Bisetto, A., & Scalise, S. (2005). The classification of compounds. *Lingue E Linguaggio*, *4*, 319–332.

Booij, G. (2005). Compounding and derivation: evidence for construction morphology. In W. Dressler, D. Kastovsky, O. E. Pfeiffer & F. Rainer (Eds.), *Morphology and its demarcations* (pp. 109–132). Amsterdam/Philadelphia: John Benjamins.

Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition*, *103*, 1–22.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, *53*, 810–842.

Estes, Z. (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language*, *48*, 304–319.

Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: a COPPER HORSE is faster when seen through a GLASS EYE. *Journal of Memory and Language*, *55*, 89–101.

Finin, T. W. (1980). *The semantic interpretation of compound nominals*. Urbana-Champaign: University of Illinois. (Ph.D. Dissertation).

Fiorentino, R., & Poeppel, D. (2007). Compound words and structure in the lexicon. *Language and Cognitive Processes*, *22*, 953–1000.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Gagné, C. L. (2000). Relation-based combinations versus property-based combinations: a test of the CARIN theory and the dual-process theory of conceptual combination. *Journal of Memory and Language*, *42*, 365–389.

Gagné, C. L. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 236–254.

Gagné, C. L. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language*, *81*, 723–735.

Gagné, C. L., & Murphy, G. L. (1996). Influence of discourse context on feature availability in conceptual combination. *Discourse Processes*, *22*, 79–101.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71–87.

Gagné, C. L., & Shoben, E. J. (2002). Priming relations in ambiguous noun-noun combinations. *Memory & Cognition*, *30*, 637–646.

Gagné, C. L., & Spalding, T. L. (2004a). Effect of discourse context and modifier relation frequency on conceptual combination. *Journal of Memory and Language*, *50*, 444–455.

Gagné, C. L., & Spalding, T. L. (2004b). Effect of relation availability on the interpretation and access of familiar noun-noun compounds. *Brain and Language*, *90*, 478–486.

Gagné, C. L., & Spalding, T. L. (2006). Relation availability was not confounded with familiarity or plausibility in Gagné and Shoben (1997): comment on Wisniewski and Murphy (2005). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1431–1437. discussion 1438–1442.

Gagné, C. L., & Spalding, T. L. (2007). The availability of noun properties during the interpretation of novel noun phrase. *Mental Lexicon*, *2*, 241–260.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20–35.

Gagné, C. L., & Spalding, T. L. (2011). Inferential processing and meta-knowledge as the bases for property inclusion in combined concepts. *Journal of Memory and Language*, *65*, 176–192.

Gagné, C. L. & Spalding, T. L. Relation diversity for opaque and transparent English compounds. In F. Rainer, W. Dressler, F. Gardani, H. C. Luschutzky (Eds.), *Current issues in linguistic theory*. John Benjamins, submitted for publication.

Gagné, C. L., Spalding, T. L., Figueredo, L., & Mullaly, A. C. (2009). Does snow man prime plastic snow? The effect of constituent position in using relational information during the interpretation of modifier-noun phrases. *Mental Lexicon*, *4*, 41–76.

Gagné, C. L., Spalding, T. L., & Gorrie, M. C. (2005). Sentential context and the interpretation of familiar open-compounds and novel modifier-noun phrases. *Language and Speech*, *48*, 203–221.

Gagné, C. L., Spalding, T. L., & Ji, H. (2005). Re-examining evidence for the use of independent relational representations during conceptual combination. *Journal of Memory and Language*, *53*, 445–455.

Gerrig, R. J. (1989). The time-course of sense creation. *Memory & Cognition*, *17*, 194–207.

Gerrig, R. J., & Bortfeld, H. (1999). Sense creation in and out of discourse contexts. *Journal of Memory and Language*, *41*, 457–468.

Gleitman, L. R., & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language*. New York: Norton.

Inhoff, A. W., Radach, R., & Heller, D. (2000). Complex compounds in German: interword spaces can facilitate segmentation but hinder assignment of meaning. *Journal of Memory and Language*, *42*, 23–50.

Ji, H. (2008). *The influence of morphological complexity on word processing*. Department of Psychology, University of Alberta. (Ph.D. dissertation).

Ji, H., & Gagné, C. L. (2007). Lexical and relational influences on the processing of Chinese modifier-noun compounds. *Mental Lexicon*, *2*, 387–417.

Ji, H., Gagné, C. L., & Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language*, *65*, 406–430.

Jönsson, M. L., & Hampton, J. A. (2008). On prototypes as defaults (Comment on Connolly, Fodor, Gleitman and Gleitman, 2007). *Cognition*, *106*, 913–923.

Kay, P., & Zimmer, K. E. (1976). On the semantics of compounds and genitives in English. In *Proceedings of the sixth California Linguistics Association*. San Diego: San Diego State University.

Koester, D., Gunter, T. C., & Wagner, S. (2007). The morphosyntactic decomposition and semantic composition of German compound words investigated by ERPs. *Brain and Language*, *203*, 64–79.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings*. Cambridge, MA: A Bradford Book, MIT Press.

Lees, R. B. (1960). *The grammar of English nominalizations*. Bloomington, IN: Indiana University.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Libben, G. (1998). Semantic transparency in the processing of compounds: consequences for representation, processing, and impairment. *Brain and Language*, *61*, 30–44.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, S. (2003). Compound fracture: the role of semantic transparency and morphological headedness. *Brain and Language*, *84*, 50–64.

Libben, G., & Jarema, G. (2006). *The representation and processing of compound words*. New York: Oxford University Press.

Maguire, P., Devereux, B., Costello, F., & Cater, A. (2007). A reanalysis of the CARIN theory of conceptual combination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 811–821.

Marchand, H. (1969). *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Wiesbaden: Otto Harrassowitz.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469–1481.

Meyer, R. (1993). *Compound comprehension in isolation and in context: the contribution of conceptual and discourse knowledge to the comprehension of German novel noun–noun compounds*. Tübingen: Niemeyer.

Mullaly, A. C., Gagné, C. L., Spalding, T. L., & Marchak, K. A. (2010). Examining ambiguous adjectives in adjective-noun phrases: evidence for representation as a shared core-meaning with sense specialization. *Mental Lexicon*, *5*, 87–114.

Olsen, S. (2004). The case of copulative compounds. In A. ter Meulen & W. Abraham (Eds.), *The composition of meaning, from lexeme to discourses* (pp. 17–37). Amsterdam: Benjamins.

Olsen, S., & Van der Meer, G. (2001). Coordination in morphology and syntax: the case of copulative compounds. In A. ter Meulen (Ed.), *Making sense: From lexeme to discourse. Festschrift for Werner Abraham* (pp. 87–101). Dorderecht: Kluwer.

Scalise, S., & Bisetto, A. (2009). The classification of compounds. In R. Lieber & P. Stekauer (Eds.), *Oxford handbook of compounding* (pp. 34–53). Oxford: Oxford University Press.

Schreuder, R., & Baayen, R. H. (1995). *Modeling morphological processing*. Hillsdale, NJ, England: Lawrence Erlbaum.

Schreuder, R., & Baayen, R. H. (1997). How complex simple words can be. *Journal of Memory and Language*, *37*, 118–139.

Shoben, E. J. (1991). Predicating and nonpredicating combinations. In P. J. Schwanenflugal (Ed.), *The psychology of word meanings* (pp. 117–135). Hillsdale, NJ: Erlbaum.

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences*, *3*, 99–105.

Spalding, T. L., & Gagné, C. L. (2007). Semantic property activation during the interpretation of combined concepts. *Mental Lexicon*, *2*, 25–47.

Spalding, T. L., & Gagné, C. L. (2008). CARIN theory reanalysis reanalyzed: a comment on Maguire, Devereux, Costello, and Cater (2007). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1573–1578.

Spalding, T. L., & Gagné, C. L. (2011). Relation priming in established compounds: facilitation? *Memory & Cognition*, *39*, 1472–1486.

Spalding, T. L., & Gagné, C. L. Concepts in Aristotle and Aquinas: implications for current theoretical approaches. *Journal of Theoretical and Philosophical Psychology*, in press.

Spalding, T. L., Gagné, C. L., Mullaly, A. C., & Ji, H. (2010). Relation-based interpretations of noun-noun phrases: a new theoretical approach. In S. Olson (Ed.), *New impulses in word-formation (Linguistische Berichte Sonderheft 17)* (pp. 283–315). Hamburg: Buske.

Storms, G., & Wisniewski, E. J. (2005). Does the order of head noun and modifier explain response times in conceptual combination? *Memory & Cognition*, *33*, 852–861.

Štekauer, P. (2005). *Meaning predictability in word formation: Novel, context-free naming units*. Amsterdam: John Benjamins Publishing Company.

Štekauer, P. (2006). On the meaning predictability of novel context-free converted naming units. *Linguistics*, *44*, 489–539.

Štekauer, P. (2009). Meaning predictability of novel context-free compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding* (pp. 272–298). Oxford, UK: Oxford University Press.

Taft, M. (2003). Morphological representation as a correlation between form and meaning. In E. Assink & D. Sandra (Eds.), *Reading complex words: Cross-language studies* (pp. 113–137). Amsterdam: Kluwer Academic.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology*, *57*, 745–765.

Taft, M., & Kougious, P. (2004). The processing of morpheme-like units in monomorphemic words. *Brain and Language*, *90*, 9–16.

Teall, F. (1892). *English compound words and phrases*. New York: Funk & Wagnalls.

Warren, B. (1978). *Semantic patterns of noun-noun compounds*. Goteborg: Acta Universitatis Gothoburgensis.

Wisniewski, E. J. (1996). Construal and similarity in conceptual combination. *Journal of Memory and Language*, *35*, 434–453.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin and Review*, *4*, 167–183.

Wisniewski, E. J., & Murphy, G. L. (2005). Frequency of relation type as a determinant of conceptual combination: a reanalysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 169–174.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes*, *9*, 341–368.

Zwitserlood, P., Bolwiender, A., & Drews, E. (2005). Priming morphologically complex verbs by sentence contexts: effects of semantic transparency and ambiguity. *Language and Cognitive Processes*, *20*, 395–415.

CHAPTER FOUR

List-Method Directed Forgetting in Cognitive and Clinical Research: A Theoretical and Methodological Review

Lili Sahakyan[1], Peter F. Delaney, Nathaniel L. Foster, Branden Abushanab

Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, USA
[1]Corresponding author: E-mail: l_sahaky@uncg.edu

Contents

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00004-6

Abstract

The primary purpose of this chapter is to provide an up-to-date review of the twenty-first century research and theory on list-method directed forgetting (DF) and related phenomena like the context-change effect. Many researchers have assumed that DF is diagnostic of inhibition, but we argue for an alternative, noninhibitory account and suggest reinterpretation of earlier findings. We first describe what DF is and the state of the art with regard to measuring the effect. Then, we review recent evidence that brings DF into the family of effects that can be explained by global memory models. The process-based theory we advocate is that the DF impairment arises from mental context change and that the DF benefits emerge mainly but perhaps not exclusively from changes in encoding strategy. We review evidence (some new to this paper) that strongly suggests that DF arises from the engagement of controlled forgetting strategies that are independent of whether people believed the forget cue or not. Then we describe the vast body of literature supporting that forgetting strategies result in contextual change effects, as well as point out some inconsistencies in the DF literature that need to be addressed in future research. Next, we provide evidence—again, some of it new to this chapter—that the reason people show better memory after a forget cue is that they change encoding strategies. In addition to reviewing

the basic research with healthy population, we reinterpret the evidence from the literature on certain clinical populations, providing a critique of the work done to date and outlining ways of improving the methodology for the study of DF in special populations. We conclude with a critical discussion of alternative approaches to understanding DF.

1. INTRODUCTION

Forgetting is often a passive process that happens regardless of our conscious intentions. However, this is not always the case; forgetting sometimes is an active process of intentionally engaging strategies that reduce memory for material that interferes with one's current goals. To this end, forgetting can be an adaptive process. For example, to use software that has been recently updated, the user must forget the prior ways of doing things to avoid errors. Similarly, we are often told "never mind, forget what I just said" following exposure to erroneous information.

Intentional forgetting has been studied extensively in the laboratory via directed forgetting (DF) manipulations, which produce impaired memory for material following an instruction to try to forget. In DF studies, participants study some information and are subsequently told to forget certain portions of it. The *list-method* instructs people to forget an entire list of earlier studied items, whereas *the item-method* instructs people to forget or remember on item-by-item basis. Both methods demonstrate forgetting of unwanted information on demand, although the way by which forgetting is accomplished differs between procedures.

This chapter primarily centers on list-method DF, partly because the mechanisms proposed to explain it have been a topic of active debate in the recent decade, and partly because the presumed connection with inhibitory processes has made the paradigm an attractive diagnostic tool for investigating clinical and special populations. We provide an up-to-date review of DF and related phenomena (e.g. the context-change paradigm) with a focus on studies conducted since 2000. There is already an excellent edited volume describing twentieth century DF research (Golding & MacLeod, 1998).

We will begin by describing methodological issues in list-method DF. The designs used to study DF in cognitive research have evolved to keep pace with theoretical and methodological developments, and so we will review the difficulties posed by earlier designs and suggest better-controlled procedures. We will then outline our broad theory of DF and describe the large number of studies supporting that broad theory, as well as areas of current controversy in DF. Even researchers who are familiar with our previous

manifestos on this topic may find this section interesting, as our view has evolved and expanded even since 2010, and because we will present previously unpublished data that bear on a number of issues. Finally, we will explore how the theory can inform studies of individual differences, especially in clinical populations, and make recommendations for conducting such studies.

2. LIST-METHOD DF: DESIGN AND MEASUREMENT

2.1. Basic Design Issues

DF is generally assessed in terms of two separate components of the effect: the costs and the benefits. To understand these terms, we first need to explain how most DF studies are set up, and then we will explain the usual findings.

The most frequently used design is the *two-list design*, which presents two lists of items (usually words) to study for a later memory test. The first list we will call L1 and the second list we will call L2. Following L1, a *cue* is presented either to keep remembering L1 or to forget L1. Because in the two-list design the cue always occurs between L1 and L2, we sometimes refer to L1 as the precue items and L2 as the postcue items, which is helpful when considering designs that include more lists (see below for a discussion of these designs). There are a variety of ways to present the forget cue, but a typical instruction is to try to forget L1 because it was "just for practice". The remember cue provides a control group against which to assess the effects of the forget instruction, and it typically involves telling people to keep remembering L1 because it was only the first half of the study material. A filler task is usually inserted after learning L2 to reduce recency effects. Afterward, participants receive a recall test for both lists, including items that they were told to forget. In free recall, the forget instruction reduces recall of L1 items relative to the remember group—a phenomenon called the *costs* of DF. Similarly, the forget instruction produces better memory of L2 items relative to the remember group—a phenomenon called the *benefits* of DF (for reviews, see Bäuml, 2008; Bjork, Bjork, & Anderson, 1998; MacLeod, 1998).

Whether participants are instructed to recall L1 or L2 first varies across studies. Some studies have required participants to recall L1 first and then recall L2 (e.g. Delaney & Sahakyan, 2007; Foster & Sahakyan, 2011; Mulji & Bodner, 2010; Pastötter & Bäuml, 2007; Spillers & Unsworth, 2011). Other studies have fully counterbalanced the test orders (e.g. Kimball & Bjork, 2002; Lehman & Malmberg, 2009, 2011a; Pastötter, Kliegl, & Bäuml, 2012; Sahakyan & Delaney, 2010; Zellner & Bäuml, 2006), and some studies

have allowed participants to recall all items in any order (Golding & Gottlob, 2005; Joslyn & Oakes, 2005; Minnema & Knowlton, 2008; Wessel & Merckelbach, 2006; Zellner & Bäuml, 2006).

One obvious concern is that differences in output order might affect DF through output interference (Anderson, 2005). Output interference is the finding that retrieving some items from a set first reduces the likelihood of retrieving the rest of the set (Dong, 1972; Roediger, 1974; Roediger & Schmidt, 1980; Tulving & Arbuckle, 1966). In DF, when test order is not controlled people might be inclined to begin recall with L2 items due to recency, causing output interference on L1 items. Given that L2 is better remembered in the forget group than the remember group, there may be greater output interference on L1 in the forget group. Indeed, when left to recall lists in any order, participants prefer to initiate recall with postcue remember items, thereby resulting in an effect of output order (Golding & Gottlob, 2005). However, a recent meta-analysis of list-method DF studies that included both forget and remember groups did not support this notion; initiating retrieval with L2 did not exaggerate the magnitude of DF costs, although it made it easier to detect the DF benefits (Pastötter et al., 2012). Although these results suggest that output order affects the benefits more than the costs (at least with unrelated items across the lists), we nevertheless suggest fully controlling the test order of each list to obtain more "pure" measures of recall uncontaminated by prior retrieval. Output interference could become a particularly thorny issue if the lists contain a mixture of items (e.g. neutral and emotional words), which could create different degrees of output interference if the test order is left uncontrolled.

Another procedural inconsistency across DF studies is whether a control group is included in the design. In the past, many DF paradigms have failed to include a control group wherein both L1 and L2 presentations are followed by a remember cue. In such paradigms, the magnitude of forgetting was quantified by comparing recall performance for the forget list (L1) relative to the remember list (L2). This measure is referred to as the "R–F" measure. The justification for not including a control group is based on the notion that the only ostensible difference between L1 and L2 is the type of cue received. However, there are several confounding variables that highlight the necessity for a control group. First, heightened recall performance for L2 items could be attributed to a recency effect, making it difficult to assume that L2 gains are a result of L1 forgetting. Second, increased accessibility for L2 items may be attributed to learning effects. That is, following L1 exposure, participants may be more adapted to the procedure, leading

them to shift their encoding strategies for L2 items. Finally, reduced recall performance for L1 items may be attributed to retroactive interference accrued from L2 items. In other words, at the time of recall, L1 accessibility is reduced due to the interference built up from L2 encoding. Given the number of alternative explanations for reduced L1 recall and heightened L2 recall, it is challenging to attribute these findings solely to DF without comparing recall performance to a control group that was required to remember both lists.

2.2. Multilist Designs in DF

A variation on the traditional two-list procedure is the use of a *three-list design*, which was put forth to reduce potential confounds in the two-list procedure (Lehman & Malmberg, 2009). The sources of these confounds are as follows. First, L2 presentation is typically followed by a filler task, whereas L1 presentation is not, which means that L2 memory could be penalized in comparison with L1. For example, participants can covertly rehearse the end of L1 during the beginning of L2 encoding, increasing memory for the last few L1 items, while hindering encoding of the first few L2 items. Additionally, in a traditional two-list design, L2 is subjected to proactive interference from L1, but L1 receives no interference from a prior list. Therefore, the two-list approach may benefit L1 and harm L2. To control these factors, in a three-list design, participants study three lists, each followed by a brief filler activity of equal duration before a forget or a remember cue is presented. For half of the participants, the forget cue is presented after L2, whereas the remaining half are told to remember all lists. At the time of final test, half of the participants are required to recall either L2 or L3 (L1 is never tested). Note that the three-list design could lead to potential floor effects on L2 if presentation rates are not substantially increased (e.g. Lehman & Malmberg, 2011a). Both the two-list and the three-list designs manipulate the forget/remember cue between-subjects.

An alternative approach is the use of a *four-list design*, where the forget/remember cue is varied within-subjects (e.g. Zellner & Bäuml, 2006). In this procedure, two lists are assigned to the first block, and two lists are assigned to the second block. The forget cue occurs either during the first block (e.g. after presentation of L1) or during the second block (e.g. after presentation of L3). Participants complete the forget condition (F, R) for the first block and the remember condition (R, R) for the second block, or vice versa. Memory for both lists is tested after each block, with counterbalanced test order of the lists. This procedure produces findings similar

to the traditional two-list version. One caveat concerns a potential shift to better encoding strategies after the first block has been completed and memory for both lists has been tested (Delaney & Knowles, 2005; Sahakyan & Delaney, 2003). With that in mind, there may be some order effects depending on whether forgetting is inferred from the RR–FR order or the FR–RR order. For example, the size of the DF impairment may be exaggerated in the FR–RR order because by the time of L3, participants may adopt better encoding strategies on L3. Although Bäuml et al. typically do not obtain session order effects, faster presentation rates that are employed in some of their studies may mask such effects because with faster rates there is leaving little room for employing a better encoding strategy than rehearsal. An order effect has been reported in one study that used a slow presentation rate and obtained larger DF when the forget cue occurred in the first block compared to the third block (Hanczakowski, Pasek, & Zawadska, 2012). Therefore, researchers should be mindful of potential order effects that may arise from use of the four-list, within-subjects approach to DF.

2.3. Additional Dependent Measures in DF

DF has been assessed primarily in terms of the number of correctly recalled or recognized items from each list (e.g. costs/benefits). In its early days, it was inferred through an even less-sensitive measure (e.g. "R–F" difference). However, recently researchers have strived to include a wider array of dependent measures with the goal of constraining the interpretations. Such measures include investigations of serial position functions (Geiselman, Bjork, & Fishman, 1983; Lehman & Malmberg, 2009; Pastötter & Bäuml, 2010; Pastötter et al., 2012; Sahakyan & Foster, 2009; Sheard & MacLeod, 2005), intrusion errors (Lehman & Malmberg, 2009; Sahakyan & Delaney, 2010; Spillers & Unsworth, 2011), recall latencies, which indicate an average time point in the recall period when responses were emitted (Spillers & Unsworth, 2011; Unsworth, Spillers, & Brewer, 2012), first response functions, which provide a measure of where in the list participants initiate their recall (e.g. Lehman & Malmberg, 2009), and conditional response probabilities, which indicate how people transition between responses during recall (e.g. Unsworth et al., 2012). The use of multiple measures provides a more detailed assessment of the recall process, thereby facilitating a deeper understanding of the mechanisms underlying list-method DF. Together, they coalesce to provide a broader perspective on list-method DF, which in turn can be used to guide and advance future research.

3. OUR FRAMEWORK OF LIST-METHOD DF

3.1. Older Ideas about the Causes of DF

The earliest theory of DF was the *selective rehearsal account* (e.g. Bjork, 1970, 1972), which was proposed at the time when not much was known about the distinction between the list-method and item-method DF. According to this view, the costs arise because participants stop rehearsing to-be-forgotten items in response to the forget cue, and they devote their rehearsal and mnemonic activity more effectively to to-be-remembered items, producing the benefits. This works well in explaining item-method DF, but its suitability for explaining list-method DF started to wane when it became clear that the list-method and item-method were differentially sensitive to recognition and indirect tests of memory (e.g. Basden, Basden, & Gargano, 1993; MacLeod, 1999). Additionally, the presence of significant DF in incidental learning (Geiselman et al., 1983; Sahakyan & Delaney, 2005, 2010; Sahakyan, Delaney, & Goodmon, 2008) made it increasingly difficult for the selective rehearsal account to encompass list-method DF, where it was soon supplemented by the *retrieval inhibition* account (Bjork, 1989; Geiselman et al., 1983).

According to the retrieval inhibition account, when participants are instructed to forget L1, they initiate an inhibitory process that suppresses or deselects that list so as to facilitate the learning of subsequent lists. Therefore, L1 memory suffers from inhibition, and L2 memory benefits because inhibited L1 items do not cause proactive interference on L2.

The term *inhibition* is interpreted differently among DF researchers. Some interpret inhibition as suppressing the activation level of L1 items (e.g. Barnier, Conway, Mayoh, Speyer, Avizmil, & Harris, 2007; Conway, Harries, Noyes, Racsmány, & Frankish, 2000; Racsmány & Conway, 2006; Racsmány et al., 2008); others propose that items reside in memory at their full strength as evidenced by indices of memory other than free recall, but retrieval of the L1 episode is inhibited (e.g. Bjork, 1989; Bjork & Bjork, 1996), and recently some have argued that the L1 context may be inhibited (e.g. Anderson, 2005; Bäuml, 2008; Pastötter & Bäuml, 2010). Multiple meanings of the term *inhibition* can be misleading, especially to researchers outside of the memory field who employ the DF paradigm as a test of inhibitory function in various populations. Equating an empirical pattern of findings with an underlying inhibitory mechanism is unwarranted because mechanisms other than inhibition could lead to impaired access (e.g. context change and blocking).

Historically, in the mid-to-late 1980s, inhibition was used to explain the discrepant findings in item-method and list-method DF. However, nothing was assumed about the nature of inhibition or the conditions under which inhibition should be more or less evident. The unspecified nature of inhibition makes it difficult to think of it as a mechanism that explains DF as opposed to providing a description of empirical findings. Later, in the 1990s, inhibition was developed into a full theoretical mechanism explaining retrieval-induced forgetting (RIF) (e.g. Anderson, 2003; Anderson, Bjork, & Bjork, 1994), which is a memory impairment phenomenon that occurs at the item-level, unlike the list-method DF, which occurs at the list-level. Nowadays inhibition account makes testable predictions for RIF, but it fails to make predictions for DF because it is unclear what exactly is being inhibited in DF. Just about any outcome can be interpreted to be consistent with inhibition, which significantly reduces its appeal as a theory of DF. That DF is frequently reviewed together with other inhibitory paradigms (e.g. RIF and think-no-think procedure) may have contributed to inadvertent blending of the use of the term *inhibition* across paradigms (e.g. Anderson, 2003, 2005; Depue, 2012).

We take issue with the notion that the "L1 episode" is suppressed as a result of the forget cue. Someone not well-versed in the nuances of memory theory is likely to interpret "inhibition of L1 episode" as analogous to "inhibition of items within that episode". However, the L1 episode contains more than just the studied items within that list. It contains contextual information, which distinguishes that particular episode from other similar episodes (e.g. the L2 episode). Thus, an episode is a complex term that refers to both the items and the context within which those items were experienced. We argue that there is no evidence for inhibition of items, and we further argue that there is no evidence for inhibition of context. Instead, we explain the impaired accessibility to L1 items by invoking the notion of a mismatch between the retrieval cues being used to search the memory and the contents of memory.

3.2. A Two-Factor, Process-Based Account

Figure 4.1 shows how we currently understand list-method DF. Assuming a two-list design, the forget cue takes place after L1 and triggers two types of reflective processing. First, it triggers a decision about how to comply with the forget instruction. We argue that the primary mechanism producing forgetting involves a change of mental context between the two lists, and that some strategies (but not all) bring about this contextual change.

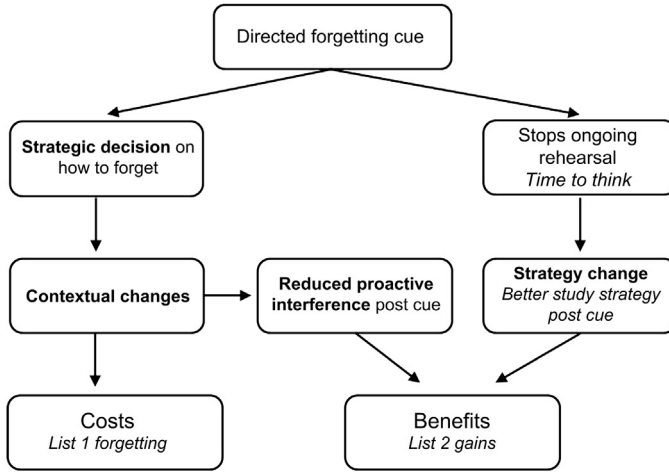


Figure 4.1 A two-factor, process-based framework for list-method directed forgetting.

Context change also contributes somewhat to the benefits by reducing proactive interference on L2, but this effect is often small and difficult to detect, particularly when there are only two lists. Second, the forget cue stops ongoing rehearsal processes on L1, which gives participants time to think about what they should do to learn L2. This leads some participants to change study strategies and consequently improve L2 learning. We think that this strategy-change component accounts for the improved memory for postcue material.

Sections 4–6 provide the rationale for each of these mechanisms and the empirical support for them. Section 4 discusses the importance of controlled strategies for obtaining DF, and how metacognitive beliefs affect the decision to deploy those strategies. Section 5 details the mental context-change account and Section 6 explains the strategy-change account. We will therefore wait to explain in detail why we think these mechanisms provide a better explanation for list-method DF phenomena than earlier accounts do.

4. FORGETTING IS A STRATEGIC DECISION

In a series of studies from our lab, we have explored what people do in response to the forget cue. Specifically, we have accumulated a wealth of information regarding how participants interpret and comply with forget instructions. People may think that they already have forgotten, and therefore fail to deploy any strategy to forget. Likewise, they may not know what kind of strategy to deploy in order to forget, and therefore do nothing.

Such results are important not only for demonstrating that forgetting is an active process that requires conscious activity in order to forget, but also for explaining individual differences—including between-group differences—in the ability to forget unwanted information.

4.1. Importance of Controlled Strategies

Many studies in our lab have revealed that obtaining DF requires that participants engage in controlled behaviors to reduce accessibility of unwanted information. For example, Foster and Sahakyan (2011) collected retrospective verbal reports regarding what participants did in response to the forget instruction, including whether they believed/trusted the forget cue. In a group of 80 forget-group participants, 25% reported "doing nothing", whereas the remaining participants reported a variety of strategies, which we collectively termed as "doing something". Nearly 34% of the forget-group participants reported "not believing" the forget cue. Surprisingly, however, there was virtually no correlation ($\varphi = 0.08$) between forget-cue belief and whether participants engaged in controlled strategies in order to forget. Some participants trusted the forget cue but did nothing in response to it, whereas others engaged in forgetting strategies despite not trusting the cue. It seemed like the latter group decided to "play along" despite having doubts about the truthfulness of the forget cue. Most importantly, L1 recall in the forget group was entirely determined by whether participants engaged in controlled strategies or not, whereas believing in the forget cue was completely irrelevant. When the forgetting strategy (doing something vs nothing) and the belief in the forget cue (yes vs no) were simultaneously entered as predictors of L1 recall in the forget group, only the strategy variable was a significant predictor; the belief variable was not. The results are summarized in Figure 4.2.

Instructions to forget in the real world are sometimes explicit. For example, if construction is blocking a geographical route, one must temporarily "forget the original route" in favor of a more contextually appropriate route (e.g. Golding & Keenan, 1985). However, an implicit instruction to forget may be just as relevant, such as forgetting a prior credit card number to facilitate memory for a recently issued card. We compared the effectiveness of explicit and implicit forget cues and showed that the magnitude of DF is unaffected by the nature of the forget instruction (Foster & Sahakyan, 2011). Regardless of whether forgetting comes from implicit needs of the environment, or explicit prompts of others, it leads to similar behavioral effects.

These findings highlight several key points about DF and suggest new directions for future research. First, similar to prior research, which shows

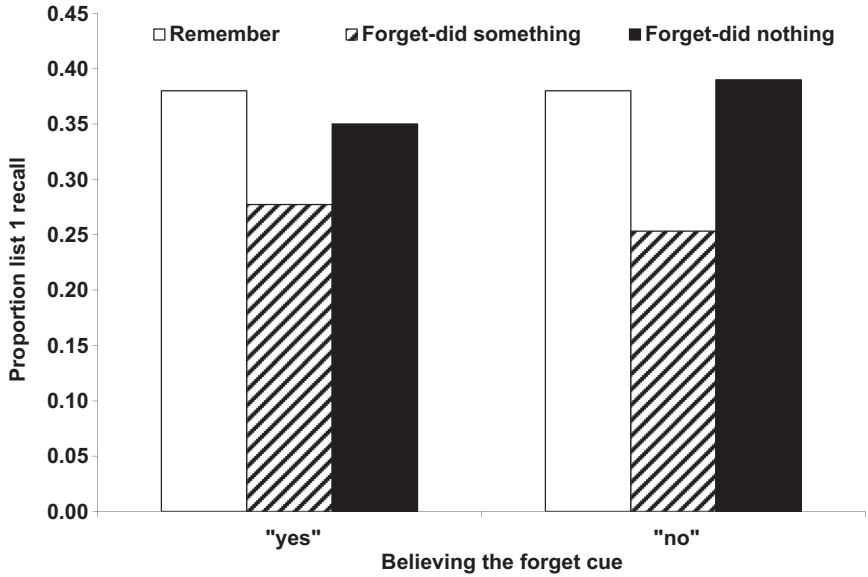


Figure 4.2 Directed forgetting impairment as a function of belief in the cue and controlled forgetting strategies. *(New figure based on the data reported in Foster and Sahakyan (2011)).*

that intention to learn, per se, is not important for remembering, but rather what people do to learn the information is what matters (Hyde & Jenkins, 1973; Mandler, 1967; Postman, 1964), the findings of Foster and Sahakyan (2011) demonstrate that merely believing the forget cue is not sufficient for demonstrating DF; what matters is what people do in response to the forget cue (see also Mulji & Bodner, 2010). Second, because believing the forget cue is not critical for obtaining DF, future research should be less concerned about staging an event to make the forget cue more believable (e.g. simulating a computer crash in the middle of the experiment), and should instead be more vigilant about reporting whether participants engaged in DF or not. For example, the absence of DF in certain conditions or certain populations could simply be linked to "doing nothing". Finally, understanding why some participants engage in DF whereas others do not should become a priority for future research as it will help to provide insights for understanding DF in special populations.

4.2. Why do Some People Fail to Engage in DF?

Are "do nothing" participants simply unmotivated to be in an experiment? If so, the remember group should recall much more than they do. In fact,

though, the "do nothing" people show recall levels comparable to that of the remember group. Maybe some people simply cannot formulate a strategy in a brief period of time between the lists. Most experiments deliver the forget cue and move on to the second list right away. Indeed, some of the participants in our experiment reported "*I didn't do anything in particular, there wasn't enough time*," or "*I wanted to forget but didn't know how to forget, so I didn't do anything*," suggesting that extending the postcue time might increase the chances of obtaining DF or produce larger-than-normal effects. These reports also indicate that one may want to forget but be unable to formulate a strategy. It is essential to be mindful of these issues when interpreting DF performance among special populations, who may have difficulty self-initiating a controlled strategy.

Metacognitive beliefs could also play a role in why some participants fail to engage in DF. For example, if participants have a preexisting belief that their memory is not good even before they attempt to memorize the material in an experiment, then they may be less likely to engage in DF because they may believe that they already forgot. Indeed, a study in our lab with older adults revealed this problem (Sahakyan, Delaney, & Goodmon, 2008). Many older adults, in response to the forget cue, spontaneously volunteered that they had already forgotten. Not surprisingly, they did not engage in DF. We consequently modified DF instructions that emphasized the need to engage in forgetting despite the beliefs that they already forgot, and we found that such instructions were more conducive for obtaining DF with older adults than were the standard instructions.

4.3. Beliefs about Memory Developed during the Experiment

Participants could also develop beliefs about their memory *during* the experiment as they gain experience memorizing the materials. Such beliefs might also play a role in whether they engage in intentional forgetting. For example, if some participants do not feel confident about their ability to remember L1 after attempting to memorize it, they may see no need to engage in DF. Support for this hypothesis comes both from older adults (Sahakyan, Delaney, & Goodmon, 2008) and from a study involving mainly college adults (Sahakyan, Delaney, & Kelley, 2004). After presenting L1, we asked college participants to predict how many words from that list they would be able to recall during the final test (i.e. an aggregate judgment of learning (JOL)). Participants' predictions ranged from the minimum of 20% (equivalent to three words) all the way to 80% (equivalent to 12 words). After providing the JOLs, they received a forget or remember cue, followed by L2.

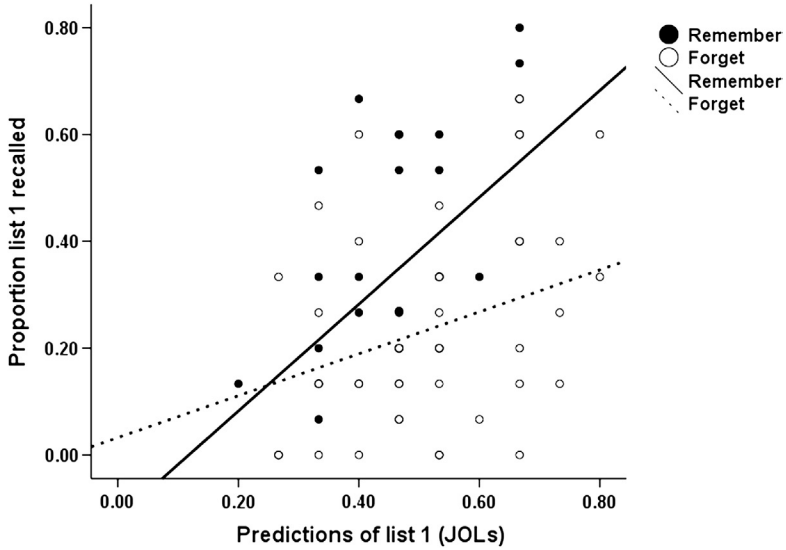


Figure 4.3 Directed forgetting impairment as a function of cue and aggregate predictions of list 1 learning. *(New figure based on the data reported in Sahakyan, Delaney, and Kelley (2004)).*

Figure 4.3 shows L1 recall during the final test as a function of JOLs and the cue. As the figure demonstrates, DF impairment was larger when participants anticipated recalling many L1 items than when they anticipated recalling few L1 items. Although we did not collect forgetting strategy reports in that study, it is possible that the more people expected to remember from the first list, the more likely they were to engage in DF. In contrast, if they felt they could not remember L1 that well, they may have been less likely to engage in DF. Individual differences in JOLs could be correlated with other variables that are involved in this effect (e.g. higher intelligence or working memory capacity), but these data are consistent with the idea that beliefs about memory could contribute to list-method DF (for a demonstration of how item-by-item JOLs affect item-method DF, see Foster & Sahakyan, 2012).

4.4. Preexisting Beliefs about Intentional Forgetting

In an unpublished study, Foster and Sahakyan investigated whether preexisting beliefs about whether it is possible to intentionally forget something might play a role in who engages in DF and who does not. People who have beliefs that it is impossible to forget something after it has been learned might be less likely to attempt DF. We asked participants who were

participating in a departmental mass screening "*Do you think it is possible to make yourself forget something after you have learned it*?" The answer required a yes/no response. Out of 130 participants who responded to this question, 49% indicated "yes", and 51% indicated "no". These participants were later invited to take part in a DF experiment in our lab (they were unaware of the nature of the experiment). At the end of the experiment, they indicated what they did in response to the forget instruction. Their responses were coded as "doing something" or "doing nothing" (similar to Foster & Sahakyan, 2011). Replicating previous studies, 29% of the forget group reported "doing nothing". Interestingly, we found no correlation between peoples' beliefs about whether it is possible to intentionally forget and whether they engaged in DF ($\varphi = -0.04$). Most importantly, the magnitude of DF impairment was virtually identical among those who believed it is possible to intentionally forget and those who did not believe it is possible. What mattered for DF was whether participants engaged in deliberate strategies to forget, replicating previous work of Foster and Sahakyan (2011). The results are shown in Figure 4.4.

Overall, studies from our lab have consistently found that DF does not emerge among participants who report "doing nothing" in response to the

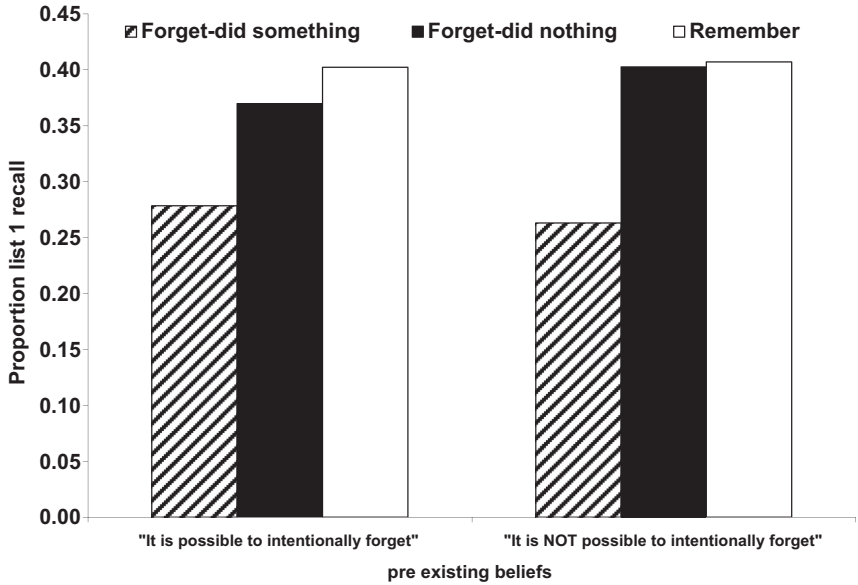


Figure 4.4 Directed forgetting impairment as a function of preexisting beliefs about intentional forgetting and controlled forgetting strategies. *(Based on unpublished data collected by Foster and Sahakyan.)*

forget cue. Engaging in DF does not appear to be linked to whether participants trusted the forget cue, or whether they think it is possible to forget things once they have been learned. A better understanding of why some healthy college adults do not engage in DF is warranted, because it could help explain DF performance in special populations. Failures to obtain DF with certain populations may not necessarily reflect an inability to inhibit. Instead, they may indicate that certain populations may select to do "nothing" in response to the forget cue, but the reasons for such decisions could be more complex and not necessarily linked to their ability to purposefully forget. They may be unable to formulate a strategy, or they may not see a need to engage in forgetting.

4.5. Forgetting Strategies and Their Relative Effectiveness

In recent years, Nathan Foster and Lili Sahakyan collected many strategy reports from participants in list-method DF studies. We did not analyze them in published papers due to smaller sample sizes. For this review, we combined several studies to compile a dataset of 290 participants (with 145 in the forget group and 145 in the remember group) to investigate whether DF depends on forgetting strategy or not. In each study, participants studied 15 unrelated nouns per list at 4 s/item, with the remember or forget cue between the lists, but no other break/delay. A short filler task followed L2, and then L1 was tested before L2. Then participants in the forget group were asked, "*What did you do to forget L1*? If you clearly remember what you did, please write that down. If you do not remember clearly, please do not make up anything at this point. If you did not do anything in particular to forget the words, or if you specifically tried to remember them, please indicate this."

The responses were grouped into six categories, and recall was assessed as a function of different strategies. The results are summarized in Figure 4.5. Reports of "stopping repeating L1 words," or "engaging in diverting thoughts" are more unambiguously interpretable compared to reports like "stopping thinking about L1 words," "clearing one's head," or "focusing on the upcoming list," because it is less clear how they went about those processes. Thus, interpreting the relative effectiveness of these strategies requires that we be mindful that some of these strategies may not be mutually exclusive.

It is nevertheless clear from Figure 4.5a that to obtain DF impairment, participants had to engage in some controlled behaviors; doing nothing in response to the forget cue did not produce DF impairment. Additionally,

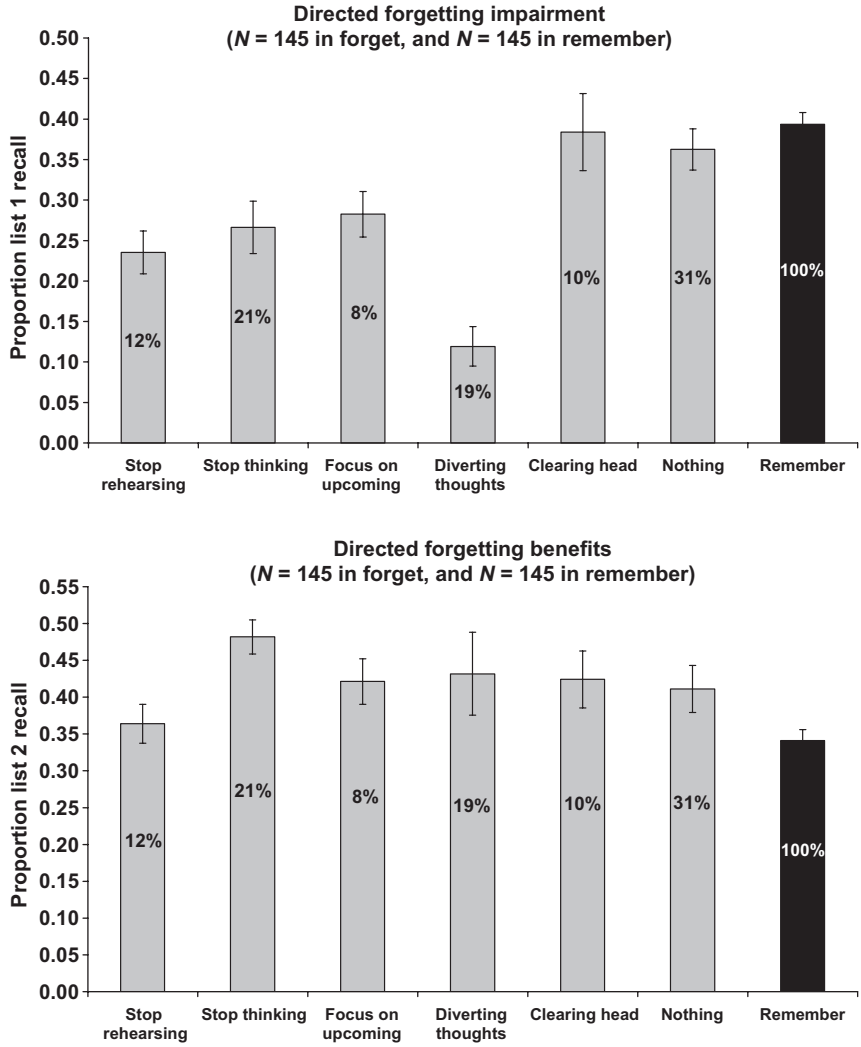


Figure 4.5 Directed forgetting costs (a) and benefits (b) as a function of forgetting strategies reported by forget-group participants in a dataset containing $N = 145$ forget-group participants and $N = 145$ remember-group participants. The black bar represents the remember condition, and the gray bars represent the subgroups of the forget condition broken down by reported strategy. The numbers inside the bars indicate the percentage of participants in each condition.

engaging in diversionary thoughts produced the biggest DF impairment compared to other strategies, some of which did not even produce DF (e.g. clearing one's head). Finally, despite the variability in the size of the DF impairment as a function of various strategies, nearly all strategies led

to DF benefits, including "doing nothing" to forget (Figure 4.5b). We will revisit this issue in Section 7, where we specifically discuss the mechanisms producing DF benefits.

5. CONTEXT CHANGE AS AN EXPLANATION FOR DF IMPAIRMENT

Once forgetting is strategically initiated, we believe that it takes place because of changes in mental context—an explanation first proposed by Sahakyan and Kelley (2002). Our thinking was greatly influenced by formal models of episodic memory, according to which people encode both the *content* of the item during learning and various *contextual features* or attributes that are present in the background (e.g. Anderson & Bower, 1972; Estes, 1955; Gillund & Shiffrin, 1984; Hintzman, 1988; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Tulving, 1983). Contextual information includes environmental, spatial–temporal, emotional, and mental states in which the item is experienced. Context cues play a particularly important role in free recall because the initial retrieval attempt is solely guided by the context cues that participants use to search their memory (e.g. Howard & Kahana, 2002; Mensink & Raaijmakers, 1988). The success of retrieval is a function of the overlap between the cues used to search memory and the contents of memory. Recall suffers when there is a low overlap between the contextual features present at encoding and the contextual features present at test (e.g. Godden & Baddeley, 1975).

Based on these ideas, Sahakyan and Kelley (2002) proposed a mental context-change account of DF, according to which the forget instruction encourages participants to abandon the contextual cues that were prevalent during L1 encoding and to sample new contextual cues for L2 encoding, thereby segregating the two lists as separate events. The DF impairment arises because retrieval context better matches L2 than L1 encoding context, producing forgetting of L1 items in the forget group. The original context-change account further explained that DF benefits arise because of reduced proactive interference on L2 due to contextual differentiation (Sahakyan & Kelley, 2002). We will argue later that changes in encoding strategy produce a larger impact than the reduced proactive interference in most studies, but that the reduced proactive interference effect is likely a real effect that is difficult to assess through overall recall rates.

The rationale for proposing the mental context-change account was triggered by the reports of participants who indicated engaging in distracting

thoughts in order to forget. Figure 4.5 confirms that this strategy leads to the largest DF impairment (although other strategies could also contribute to DF). We reasoned that engaging in diversionary thoughts could change peoples' mental context between L1 and L2. Recently, Lehman and Malmberg (2009) developed a computational model of DF within the framework of search of associative memory, by capitalizing on our notion of context change between the lists, and the difficulty reinstating L1 context during the test in the forget group.

In this section, we review the findings from various conditions that were predicted to interact with list-method DF based on the context-change mechanism along with other theoretically relevant findings.

5.1. Mental Context-Change Paradigm and Its Relation to DF

In the process of empirically testing the context-change account of DF, Sahakyan and Kelley invented what has come to be known as the *mental context-change/diversion paradigm*. This paradigm involves asking participants to engage in diversionary thoughts following L1 learning, and examining the impact of such activity on memory. Several years of investigation have demonstrated that mental context-change produces forgetting much like the explicit instructions to forget. A variety of tasks were shown to create forgetting, such as imagining being invisible (Sahakyan & Kelley, 2002), thinking about the childhood home (e.g. Sahakyan & Delaney, 2003), daydreaming about vacations (e.g. Delaney, Sahakyan, Kelley, & Zimmerman, 2010), or wiping a computer monitor (e.g. Mulji & Bodner, 2010). In contrast, other tasks do not lead to forgetting, such as solving arithmetic problems (Sahakyan & Delaney, 2003), counting forward or backward (Sahakyan, Delaney, & Goodmon, 2008), briefly chatting with an experimenter (Aslan & Bäuml, 2008), or waiting quietly (Sahakyan & Kelley, 2002). Why some tasks create mental context-change whereas others do not is not fully clear and requires future research.

Whenever context-change effects were found, they were found irrespective of encoding strategies (Sahakyan & Delaney, 2003), across individual differences in working memory capacity (Delaney & Sahakyan, 2007), across age-related differences (Aslan & Bäuml, 2008; Sahakyan, Delaney, & Goodmon, 2008), across serial position effects (Sahakyan & Foster, 2009), and the boundary conditions that determine whether DF is obtained, such as the need for L2 learning (Pastötter & Bäuml, 2007; but see Unsworth et al., 2012). Thus, the mental context-change paradigm produces many behavioral effects that are similar to DF. Interestingly, the retrieval dynamics

in the mental context-change paradigm were shown to be similar to the environmental context-change paradigm across several dependent measures including recall latencies, serial position effects, first response functions, and the patterns of response transitions during recall captured through conditional response probability functions (Unsworth et al., 2012). Thus, the mental context-change paradigm produces behavioral effects similar to not only DF but also the environmental context-change paradigm.

Some recent findings, however, have revealed differences across DF and the mental context-change paradigm. For example, although Pastötter and Bäuml (2007) reported that L2 learning is critical for obtaining both DF and mental context-change effects, Unsworth et al. (2012) obtained significant forgetting using a mental context-change paradigm without involving L2 learning. Several methodological differences exist between these two studies, most notably the list length, which was much longer in the Unsworth et al. study (40 items) than is typically employed in the DF studies. The use of such long lists may affect the likelihood of reinstating the context of the items from the beginning of the list because those items were studied long ago. Future research should consider more rigorous investigation of the list-length factor, including examining whether long lists might also lead to a DF effect without the need for L2 learning.

5.2. Recognition Tests and Indirect Tests of Memory

The majority of list-method DF studies failed to obtain a DF effect on a standard recognition test (e.g. Basden et al., 1993; Bjork & Bjork, 2003; Block, 1971; Conway et al., 2000; Elmes, Adams, & Roediger, 1970; Geiselman et al., 1983; Gottlob & Golding, 2007; Smith, Barresi, Gross, & 1971; Racsmány, Conway, Garab, & Nagymáté, 2008; Reitman, Malin, Bjork, & Higman, 1973; Whetstone, Cross, & Whetstone, 1996; Zellner & Bäuml, 2006). Although two relatively recent studies obtained DF benefits on L2 using longer study lists than were previously employed (e.g. Benjamin, 2006; Sahakyan & Delaney, 2005), they too failed to obtain DF impairment. In contrast to the list-method, the item-method DF effect was always obtained on recognition tests (e.g. Basden et al., 1993; MacLeod, 1999). Implicit memory tests have also been shown to be insensitive to the list-method DF, including both word stem and fragment completion tests (Basden et al., 1993; Bjork & Bjork, 1996; but see Koppel & Storm, 2012), tests of primed word association or general knowledge (Basden & Basden, 1996, 1998), and lexical decision (Racsmány & Conway, 2006). In all these tasks, forget items show priming that is equivalent to remember items.

Because performance on the forget and the remember items did not differ on recognition tests (or on implicit tests), the inhibitory account proposed that those items were released from inhibition by virtue of being represented during the test (e.g. Bjork, 1989; Bjork & Bjork, 1996, 2003; Geiselman et al., 1983; Zellner & Bäuml, 2006). Although this explanation is still rather popular, we think of it more as a description rather than a specific assumption or an explanation offered by theory. The release of inhibition was proposed mainly to reconcile the discrepant findings across the list and the item methods. However, there is nothing *a priori* in the inhibitory account to propose a release from inhibition on certain tests. For example, assumptions of release are not made for the RIF phenomenon, which was discovered much later than DF, and which is usually observed on recognition and lexical decision tests (e.g. Aslan & Bäuml, 2011; Hicks & Starns, 2004; Soriano, Jiménez, Román, & Bajo, 2009; Spitzer & Bäuml, 2007, 2009; Veling & van Knippenberg, 2004; Verde, 2004). If list-method DF was observed on various memory tests just like the item-method, assumptions of release would not be made and the inhibitory interpretation of DF would probably not even be proposed.

We have interpreted DF from the perspective of our contextual account, and have a different take on the null effects of DF in various direct and indirect tests. For example, the absence of DF in implicit memory tests is consistent with the contextual account. Implicit tests do not require retrieval of contextual information because they do not direct attention to the original study episode. Since they do not require reinstating prior episodic context, we would not expect to observe DF on such tests. Environmental context effects also are not found on implicit tests (e.g. Parker, Gellatly, & Waterman, 1999).

The absence of DF in recognition is consistent with a body of literature that shows that environmental context effects typically are not detected in recognition (e.g. Godden & Baddeley, 1975). That said, the meta-analysis of that literature revealed that under certain conditions, environmental context effects do emerge in recognition, and those effects are larger when the encoding processes are primarily nonassociative in nature (for a review, see Smith & Vela, 2001). For example, a number of studies using nonwords or unfamiliar faces as stimuli detected environmental context effects in recognition (e.g. Dalton, 1993; Krafka & Penrod, 1985; Macken, 2002; Malpass & Devine, 1981; Russo, Ward, Geurts, & Scheres, 1999; Smith & Vela, 1992). Building on those findings, we predicted and obtained DF in recognition using nonword stimuli, despite not obtaining such effects using words (Sahakyan, Waldum, Benjamin, & Bickett, 2009).

In another experiment, we obtained DF in recognition using word stimuli by manipulating whether the test required distinguishing the targets from similar or dissimilar distractors (Sahakyan et al., 2009, Experiment 3). Similar distractors involved plurality-reversed versions of the target items, whereas dissimilar distractors were novel words. We hypothesized that plurality discrimination would engage more direct retrieval of contextual information during the test, enabling us to detect DF in recognition even with word stimuli. Indeed, the results confirmed our predictions: There was no DF in the dissimilar distractor condition, but there was DF in the similar distractor condition.

In the same year, Lehman and Malmberg (2009) reported DF in recognition of word stimuli using an exclusion/inclusion manipulation. They reasoned that typical recognition tests ask participants to endorse any item studied during the experiment, regardless of which list it appeared on (i.e. inclusion condition). Under such conditions, participants do not have to rely on context cues to differentiate one study episode from another, and they can rely on item familiarity as the basis of recognition. In contrast, in an exclusion condition, participants have to endorse only the words from a designated list (either L1 or L2), and to reject the words that come from a different list (or new words). To perform such a task accurately, participants have to rely more heavily on context cues to differentiate one list from another. Overall, the exclusion recognition is closer to what is needed to perform free recall than inclusion recognition is, and indeed Lehman and Malmberg (2009) predicted and obtained robust DF under exclusion instructions.

Plurality discrimination, exclusion recognition, and nonword recognition all rely on retrieval of contextual information, and hence all these conditions showed DF in recognition. Recently, Racsmány, Conway, Garab, and Nagymate (2008) used a remember/know procedure and also obtained DF in recognition, but only for "remember" responses. Although one might be tempted to conclude that DF impairs all types of contextual information, the results from associative recognition further refine the nature of contextual information that is important in DF. Specifically, Hanczakowski et al. (2012) did not obtain DF on an associative recognition test, despite it being a recollection-laden test. Although there was no effect in associative recognition, within the same study, DF affected list discrimination judgments. Therefore, the authors argued that DF does not lead to a generalized impairment in recollection of all contextual details. Instead, DF affects primarily global contextual information that differentiates L1 from L2, without necessarily impairing interitem associations between the words within the

list. Finally, Gottlob and Golding (2007) demonstrated that despite the lack of list-method DF in the item recognition, the forget instruction impaired source memory for color and case of L1 items.

Overall, the reports of significant list-method DF in item recognition suggest that the typical conditions of recognition rather than the recognition test itself were the reasons for the observed lack of DF in previous research. When stimuli or test conditions encouraged greater utilization of global contextual information, DF was observed in recognition. These results challenge the inhibitory interpretation because it does not *a priori* predict the conditions under which the release can or cannot be obtained. In contrast, these results fit well with the context account. As the need to retrieve context increases, the likelihood of seeing list-method DF in item recognition or recognition of source also increases.

One of the earlier explanations of DF involved selective rehearsal, and some researchers lately acknowledged that this account should not be prematurely dismissed for intentional learning (e.g. Benjamin, 2006; Sheard & MacLeod, 2005). Considering that people often report stopping rehearsal in response to the forget cue and that engaging in this type of strategy produces DF impairment, we briefly discuss the results of recognition findings from the perspective of selective rehearsal. The absence of DF in recognition tests is particularly problematic for the selective rehearsal account because recognition and recall are sensitive to elaborative rehearsal (Geiselman & Bjork, 1980), and recognition is even more sensitive than recall to rote rehearsal (Benjamin & Bjork, 2000; Craik & Watkins, 1973). Thus, regardless of nature of the rehearsal process involved in DF, terminating rehearsal of to-be-forgotten items should have negative consequences for recognition memory. We note that recognition is typically not at ceiling despite the short lists used in many DF studies, and even longer lists do not lead to DF impairment in recognition (Benjamin, 2006; Sahakyan & Delaney, 2005). Thus, any explanation that attempts to revive the selective rehearsal account must successfully deal with this limitation, and should also be able to explain why recognition effects emerge under the specific conditions described above.

5.3. Mental Reinstatement of Context

The context-change account assumes that participants rely on the recent L2 context to probe their memory for L1, and because that context mismatches the L1 context in the forget group, it leads to impaired access to L1 items. The context account predicts that if participants could reinstate L1 context at the time of test, this should reduce DF.

The environmental context-change studies imply that participants do not automatically reinstate the original environmental context when they are tested in a new context, otherwise we would not be observing forgetting (e.g. Godden & Baddeley, 1975). Either participants do not attempt contextual reinstatement in those studies, or they attempt to reinstate the original context but are unsuccessful at it. We suspected that reinstatement of mental context (i.e. internally generated context) might be even more challenging than reinstatement of environmental context. To make L1 context more memorable and distinctive, we played the music from Star Wars at the start of the experiment. Before final test, Sahakyan and Kelley (2002) asked participants to mentally reinstate the episodic context that was prevalent during L1 encoding by guiding them through similar types of steps as previously employed in the environmental context literature (e.g. Smith, 1979). Reinstatement of L1 context did not eliminate the DF effect. However, it significantly reduced it both in the mental context-change condition and in the forget group, suggesting that the interaction between the contents of memory and the contextual cues used during the test is a critical ingredient for obtaining DF.

If L1 items were inhibited as a result of DF, then why did they recover from contextual reinstatement not just in the context-change group but also in the *forget* group? To explain the recovery, the inhibitory account would have to assume that context plays an important role in retrieval, and that changing or reinstating the context at the time of final test has consequences for memory. Because these are general claims made by many formal theories of memory, it is unclear what is gained by postulating an inhibitory mechanism if it must still explain the results by resorting to the contextual mechanisms.

5.4. Cuing with Context Cues vs Category Cues

Recent research manipulated the structure of the study list combined with specific ways of cuing memory at the time of test, and predicted the factors that should eliminate DF vs preserve it (Lehman & Malmberg, 2011a). Participants studied two lists. L2 always contained unrelated items, but L1 consisted of unrelated items for some participants vs items drawn from a single category for others. On the test, memory was probed either with a context cue (e.g. "retrieve the items from L1") or using the category cue (e.g. "use the word *clothing* to help you retrieve the words from L1"). The authors predicted that if DF impairment is due to context change, then it should be reduced with category cues, which will reinstate the specific L1 context

more than will more general temporal context cues (e.g. Raaijmakers & Shiffrin, 1981). Successful initial recall would in turn provide a good retrieval cue for subsequent items. Replicating prior work, DF was obtained with unrelated lists. There was also significant DF with the categorized list as long as the test cue was just the temporal context. In contrast, when a category cue was used to test memory, the costs were eliminated, as predicted. Lehman and Malmberg's (2011a) findings highlight the importance of the interaction between retrieval cues and the contents of memory at test: The same categorical list can either show DF or be released from DF depending on the type of retrieval cues utilized during the test. The more specific the cue is to the L1 study context, the less likely DF is to be observed.

Note that another study conducted by Wilson, Kipp, and Chapman (2003) obtained a different outcome with categorical lists, arguing that categorical lists were resistant to DF. However, before the study of each list, Wilson et al. (2003) drew special attention to the fact that the lists in their study were either unrelated or categorical in nature. Their test instructions were not described in sufficient detail to infer what type of retrieval cues were given to participants. Thus, methodological differences between the two studies may be responsible for the discrepant findings. Note that Sahakyan (2004) also obtained intact DF using different categorical items across the lists while probing memory with temporal context cues.

5.5. Encoding Strength: Item vs Context Strength

Does deeper/elaborate encoding of material protect against DF? Sahakyan and Delaney (2003) had participants encode both lists either via the story mnemonic or via rote rehearsal. Despite robust differences in overall memory performance, they obtained equivalent DF impairment in both encoding conditions, suggesting that interactive encoding does not protect against DF.

Our subsequent research suggests that certain ways of strengthening items make them more susceptible to DF than other ways of strengthening (Sahakyan, Delaney, & Waldum, 2008). Relying on research by Malmberg and Shiffrin (2005), who suggested a more complex view of encoding strength by differentiating it into an item strength and a context strength, we derived and tested predictions of the contextual account of DF. Specifically, Malmberg and Shiffrin (2005) suggested that certain strengthening manipulations increase the strength of the items without affecting their associations with the list-context (e.g. depth of processing or extrastudy time), whereas other manipulations lead to strengthening of both item and contextual information in the memory trace (e.g. spaced presentations).

Based on their findings, we predicted that spaced items should suffer more when context changes because item-to-context associations are stronger for spaced items. Indeed, our study showed that spaced items showed larger DF impairment than did massed items. Similar results have also been obtained in an environmental context-change study (Isarida & Morii, 1986). In contrast to the spacing manipulation, we predicted that the depth of processing or the extrastudy time manipulations should not interact with DF, and indeed the results confirmed our predictions. Even though items studied for longer duration were remembered better than items studied for shorter duration, they showed equivalent impairment from DF. Taken together, these results suggest that what determines interactions with DF is not the strength of the items per se, but rather the strength of item-to-context associations. Along the same lines, Waldum and Sahakyan (2012) found that statements incongruent with one's attitude show greater DF impairment than do congruent statements. Using a variety of memory paradigms, they confirmed that greater DF of incongruent statements was driven by stronger associations with episodic context, entirely consistent with the context account.

5.6. Contextual Variability

In an unpublished study, Sahakyan and Foster crossed DF with manipulation of context variability (CV), which is the number of different semantic contexts in which a given item occurs (e.g. Dennis & Humphreys, 2001; Steyvers & Malmberg, 2003). The rationale for their study was that although low-CV items have a memory advantage in both recognition and free recall, they tend to suffer more from environmental context manipulations between study and test (Marsh, Meeks, Hicks, Cook, & Clark-Foos, 2006). Marsh et al. (2006) argued that during encoding, low-CV items form stronger item-to-context associations compared to high-CV items due to a smaller fan of preexisting contextual associations. Therefore, when context changes between the study and the test, low-CV items suffer more than high-CV items do.

Consequently, we expected that DF should also affect low-CV items more than high-CV items. We tested 46 participants randomly assigned to forget or remember groups. Each participant studied two lists of 14 items selected from Steyvers and Malmberg's (2003) appendix. Half of the items within each list were low-CV and the other half was high-CV words. Because CV correlates with word frequency (Steyvers & Malmberg, 2003), we controlled word frequency by selecting the items from the low-frequency category similar to what was done by Marsh et al. (2006). During

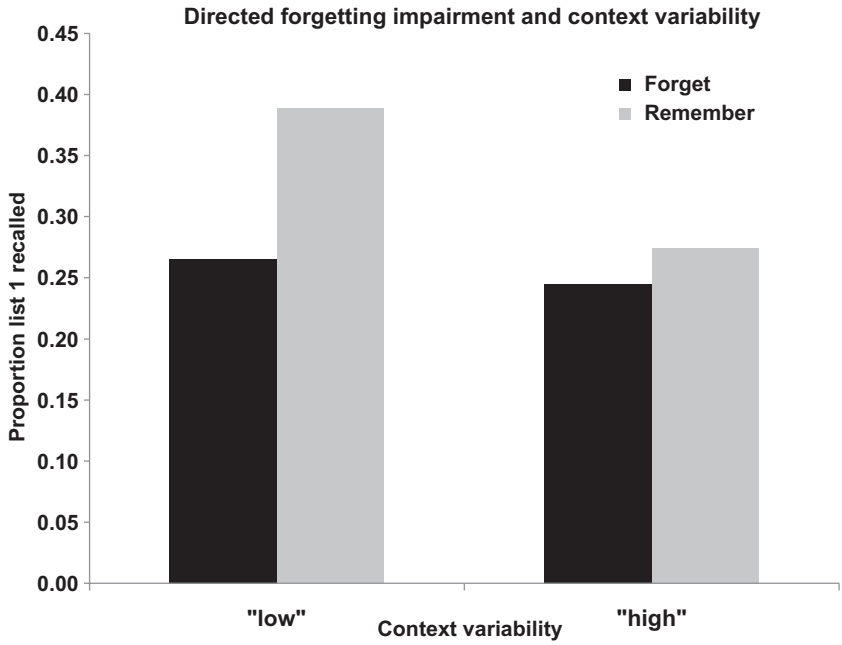


Figure 4.6 Directed forgetting impairment as a function of context variability of studied items. *(Based on unpublished data.)*

the test, L1 was always tested before L2. Figure 4.6 summarizes the findings. DF impairment was much greater for low-CV items than for high-CV items, consistent with the predictions of the context account.

5.7. Extralist Cued Recall

Most list-method DF studies employ free recall tasks partly because recognition and indirect memory tests are insensitive to DF. Sahakyan and Goodmon (2010) used a different testing procedure that allowed them to contrast the predictions of the inhibitory and context accounts. Specifically, participants studied two lists of items (called targets). The test provided cues that were meaningfully related to the targets, but were never studied with them (hence the name extralist cued recall). For example, the studied word *planet* could be tested with the cue *universe*. Extralist cued recall is affected both by target characteristics and by cue characteristics (for reviews, see Nelson, McKinney, Gee, & Janczura, 1998; Nelson & McEvoy, 2005), and it provides an excellent opportunity to test the opposite predictions of the inhibitory account and the context account within the same study.

Sahakyan and Goodmon (2010) conducted five experiments manipulating different sources of implicit associative strength. Crucially, some of their experiments varied the implicit strength of the targets while holding the cue strength constant, whereas other experiments held the implicit strength of targets constant while varying the strength levels of the test cues. Across all experiments, they found greater DF from the strong conditions compared to the weak conditions, regardless of whether strength referred to the targets or the test cues. This pattern of findings is exactly what is predicted by the contextual account: Presentation of the extralist cue should not diminish the importance of recovering contextual information during the test. Reinstating the context helps to differentiate the studied item from other items implicitly activated by the test cue. According to the PIER model which explains extralist cued recall (Nelson, Goodmon, & Ceo, 2007), context cues and extralist cues combine together to elicit recall of the target. Thus, the context account makes a strong prediction that there should be greater effects of DF in the strong than weak conditions regardless of whether the study varied the target strength level or the cue strength level.

In contrast, the inhibitory view assumes that items can be released from inhibition with provision of certain cues (e.g. copy cues of the items on the recognition test). Hence, the inhibition account would predict that stronger test cues would be more successful at releasing the targets from inhibition than weaker cues, and hence DF should be smaller not greater in the conditions where cue strength is high. The results did not support this prediction. The findings of Sahakyan and Goodmon (2010) suggest that items/targets were not inhibited at the level of individual item representation because there is no way for the cognitive system to determine which targets should be inhibited more and which targets should be inhibited less until a test cue is provided. The effects in their study were instead entirely driven by the combination of test cues and context cues, consistent with predictions of the context account.

5.8. Importance of L2 Learning

A number of studies revealed boundary conditions that are needed to observe DF impairment. For example, replicating prior work by Gelfand and Bjork (1985; cited in Bjork, 1989), Pastötter and Bäuml (2007) showed that in the absence of L2 learning, one does not observe DF impairment. They later showed that DF is reduced with decreasing length of L2 (Pastötter & Bäuml, 2010). Finally, Conway et al. (2000) as well as Macrae, Bodenhausen, Milne,

and Ford (1997) demonstrated that divided attention during L2 learning by a concurrent task reduces DF impairment.

According to the inhibitory view, in the absence of L2 learning, there is no need to inhibit L1 items because inhibition is an adaptive mechanism invoked to reduce interference (e.g. Conway et al., 2000; Barnier et al., 2007). When there is no L2 learning, there is no need to invoke inhibition. The inhibitory view thus assumes that L1 items intrude during L2 learning and hence must be inhibited (e.g. Conway & Fthenaki, 2003; Racsmány & Conway, 2006). However, in the domain of RIF, it is generally assumed that simply studying a list of items should not trigger inhibitory processes; inhibition should be triggered only in response to competitive retrieval (e.g. Anderson, 2003; Storm & Levy, 2012). It is not fully clear why the mere study of L2 would trigger competitive retrieval of L1 items.

According to the context-change account, DF arises not simply because of context change between the lists, but because participants experience difficulty reinstating the context of the previous list. In the absence of L2 learning, reinstating the context of the only list that was studied is not an issue because it was the only list that was studied (e.g. Jang & Huber, 2008). Thus, according to the context-change view, one would not expect to observe DF. In contrast, in the presence of L2, participants have to rely on context cues to distinguish between the two lists and reinstate the context of the previous list. Also, in the environmental context literature, the effects of the first-order paradigms, where a single list is studied and tested either in the same or a different context, are much smaller compared to the second-order paradigms that involve multiple lists (e.g. Eich, 1985; Fernandez & Glenberg, 1985; for a review of first- and second-order paradigms, see Bjork & Richardson-Klavehn, 1989). We interpret the impact of L2 length from a similar perspective. Enhanced length of L2 contributes to greater distancing of L1 context from the time of test and, as it gets more distant, the reinstatement of that context would also get more difficult, explaining why shorter L2 reduces DF. Finally, we interpret the results of divided attention as reflecting attentional demands of context encoding and context change. We expand on this view in the section discussing working memory interactions.

5.9. Related Items Across Lists

Several studies have shown that when items are related across the lists, DF impairment is eliminated (e.g. Barnier et al., 2007, Experiment 5; Conway et al.,

2000; Sahakyan & Goodmon, 2007). These findings are not consistent with an inhibitory viewpoint that assumes suppression at the level of items. Specifically, one would assume that when the two lists involve related items, then L1 items may be more likely to come to mind during L2 learning and require greater inhibition compared to when the two lists are unrelated. However, the results show that related items across the lists *reduce* rather than increase DF.

Although items across the lists may be related to each other in a number of ways, our research showed that when L2 items remind participants of L1 items due to backward associations (e.g. *chip*← *chisel*), DF is eliminated, but when L1 items are related to L2 items via forward associations (e.g. *chip*→ *wood*), DF remains intact (Sahakyan & Goodmon, 2007; for arelated result using the item procedure, see Golding, Long, & MacLeod, 1994). We proposed several mechanisms by which related items could reduce DF. First, reminding could reinstate the context of L1 during L2 learning, thereby preventing or reducing contextual differentiation between the lists. Second, reminding could initiate retrieval of L1 items during L2 learning, thereby linking L1 items to the context of both lists. This would enhance the probability of their retrieval even if L2 context is used as the main retrieval cue during the test. Finally, reminding could strengthen L1 items via retrieval during L2, and enhanced strength of L1 items could also make them resistant to DF.

5.10. Retrieving L2 before L1

The meta-analysis suggests that output order does *not* affect the magnitude of DF impairment (Pastötter et al., 2012). In other words, whether L1 is tested first in the recall sequence, or whether it is tested after L2, the magnitude of DF impairment remains invariant. This is inconsistent with the inhibitory explanation because if L2 is tested first, then according to the inhibitory account, L1 items should be inhibited to allow retrieval of L2. Thus, the DF impairment should be larger when L1 is tested after L2 than when it is tested first in the recall sequence. However, Pastötter et al. (2012) showed that retrieving L2 first does not increase the magnitude of forgetting.

An even more extreme manipulation was employed by Basden, Basden, and Morales (2003), who gave multiple retrieval trials on L2 before the final test. Despite multiple retrieval trials on L2, Basden, Basden, and Morales (2003) and Basden, Basden, and Wright (2003) did not observe negative effects of such retrieval on L1 memory, even when their lists consisted of categorical items from the same taxonomy. Overall, these results suggest

that retrieving L2 before L1 has no detrimental effects on DF impairment, contrary to what would be predicted by the inhibitory account.

5.11. Delay Effects

There is a scarcity of published studies that have examined DF after a delay. If DF impairment arises from the mismatch of the study and test contexts, then the impairment should dissipate with the passage of time because with delay, context no longer would favor L2 context over L1, and thus one would expect spontaneous recovery of L1 (e.g. Mensink & Raaijmakers, 1988). Similar predictions would be made also by the inhibitory account because items should not remain in an inhibited state indefinitely. We found only one published study that employed a delay manipulation with a full-DF design (Shapiro, Lindsey, & Krishan, 2006). The delay involved 15 min, and participants were tested either immediately, or after delay (but never in both). The immediate test revealed the costs and the benefits of DF. In contrast, delay eliminated DF costs, but did not affect the benefits, which remained preserved. Identical results were reported by Liu Xun in his unpublished dissertation (2001). The presence of significant L2 benefits is consistent with the encoding-based interpretation of DF benefits (Pastötter & Bäuml, 2010; Sahakyan & Delaney, 2005) because stronger encoding of L2 items should be preserved relatively in delay (e.g. MacLeod, 1975). Thus, the dissociating effect of delay is consistent with the dual-factor account of DF.

Two other delay studies were briefly discussed in two chapters (Basden & Basden, 1998; MacLeod, Dodd, Sheard, Wilson, & Bibi, 2003), and they reported significant DF in a delayed condition. However, neither study included a remember control group, and DF was inferred from the R–F measure. Given that delay preserves the benefits (Liu, 2001; Shapiro, Lindsey, & Krishnan, 2006), the difference between the two lists would probably still be significant after a delay. Most important, the brief description of the methods in both studies makes it difficult to know whether the same participants took part in both the immediate and the delayed test, or whether delay was varied strictly between subjects. This could be a critical detail given the effects of testing on retention (e.g. Roediger & Karpicke, 2006). If the same participants participated in both conditions, then the immediate test could preserve DF on the delayed test. In other words, significant DF on the delayed test in some studies may be due to a testing effect. More research is needed to fully investigate delay effects in DF.

5.12. Individual Differences in DF as a Function of Working Memory Capacity

Another feature of the contextual change account is that attentional control has been theoretically linked to effective use of context (e.g. Unsworth & Engle, 2008). Therefore, one might expect that people with higher working memory capacity, which is an index of attentional control, would show larger effects of the forget cue and of the context-change task, as they would be better able to manipulate context. Indeed, this appears to be the case in several recent studies. Delaney and Sahakyan (2007) showed that in a typical two-list design, both L1 and L2 memory were positively correlated with working memory capacity. (This effect disappeared when everyone was instructed to make up a story using all the words on the list). However, working memory capacity was negatively correlated with L1 memory in the forget and context-change conditions in two experiments. In other words, the DF and context-change manipulations produced substantially more forgetting for high-span than for low-span participants. Consistent with the context-change account, both the context-change task and the forget task interacted with working memory capacity in the same manner. This basic result has been replicated with adults and children (Aslan, Zellner, & Bäuml, 2010), and in some reports, the benefits of DF were also linked to working memory capacity (Soriano & Bajo, 2007).

The effects of dividing attention during L2 (e.g. Conway et al., 2000) are consistent with the notion that context encoding requires attentional resources. Dual task could interfere with context encoding during L2, thereby giving contextual advantage to L1 compared to L2. It could also make the two lists sufficiently differentiated even in the remember group. Note that Conway et al. (2000) did not include a condition where attention was divided during both lists. Thus their attentional manipulation was confounded with the list.

5.13. Reexposure or Retrieval of L1 Items Affects Subsequent Memory

Goernert and Larson (1994) examined how part-set cuing affects list-method DF and found that when participants were provided with a subset of L1 items on the final test to use as cues for retrieval, their recall of remaining L1 items showed improvement in the forget group, whereas it suffered in the remember group compared to the noncued condition. These effects were replicated and extended by Bäuml and Samenieh

(2010, 2012a, 2012b), who showed that both part-set cuing and also selective retrieval of a subset of L1 items can enhance recall of the remaining L1 items in the forget group, while the same manipulations impaired recall in the remember group. Similar effects were obtained in the mental context-change paradigm (Bäuml & Samenieh, 2012a, 2012b). Prior studies also showed that reexposure to L1 items on an intervening test can release subsequent DF depending upon the nature of the intervening test. The typical design of these studies involves (1) giving an initial test, (2) reexposing all or a subset of items from both lists on an intervening test, and (3) administering a final recall test. The intervening tests included recognition (Basden et al., 1993; Basden et al., 2003; Bjork & Bjork, 1996), word fragment completion (Basden et al., 2003; Bjork & Bjork, 1996), implicit free association test (Basden et al., 1993), or a lexical decision test intermixing all studied items with unstudied items and nonwords (Racsmány & Conway, 2006). Although initial recall revealed significant DF in these studies, the intervening tests did not (neither recognition, nor lexical decision, nor fragment completion showed evidence of DF). Importantly, the final recall findings differed depending upon the nature of the intervening task. When the intervening test involved a recognition test that included a subset of L1 items as lures, DF was released on a final recall test. In contrast, intervening word fragment completion or lexical decision tests did not release DF.

We interpret these findings to suggest that retrieval of contextual information associated with the study episode is critical to reinstating access to L1 items (similar arguments have been made by Bauml et al. for the part-set cuing and part-set retrieval effects in the forget group). The fragment completion test is an indirect memory test because it does not direct the subject to go back to the original learning episode, and it is considered mainly data driven. Likewise, the lexical decision task does not require reinstating prior episodic context. The intervening recognition test, however, is a direct measure of memory because the subject is instructed to think back to the learning episode and make a decision about a specific item (note that Bjork & Bjork, 1996 included an exclusion instruction on the intervening test of recognition, which would further facilitate retrieval of contextual information). Thus, DF on the final free recall test is reduced when participants have to think back to the original study episode and reinstate its episodic context. In contrast, DF is unaffected by indirect tests that do not make any connection to the earlier study episode.

5.14. Summary

The evidence that contextual change underlies DF impairment comes from a wide range of experimental sources. The idea came about because participants in DF studies self-reported thinking of something else, and subsequent research confirmed that this strategy was effective in producing forgetting. Furthermore, context is used to explain many other findings in memory, so it is not invented specifically to explain DF, and can be modeled in context-based formal models of memory. Among the findings that are predictable from a context viewpoint are that recognition does not show DF unless it relies on retrieval of contextual information; that mentally reinstating the original study context releases DF; that indirect memory tests do not show DF because they do not redirect back to the context of the study episode; that items more strongly associated with episodic context show greater DF; that cuing the items with stronger extralist cues leads to larger rather than smaller DF; that having strong reminders on L2 can reduce DF; that having no L2 eliminates DF; that categorized items can be susceptible to DF or not depending on whether the test cue involves context cue or a category cue. In our view, the account is strongly supported by the existing data, and any future theory should strive to not only explain the effects that the context-change account predicts but also make novel predictions that cannot be handled by the current theory.

Shortly after we proposed the context-change account of DF, Anderson (2005) proposed that the context-change mechanism is compatible with an inhibitory interpretation of DF if inhibition is assumed to be the mechanism by which context change takes place. According to Anderson's interpretation, inhibition is a flexible mechanism that can be targeted at different levels of representation, including at a context representation. We are reluctant to embrace this notion because we do not see it as advancing our theoretical understanding of DF. It redescribes context change with new terms (i.e. inhibition of context) without specifying a new theory of inhibition of context. There is currently no evidence to suggest that context is inhibited as opposed to being changed across the lists in DF. Impaired memory of L1 items does not constitute any evidence of inhibition of context or indeed any inhibition at all (MacLeod et al., 2003). Equating mental context change with inhibition of context implies that any time we change what we are thinking about, we engage in inhibition. By this view, everything becomes inhibition, and thus impossible to know what does *not* involve inhibition. Inhibition of context also complicates the explanation of

DF because it implies two mechanisms, not one. If we assume that context serves as retrieval cue, then when one changes from context A to context B, context match becomes poorer for the previous context A. Adding inhibition would be imposing a second mechanism that assumes that changing from context A to context B requires inhibition of context A. To the best of our knowledge, no one has offered an inhibitory theory of environmental context-dependent memory (e.g. Godden & Baddeley, 1975). The impaired memory arising from the context change is instead explained in terms of the mismatch of cues being used to search memory and the contents of memory, which leads to reduced accessibility.

6. AREAS OF DISAGREEMENT ACROSS STUDIES

While reviewing the empirical findings, two broad areas of inconsistencies emerged as in need of further investigation and more careful analyses. These include L1 serial position findings, and the source misattribution and intrusion error findings.

6.1. Serial Position Effects

We identified five published DF studies that examined serial position functions in DF (Geiselman et al., 1983; Lehman & Malmberg, 2009; Pastötter & Bäuml, 2010; Pastötter et al., 2012; Sahakyan & Foster, 2009). Some of these studies used serial positions as a tool for testing various DF theories, whereas other studies plotted those functions without formally analyzing them. Although the studies agree that the effect of the forget cue on L2 is mostly evident at the start–middle of the L2 serial position curve, their results regarding L1 functions differed to some extent. For example, studies that analyzed serial position functions did not obtain greater DF costs from the primacy region of L1 curve (e.g. Pastötter & Bäuml, 2010; Pastötter et al., 2012; Sahakyan & Foster, 2009), whereas studies that reported serial position curves without formal analyses (or deviation measures) obtained greater DF from the primacy regions of L1 curves (e.g. Geiselman et al., 1983; Lehman & Malmberg, 2009). Lehman and Malmberg (2009) argued that context is more prominent at the start of the list, which could explain greater forgetting from the primacy region. A study reported by Sheard and MacLeod (2005) also obtained greater DF costs from L1 primacy region, although their study failed to obtain overall DF impairment. Thus, some studies obtain greater DF impairment in the primacy region

whereas others do not. Note that Sahakyan and Foster (2009) obtained equivalent DF impairment across L1 serial position curves of performed actions phrases which typically do not produce primacy effects, suggesting that the magnitude of DF impairment may not be linked to the primacy effects. Considering these controversies and the methodological differences that could have given rise to these findings, more careful attention is needed to serial position effects in DF research.

6.2. List Differentiation in DF: Source vs Intrusion Errors

Some DF studies found that the forget cue impaired source discrimination judgments in recognition (e.g. Bjork & Bjork, 2003; Gottlob & Golding, 2007; Hanczakowski et al., 2012), whereas we explained these findings in terms of a bias to attribute L1 items to L2 during the recognition test (e.g. Sahakyan & Delaney, 2005). Regardless of whether this is a bias effect or a true source confusion effect, these findings overall suggest that DF impairs list differentiation. Somewhat different conclusions are reached when DF is examined via intrusion errors during free recall or list categorization errors during free recall (e.g. Geiselman et al., 1983; Golding & Gottlob, 2005; Lehman & Malmberg, 2009; Sahakyan & Delaney, 2010; Spillers & Unsworth, 2011). Some of these studies have found enhanced list differentiation as evidenced by reduced intrusion errors, whereas others have found the opposite. Although intrusion errors are notoriously difficult to study, they could nevertheless be informative because they could indicate the type of retrieval cues that participants rely on during free recall. For example, if people make many intrusions from a wrong list, this might suggest problems with the sampling stage, indication that they use "wrong retrieval cues" that better match one list than the other list (e.g. Spillers & Unsworth, 2011).

As list differentiation in DF is studied further, it is important to keep in mind that source errors and intrusion errors may not necessarily lead to the same conclusions because they represent opposite sides of the same coin. Source errors refer to the ability to retrieve the context given the item, whereas intrusion errors are more diagnostic of the cuing property of context, referring to the ability to retrieve the item given the context cue. If item-to-context and context-to-item associations are not necessarily equivalent in strength, or if ability to retrieve the context given the item as opposed to retrieving the item given the context were not identical, then the findings from source errors and intrusion errors may not necessarily produce the same empirical outcomes.

7. STRATEGY CHANGE EXPLAINS DF BENEFITS

Sahakyan and Delaney (2003) were the first to propose that the costs and benefits of DF might be dissociable, and that they might be explained by different mechanisms. A critical piece of evidence for their proposal was that it is possible to dissociate the costs and benefits (e.g. Benjamin, 2006; Conway et al., 2000; Joslyn & Oakes, 2005; Pastötter et al., 2012; Sahakyan & Delaney, 2005, 2010; Shapiro et al., 2006; Spillers & Unsworth, 2011; Zellner & Bäuml, 2006). How could that be true if a single process accounted for both results? While not all researchers agree what the two processes are, most now agree that separate mechanisms are needed to explain costs and benefits.

7.1. The Advent of the Dual-Process Account

The Sahakyan and Delaney (2003) dual-process account explains the benefits by the strategy-change mechanism. According to this explanation, the benefits of DF emerge because the forget cue encourages participants in the forget group to adopt a more elaborate encoding strategy for L2 items compared to those in the remember group. Thus, L2 benefits reflect a strategic encoding enhancement in the forget group.

We initially proposed this explanation because we evaluated participants' verbal reports of study strategies from Sahakyan and Kelley's (2002) study, which revealed that many participants adopted a better study strategy on L2 than L1, and the forget group was almost twice as likely as the remember group to switch encoding strategies (38% vs 22% in the Sahakyan and Kelley study). Therefore, in Sahakyan and Delaney (2003), we first reanalyzed Sahakyan and Kelley's (2002) data by statistically controlling the strategies, and we found that the effect of cue (F vs R) was much smaller compared to the effect of strategy switch, which accounted for the majority of variance in L2 recall. We also conducted two experiments that experimentally controlled the encoding strategy, with the same outcome. In Experiment 1, participants studied both lists using the same encoding strategy, and we did not obtain the benefits because everybody suffered massive proactive interference. In Experiment 2, everybody was told to switch study strategy between L1 and L2 (including the remember group), and everybody benefited from such switch and escaped proactive interference. In both experiments, we always obtained DF costs.

This is important to emphasize because we often hear from our colleagues that a switch of study strategies between the two lists may be part of mental context change and could explain the DF impairment. Although strategy switch could contribute to mental context change, it cannot explain why the DF impairment was still obtained in Experiment 2, when everybody switched strategies, including the remember group. According to this reasoning, there should be no DF costs if *everybody* changes mental context due to strategy shift. However, the costs are present regardless of whether people retain the same strategy on both lists, or change the strategy across the lists, which led us to propose the dual-process account. Studies investigating neural bases of DF also provide support for the two-factor account, and suggest different neural origins for the costs and the benefits of DF and mental context change (Bäuml, Hanslmayr, Pastotter, & Klimesch, 2008; Hanslmayr et al., 2012; Pastötter, Bäuml, & Hanslmayr, 2008).

In general, whenever we controlled encoding strategy in subsequent studies, we never obtained the benefits (e.g. Sahakyan & Delaney, 2005, 2010; Sahakyan, Delaney, & Waldum, 2008). Along the same lines, when participants performed simple action phrases during encoding (e.g. *break a toothpick*), we did not obtain the benefits, but the same action phrases encoded verbally without controlling for study strategy produced the benefits (e.g. Sahakyan & Foster, 2009). Although we explain L2 benefits in terms of encoding strategy switch, a related and similar account was proposed by Pastötter and Bäuml (2010), who argued that the benefits arise because the forget cue resets the encoding processes at the start of L2 learning. This proposal is analogous to the earlier selective rehearsal account, with an exception that it reserves selective rehearsal mechanism strictly for explaining L2 benefits (as opposed to the costs and the benefits as was initially assumed by the selective rehearsal). The primary reason for proposing a reset of encoding (as opposed to a strategy change) is that L2 serial position curves typically show much larger DF benefits in the primacy regions of the curve than elsewhere in the list (e.g. Geiselman et al., 1983; Lehman & Malmberg, 2009; Pastötter & Bäuml, 2010; Sahakyan & Foster, 2009). Pastötter and Bäuml (2010) argued that a switch to a better encoding strategy should be evident throughout the entire list, not just at the beginning of L2. This interpretation of strategy change is reasonable, but might be a bit extreme. Although many participants spontaneously report doing something different to encode L2 (or in addition to what they were doing to encode L1), it is quite possible that they do not persist with that strategy throughout the entire L2 learning. For instance, often participants report

starting out with some strategy on L2, but further down the list they report being overwhelmed and being unable to keep it up. This could explain why the benefits of DF are more prominent in the early sections of L2 serial position curves than in later sections.

7.2. A Large-Scale Replication Study

Our dataset of $N = 290$ participants (Figure 4.5) provides reports of strategy in response to the forget cue, and also contains reports of participants' encoding strategy for both lists. We therefore reevaluated the strategy-change account in our dataset with a substantially larger sample. As Figure 4.5 demonstrates, despite using a variety of forgetting strategies with some being more effective than others in leading to DF impairment, nearly all strategies were associated with almost invariant magnitude of DF benefits. Interestingly, even the group that reported doing nothing to forget also showed the benefits. The "doing-nothing" group essentially shows a dissociation, whereby they show no DF impairment compared to the remember group, but somehow they show the benefits on L2 despite no forgetting of L1. If the benefits are driven by strategy change, then it is quite possible that these participants changed study strategy between the lists (perhaps in hopes that doing so might help them forget L1). We coded participants' encoding strategies for both lists, and created a variable that captured whether participants employed a more efficient strategy on L2 compared to L1. For example, any mention of doing something in addition to what they were doing during L1 encoding was coded as a *strategy switch* (e.g. rehearsing on L1, and rehearsing plus also grouping words together on L2).

Confirming previous work from our lab, we found that 28% of the forget group switched to a better study strategy on L2 compared to 10% of the remember group, X^2 (1, $N = 290$) = 14.43, $p < 0.001$. Although the rates of strategy switch were much lower in the University of North Carolina at Greensboro sample than in the previous samples of Florida State University and University of Florida participants, the basic pattern was similar—the forget group was more likely than the remember group to switch to better encoding on L2. Confirming the results of Foster and Sahakyan (2011), we also found that within the forget group, there was no relation between the likelihood of switching to a better study strategy on L2 and the likelihood of engaging in the forgetting strategy such as doing something vs nothing to forget L1 ($\varphi = -0.03$, $p = 0.73$). We therefore collapsed across that factor in the forget group, and analyzed L2 recall using cue (forget vs remember) and strategy switch (yes vs no) as predictors. Strategy switch was the only

significant predictor ($p = 0.007$). People who switched study strategy recalled significantly more from L2 than those who did not change study strategy. The main effect of cue was not quite significant ($p = 0.06$) despite a large sample size. These findings largely replicate our previous research conducted at another institution (Sahakyan & Delaney, 2003), and they suggest that strategy switch provides a better explanation for the DF benefits than does the cue. However, it is important to note that we have always obtained cue effects both in prior research and in the current dataset, just not at the conventional levels of significance. This suggests that there is probably some escape from proactive interference that arises from context change. However, such effects are difficult to detect in overall recall rates (at least in a two-list design), and they are much smaller than the encoding effects arising from the strategy change.

7.3. Test Order and the Benefits of DF

One factor that could contribute to failures to detect DF benefits may be linked to test order. A recent meta-analysis on 20 studies that reported data on output order concluded that failure to detect DF benefits in many studies could be driven by the fact that L2 was tested after L1 in many prior studies (Pastötter et al., 2012). In addition to meta-analyses, Pastötter et al. (2012) conducted two studies, where recall order was manipulated, and they showed that the benefits of DF are much larger when L2 is tested first in the recall sequence than when it is tested after L1.

Although several studies from our lab controlled the encoding strategy and tested L2 after L1 (Sahakyan & Delaney, 2003; Sahakyan, Delaney, & Waldum, 2008), it is unlikely that our chosen test order could account for the absence of benefits in those studies. Sahakyan and Delaney (2010) counterbalanced test order while controlling study strategy for some participants and not for others. Although their published data were collapsed across test order, Figure 4.7 replots the data broken out by test order, and shows that the presence or absence of DF benefits was driven entirely by whether the study strategy was controlled or not; test order had no impact on the findings.

7.4. Summary

Most researchers now agree that DF benefits on L2 arise from a different mechanism than the costs. We argued that the underlying factor is the strategy change triggered by the forget cue. As with the context-change mechanism, the idea for this came from verbal reports given by participants

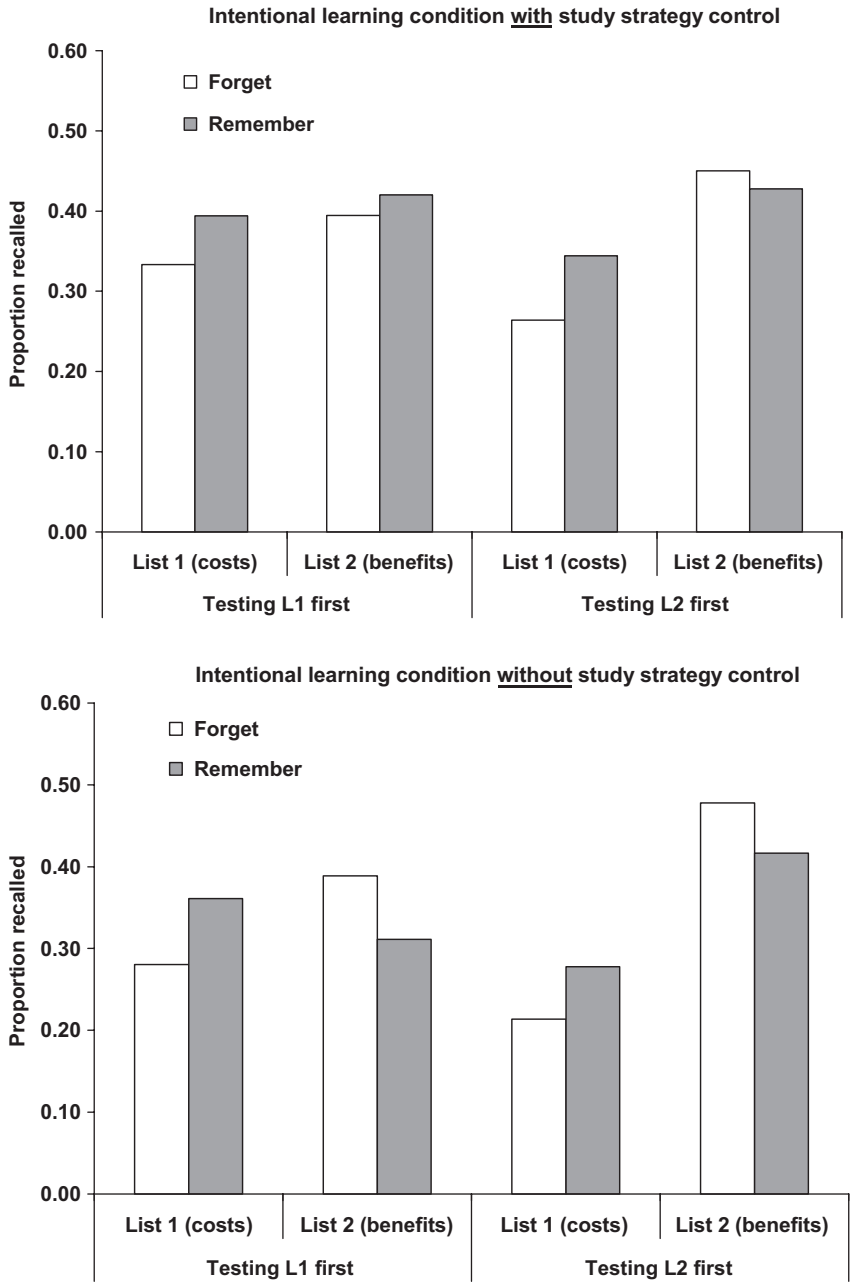


Figure 4.7 Directed forgetting costs and benefits as a function of test order, controlled study strategy (a), and uncontrolled study strategy (b). *(New figures based on data reported in Sahakyan and Delaney (2010)).*

in our experiments, and was supported by successful predictions in other experiments. For example, controlling participants' encoding strategy via instructions reduced or eliminated the benefits, as does the use of incidental learning procedures. This review also reports on a large-scale replication of the original work establishing the benefits were due to strategy changes: even though strategy switches are by no means common, they produce such large benefits to memory that they show up at the group level. We think that reducing build-up of proactive interference due to contextual isolation (the mechanism originally proposed by Sahakyan & Kelley, 2002 to explain the benefits) may also be important in some cases, but we have at best limited empirical evidence for that mechanism.

8. IMPLICATIONS FOR CLINICAL POPULATIONS

People with clinical disorders sometimes behave atypically on cognitive tasks. Demonstrations that clinical disorders alter behavior in cognitive tasks are useful because they can improve diagnosis (if the difference is sufficiently robust), differentiate subtypes of disorders, and clarify what functions are impaired in clinical populations. Cognitive difficulties can suggest treatments that may remediate symptoms, or predict the kinds of everyday problems that people may have. Of course, it is insufficient to know that a particular cognitive task is affected by a clinical diagnosis; we must have a theory as to the reason why the disorder affects behavior in the cognitive task. Hence, our theories about why cognitive effects occur are important in interpreting why clinical populations perform the task differently.

List-method DF is one cognitive task that people with clinical diagnoses sometimes perform differently from other people. Until the past decade, list-method DF deficits have generally been interpreted in terms of deficits in inhibition (for earlier reviews of the DF literature in clinical populations, see Cloitre, 1998; McNally, 2005). Significant DF is taken as evidence that inhibition is intact, while difficulties in DF are taken as evidence for impairments in inhibitory control. However, our review provides an alternative interpretation. Some clinical populations may have problems with context processing or context change. Although not everyone is convinced that the case for inhibition is well-established even in the other tasks (e.g. MacLeod et al., 2003; Raaijmakers & Jakab, 2013), many researchers assume that impairment on tasks like the stop-signal paradigm and RIF reflect inhibitory control, and that deficits on these tasks should be diagnostic of inhibitory deficits. If so, then if DF is also inhibitory in nature, we would expect

that the same inhibitory deficits should occur there as well. We will review evidence that is consistent with a context-based interpretation of clinical difficulties in DF, and that other putatively "inhibitory" tasks do not show the same pattern of deficits as DF, consistent with our theory. The notion that DF impairments are linked to context processing (and not inhibition) provides new ways of thinking about how DF elucidates clinical disorders.

Having shown that context may be important to clinical disorders, we will next argue that many of the studies that have been conducted in the clinical literature are lacking in rigor, and that improving the methodological rigor of future studies will be necessary to draw firmer conclusions from DF research. We will provide specific advice for clinical researchers seeking to use DF as a tool for understanding context processing in clinical disorders. Without singling out any earlier study, we will suggest that many earlier studies could be repeated with more recently developed methodologies from the cognitive literature in order to yield more detailed conclusions about disorders.

8.1. Reinterpreting Clinical List-Method DF Results as Reflecting Context

Given that context is so important in memory theory, it is perhaps surprising that it is so underrepresented as an explanatory mechanism in clinical problems. Our ability to link events to their context plays an important role in understanding aging deficits, for example, where research suggests that binding items to their context may be extremely difficult for older adults. However, the same logic has rarely been applied to clinical disorders, where one might expect quite a lot of traction to be gained from linking disorders to mental context processing. In the next few sections, we will outline some disorders where we think that the context view may make sense of dissociations with other tasks that are assumed to reflect inhibition, and perhaps provide new ways of understanding how context is implicated in clinical problems.

8.1.1. Attention-Deficit Hyperactivity Disorder

An obvious place to begin the review is with attention-deficit hyperactivity disorder (ADHD), because it is widely believed that ADHD is in part a dysexecutive disorder (e.g. Barkley, 1997). A number of studies have argued that there are inhibitory deficits in ADHD patients, because they perform poorly in the think/no-think paradigm, stop-signal paradigm, and elsewhere (for a brief review, see Depue, Burgess, Wilcutt, Ruzic, & Banich, 2010).

Given that pattern of deficits, inhibitory theorists would likely expect that ADHD should impair list-method DF as well, but that does not appear to be true. Gaultney, Kipp, Weinstein, and McNeill (1999) compared children aged 8–15 who had ADHD and who did not. Using a standard list-method paradigm, they obtained both the costs and benefits of DF with both groups. Thus, ADHD children apparently show normal DF. White and Marks (2004) compared undergraduates with ADHD to non-ADHD undergraduates. Participants studied two lists of items using a procedure similar to Geiselman et al. (1983): Half of the items were designated as "learn" items that participants were supposed to memorize for a later test and the other half were "judge" items that they had to rate for pleasantness without intending to memorize them. The non-ADHD students replicated the results of DF costs reported by Geiselman et al. (1983)—namely, both learned and judged words suffered from DF. The results of the ADHD group, on the other hand, can be interpreted differently depending on how they are analyzed. The authors concluded that "individuals in the ADHD condition did not show reliable evidence of intentional forgetting." However, a reexamination of their data suggests a different conclusion: normal costs (11%) were present for the learn items among ADHD students, although for judge items the costs were absent (−6%). The pattern with judge words may be a false negative and may be worth following up in future work given that several studies have obtained DF in incidental encoding (e.g. Sahakyan & Delaney, 2005, 2010). The results with the learn words are largely consistent with the Gaultney et al. (1999) finding of intact DF in ADHD. In sum, while ADHD patients show deficits in other tasks that have been linked to inhibition, DF appears to be largely intact in this population.

8.1.2. Schizophrenia

Impairments in DF sometimes co-occur with deficits in other inhibitory tasks. For example, schizophrenia is thought to be associated with deficits in memory and inhibitory control (for reviews, see Hoff & Krennan, 2002; Perlstein, Carter, Barch, & Baird, 1998), and therefore researchers who believe that DF reflects inhibition should predict deficits in list-method DF in schizophrenics. Several studies have tested this possibility. In our judgment, these are among the best-conducted list-method DF studies in the clinical literature, and should be viewed as models for others.

Racsmány et al. (2008) examined medicated schizophrenics. While the control group showed robust costs and benefits, the schizophrenic patients showed neither costs nor benefits. Their overall memory was also lower.

Importantly, in a second experiment, they obtained robust RIF effects, suggesting dissociation between DF and RIF.

Soriano et al. (2009) examined medicated patients diagnosed with schizophrenia, schizoaffective disorder, and schizophreniform disorder. They were split according to whether they reported hallucinations on the Positive and Negative Syndrome Scale (PANSS). Although the omnibus statistics did not always work out due to low power, the planned comparisons suggested that the costs and benefits were present for both controls and nonhallucinators, whereas the patients who were hallucinating showed neither. Both groups of patients had impaired memory compared to controls. Thus, Soriano et al.'s (2009) findings suggest that the presence or absence of hallucinations is a critical variable in terms of whether schizophrenics show DF or not.

8.1.3. Generalized Anxiety Disorder

Few studies have found any executive functioning difficulties in generalized anxiety disorder (GAD). A recent study, for example, looked at postpartum women's GAD symptomology and found no correlation with a range of executive functions other than working memory capacity (Vadnais et al., 2012). Thus, there is not much evidence to suggest that GAD is associated with impaired inhibition. Nevertheless, a person with high anxiety may find it aversive to "think of something else" and may have difficulty encoding or shifting contexts. In other words, anxiety disorders may provide an example of difficulties shifting context despite intact inhibitory abilities. A study by Albu (2008) used the full design of the list method to examine GAD patients and patients without anxiety difficulties. Their L1 consisted of anxiety-related words and L2 consisted of neutral words. The major problem with the study is that the authors did not counterbalance which items were assigned to the forget and remember conditions. Therefore, any differences between the groups must be interpreted in light of possible differences due to item effects (we will discuss item effects more fully in Section 8.3). The non-GAD group showed robust costs (and, unusually, antibenefits). In contrast, the GAD group showed neither costs nor benefits. These findings suggest that the GAD patients might have a deficit in context change despite intact performance on other inhibitory tasks.

8.1.4. Depression

An even more interesting pattern would occur if a clinical problem improved DF. A study by Power, Dalgeish, Claudio, Tata, and Kentish (2000) examined subclinically dysphoric people, comparing people with low-Beck Depression

Inventory (BDI) scores to high-BDI scorers (dysphorics). In all their studies, they used mixed lists of positive and negative adjectives. Their Experiments 1 and 2 are difficult to interpret, as they did not include a remember control group or control for test order of the lists, but the "R–F" measure was positive and significant in both groups. However, their Experiment 3 used clinically depressed, high anxious, and healthy controls, and included a remember control group for each sample. The number of participants in each cell was extremely small (in some cases as few as five people), making interpretation difficult. Although most of their effects were in the correct direction for obtaining the costs and benefits, nothing was significant except an anti-forgetting effect for negative words among the depressed participants. This study suggests that negative words may be difficult to forget for depressed people, perhaps because they trigger rumination. Consistent with a bias to ruminate on negative words, Experiment 2 found that dysphorics recalled approximately equal numbers of positive and negative words, whereas the low-BDI participants showed a positivity bias and favored positive words. Indeed, negative words are difficult for depressives to update in working memory, which has led some authors to call them "sticky"—difficult to turn attention away from (Joormann, Levens, & Gotlib, 2011).

The Power et al. (2000) study therefore does not provide a clean test of whether depressives show greater DF than the controls. A recent poster by Lehman and Malmberg (2011b) provided a progress report on their large-scale study examining depressed peoples' DF, and suggested that depressives may be better able to change context. They used the full list-method DF design with two lists of neutral words, and found that clinically depressed people showed larger DF impairment than the control participants. Their argument was that depressed patients may have a built-in mechanism for producing more context change: depressive rumination. Although Joormann and Tran (2009) studied people high on rumination in a list-method study, they used a median split on a rumination scale, and did not find higher BDI scores among their ruminators. Hence, their failure to obtain enhanced DF among ruminators (using a no-control-group "R–F" statistic) may not be surprising.

8.2. Future Directions for Clinical Research Using the List Method

In sum, ADHD produces impaired executive functioning and poor performance on tasks thought to reflect inhibitory control, but nonetheless show intact DF. Contrastingly, GAD patients show impaired list-method

DF without clear executive or inhibitory deficits, suggesting a double dissociation between inhibitory/executive measures and DF. Schizophrenia produces impairments on both putatively inhibitory tasks and on DF (but see Racsmány et al., 2008 for dissociations between DF and RIF). Finally, depressed patients may show larger than normal DF, despite some evidence that they sometimes have impaired working memory. Taken together, the clinical studies are consistent with our argument that the basis of list-method DF is not inhibition but contextual change, and that disorders that affect the likelihood of effectively thinking of something else affect list-method DF.

These studies all focused on whether or not DF occurs in a particular population, but the obvious next step is to ask *why* DF is impaired. In this chapter, we have argued that DF is caused by context change, but also that context change is under volitional control. Impaired DF might reflect inability to deploy an effective strategy for forgetting. Perseverative difficulties could make it difficult to stop ongoing processing and deploy forgetting strategies, for example. It may be useful to include measures of what people do when attempting to forget (Foster & Sahakyan, 2011), as some people may fail to initiate context change spontaneously, but could do so normally with instruction (e.g. in the aging literature; Sahakyan, Delaney, & Goodmon, 2008). Alternatively, impaired DF could reflect difficulties in context processing, as was suggested for people with low working memory capacity (Section 5.12). Superior DF, likewise, could make sense if people are more able than usual to deploy an effective forgetting strategy—a case we will suggest might apply to depression (Lehman & Malmberg, 2011b). One possible method of getting at whether generalized context-change problems are typical of a disorder or whether it is a problem of initiating context change on one's own may be to use diversionary thought tasks like imagining your childhood home (Section 5.1).

That depressives seem more effective than others at forgetting neutral words (Lehman & Malmberg, 2011b) but not mixed lists of negative and positive words (Power et al., 2000) further suggests that forgetting may be dependent on the type of material being forgotten (McNally, Metzger, Lasko, Clancy, & Pitman, 1998). Different types of material could trigger thoughts that produce forgetting or could fail to do so. Studies contrasting forgetting of neutral and disorder-related words may be informative in other disorders as well, especially if they include appropriate remember control groups to facilitate differentiating between generalized memory biases and biases specific to forgetting.

8.3. Improving Research Design for Clinical Studies of List-Method DF

One problem with interpreting studies on list-method DF is that many of them are based on older research designs that provide data that are often misinterpreted. Hence, our next task is to suggest list-method designs that provide the best chance of accurately answering research questions about clinical populations' DF. Whether or not one accepts that DF is context-based, it is a good idea to use a design that can discriminate different theoretical possibilities. We have outlined our specific recommendations as Table 4.1.

8.3.1. Inclusion of a Separate Remember Condition

As noted in Section 2.1, a common design issue in DF studies is the absence of a remember control group. In such studies, the costs and benefits of DF have been assessed using the "R–F" measure, which can be a problematic assessment for DF effects. This measurement issue is particularly problematic when assessing DF in clinical populations. Specifically, if a clinical population is impaired on the costs but shows normal benefits, this would be impossible to distinguish using the "R–F" measure. According to the theory outlined in this chapter, smaller costs but normal benefits suggest difficulties with the use of context. Intact costs but impaired benefits suggest strategy-change issues. Thus, including a remember control group allows for a systematic assessment of the costs and benefits of DF.

It is understandable that including remember control groups when patient populations are scarce may be difficult. For this reason, many designs rely on comparing "R–F" in healthy vs patient populations. We think these studies are informative in that they indicate differences in DF between patients and healthy controls. However, one has to be very careful about interpreting them in terms of impaired inhibition (or even impaired context change), because both costs and benefits are rolled into the "R–F" differences. A possible solution is to adopt the four-list procedure (Section 2.2). Especially if

Table 4.1 Suggestions for Improved List-Method Directed Forgetting Designs

1. *Use a remember control group*—or at least the within-subjects design that includes both RR and FR lists
2. *Test list 1 and list 2 separately*—or at least calculate output percentiles
3. *Consider output order*—if a study mixes different types of items (e.g. trauma and neutral words), examine output order effects
4. *Avoid or acknowledge scaling effects*—baseline memory may differ across groups
5. *Avoid floor effects*—use appropriate numbers of items in each category

one is interested in the costs, there is little evidence that the within-subjects design distorts the effect, as reviewed earlier. It is important to counterbalance the assignment of words to lists as well; at least one study obtained odd results because they failed to ensure that the words assigned to the forget condition were assigned to the remember condition for other participants, and hence item effects emerged (Albu, 2008). The four-list procedure allows the researcher to compute both costs and benefits using a powerful within-subjects comparison, which is clearly advantageous. If there are concerns about order effects, they can also be addressed using this design. If there are serious order effects, one can also focus on only the first lists studied.

8.3.2. Controlling Test Order and Measuring Output Order

Another design issue is that most clinical studies have not controlled the test order of L1 and L2. In Section 2.1, we discussed the problem of output interference in the benefits. While research suggests that in healthy controls the costs are relatively unaffected by output interference effects (Pastötter et al. 2012), this may not necessarily be the case for patients. There could be systematic differences between patients and healthy controls as to whether they start recall from L1 or L2 for a variety of reasons.

Controlling test order may be particularly important when mixing several different types of items on the same list, as when one is interested in whether trauma or neutral words are harder to forget. Comparing the relative recall rate of different item types will be complicated by output interference, especially if different groups of participants output them in different orders. For example, if we mix trauma and neutral words on each list, then patients might begin recall from all the trauma words, whereas healthy controls might have no such bias. Such differences in output order would produce varying levels of output interference on different types of items across the groups, resulting in apparent differences in recall rates. Hence, we suggest that output order of the lists be controlled whenever possible.

At the very least, one should examine output order as a possible explanation for any observed results. One way to examine output order is to calculate the average output positions. The easiest way to do this is to create output percentiles, which examine the average position within the output of the items of a given type (e.g. Sahakyan, Delaney, & Waldum, 2008). For example, imagine we mix trauma and neutral words, and the output items are in the order *trauma*, *trauma*, *trauma*, *neutral*, *trauma*, *neutral*, *neutral*. The output percentile for trauma words would be $(1 + 2 + 3 + 5)/(7 \times 4) = 0.39$. For neutral words, it would be $(4 + 6 + 7)/(7 \times 3) = 0.81$. Thus, output for

neutral words is occurring later than for trauma words for this participant. Further, significant differences in output percentiles between groups would be evidence for differences in output order of different types of items.

8.3.3. Scaling Effects

Another common problem when comparing groups of people is a scaling effect (also known as baseline effects). Some groups of people show general memory impairments, and hence have lower baseline memory. For example, this problem is well-known in the literature on aging—older adults have impaired recollection, and so they perform worse than younger adults on free recall tests. In this case, the costs might be numerically smaller in older adults than in younger adults due to reduced overall memory. If the costs are a percentage of the original memory, then Analysis of Variance (ANOVA) might incorrectly suggest that the costs are smaller for older adults than for younger adults. One solution is either to enhance the memory of the weaker group, such as by giving longer presentation times, or to provide more elaborate processing instructions before the forget cue, especially because neither study time nor depth of processing interacts with DF (Sahakyan, Delaney, & Goodmon, 2008; Sahakyan, Delaney, & Waldum, 2008). If the baseline recall rates are closer in the remember control group, then the differences in the costs may be more easily interpreted. Of course, if the costs are completely absent, it is more difficult to argue for a scaling effect (unless there is a floor or ceiling effect in the data).

8.3.4. Floor Effects

Many clinical studies also show floor effects in recall. The standard diagnosis for floor effects is that the confidence interval around the mean captures 0%. Another good indicator of floor effects is if a significant percentage of participants recall zero of a given type of item. When the range of responses is very small (e.g. 0, 1, 2 or 3), then a lot of participants would be needed to detect a meaningful difference in the means, as the measurement error is likely to exceed the size of the effect. Whenever there is an apparent floor effect, one should be prudent when interpreting mean differences, as the estimate of the variance is likely to be deflated, producing false positives.

It is difficult to avoid floor effects in recall, but a good start is to avoid very long lists; not more than 20 items are likely to yield reasonable recall rates. Likewise, if the items are fractionated into different types of items, such as traumatic and neutral words, then one should try to keep these sublists from getting too short. We suggest a minimum of eight items per

category. Of course, this restricts the kinds of tests that can be run with a single design, but it is better to get meaningful answers about one or two different types of item than to get meaningless floor effects with a large number of different types of item.

8.4. Summary and Linking Clinical to Nonclinical Research on DF

While we have argued that the current theory can explain many of the differences in DF among clinical populations, in some ways, studies with special populations provide the best evidence for individual differences in DF. Even if contextual change can account for DF impairment in everyday people, it may be less successful among some clinical groups for a variety of reasons. The absence of DF in some clinical populations may provide clues as to why some seemingly healthy participants show little or no DF. Likewise, if some clinical populations show more effective DF than healthy participants, it may give us some insight into why some healthy participants show larger-than-usual DF effects. For example, it may well be that people who ruminate more or people who are fantasy-prone are better forgetters even in the absence of any clinical symptoms.

Clinical research could also shed light on the controversies surrounding what types of items are more or less forgettable. Some populations have difficulties related to specific types of words, which could enable us to better understand how differential processing interacts with the forget cue. Thus, more insights about the nature of DF can be gained through well-designed studies with clinical populations. Although basic research provides a foundation for interpreting population differences, research with special populations is equally necessary for the future development and refinement of DF theories. We are hopeful that by incorporating better, modern research designs that were developed in the cognitive literature to investigate DF that deeper conclusions can be drawn about context processing in various clinical populations. Repeating some of the earlier studies with these modern designs should be helpful in discriminating between different theoretical possibilities, and shed light on the nature of several important clinical disorders.

9. CONCLUDING THOUGHTS

Obtaining DF requires engaging in controlled strategies, and the decision to engage in those strategies may be mediated by beliefs about one's

memory abilities or whether it is possible to intentionally forget. These findings highlight the importance of attending to participants' behavior and beliefs in understanding why memory phenomena occur. The costs and the benefits of DF can be observed or not depending on what participants choose to do. A closer attention to participant strategies is warranted in DF.

The context account as laid out here explains a substantial amount of DF data, while also serving as foundation for testable predictions regarding the impact of different manipulations on DF. The context-change mechanism is not uniquely suited to explain DF. Instead, the account brings DF into the family of effects that are explained by existing memory models, allowing us to make new testable predictions. These models and the rich empirical foundation of the environmental context literature serve as a foundation for new quantitative and qualitative predictions.

The traditional way of interpreting DF deficits in many special populations is that they have deficits in inhibitory control. The context account of DF, however, suggests new interpretations, opening up new directions for investigating context processing and/or contextual binding deficits in those populations.

Finally, a commonsense notion is that disruptions are inherently bad for memory. Work on DF and mental context-change paradigm suggests that how disruptive interruptions depend on the contents of the interruption and the extent to which they change mental context. Some interruptions are easier to recover from than others. A major challenge that lies ahead is to more fully understand why some tasks but not others create mental context change—in other words, to create a theory of what mental context consists of. We look forward to seeing how these theories develop in the next few decades. Understanding how mental context is represented and updated may turn out to be central to memory functioning in many more ways than we yet realize.

ACKNOWLEDGMENTS

The authors are grateful to Colleen Kelley, Colin MacLeod, and Tanya Jonker for their feedback and constructive comments on the earlier draft of this chapter.

REFERENCES

Albu, M. (2008). Automatic and intentional inhibition in patients with generalized anxiety disorder. *Cognition, Brain, & Behavior*, *12*, 233–249.

Anderson, M. C. (2003). Rethinking interference theory: executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445.

Anderson, M. C. (2005). The role of inhibitory control in forgetting unwanted memories: a consideration of three methods. In C. MacLeod & B. Uttl (Eds.), *Dynamic cognitive processes* (pp. 159–190). Tokyo: Springer-Verlag.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123.

Aslan, A., & Bäuml, K. T. (2008). Memorial consequences of imagination in children and adults. *Psychonomic Bulletin & Review*, *15*(4), 833–837.

Aslan, A., & Bäuml, K. T. (2011). Individual differences in working memory capacity predict retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 264–269.

Aslan, A., Zellner, M., & Bäuml, K. T. (2010). Working memory capacity predicts listwise directed forgetting in adults and children. *Memory*, *18*(4), 442–450.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*, 65–94.

Barnier, A. J., Conway, M. A., Mayoh, L., Speyer, J., Avizmil, O., & Harris, C. B. (2007). Directed forgetting of recently recalled autobiographical memories. *Journal of Experimental Psychology: General*, *136*(2), 301–322.

Basden, B. H., & Basden, D. R. (1996). Directed forgetting: further comparisons of the item and list methods. *Memory*, *4*(6), 633–653.

Basden, B. H., & Basden, D. R. (1998). Directed forgetting: a contrast of methods and interpretations. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 139–172). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Basden, B. H., Basden, D. R., & Gargano, G. J. (1993). Directed forgetting in implicit and explicit memory tests: a comparison of methods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 603–616.

Basden, B. H., Basden, D. R., & Morales, E. (2003). The role of retrieval practice in directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 389–397.

Basden, B. H., Basden, D. R., & Wright, M. J. (2003). Part-list reexposure and release of retrieval inhibition. *Consciousness and Cognition: An International Journal*, *12*(3), 354–375.

Bäuml, K. H. (2008). Inhibitory processes. In H. L. RoedigerIII (Ed.), *Cognitive psychology of memory* (pp. 195–220). Oxford: Elsevier.

Bäuml, K.-H., Hanslmayr, S., Pastotter, B., & Klimesch, W. (2008). Oscillatory correlates of intentional updating in episodic memory. *NeuroImage*, *41*, 596–604.

Bäuml, K. T., & Samenieh, A. (2010). The two faces of memory retrieval. *Psychological Science*, *21*(6), 793–795.

Bäuml, K. T., & Samenieh, A. (2012a). Influences of part-list cuing on different forms of episodic forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 366–375.

Bäuml, K. T., & Samenieh, A. (2012b). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 488–494.

Benjamin, A. S. (2006). The effects of list-method directed forgetting on recognition memory. *Psychonomic Bulletin & Review*, *13*, 831–836.

Benjamin, A. S., & Bjork, R. A. (2000). On the relationship between recognition speed and accuracy for words rehearsed via rote versus elaborative rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 638–648.

Bjork, R. A. (1970). Positive forgetting: the noninterference of items intentionally forgotten. *Journal of Verbal Learning and Verbal Behavior*, *9*, 255–268.

Bjork, R. A. (1972). Theoretical implications of directed forgetting. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 217–325). Washington, D. C.: V. H. Winston & Sons.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. Roediger & F. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309–330). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Bjork, E. L., & Bjork, R. A. (1996). Continuing influences of to-be-forgotten information. *Consciousness and Cognition*, *5*, 176–196.

Bjork, E. L., & Bjork, R. A. (2003). Intentional forgetting can increase, not decrease, residual influences of to-be-forgotten information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 524–531.

Bjork, E., Bjork, R. A., & Anderson, M. C. (1998). Varieties of goal-directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 103–137). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree symposium on cognition* (pp. 313–344). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Block, R. A. (1971). Effects of instructions to forget in short-term memory. *Journal of Experimental Psychology*, *89*(1), 1–9.

Cloitre, M. (1998). Intentional forgetting and clinical disorders. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 395–412). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Conway, M. A., & Fthenaki, A. (2003). Disruption of inhibitory control of memory following lesions to the frontal and temporal lobes. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *39*(4–5), 667–686.

Conway, M. A., Harries, K., Noyes, J., Racsmány, M., & Frankish, C. R. (2000). The disruption and dissolution of directed forgetting: inhibitory control of memory. *Journal of Memory and Language*, *43*(3), 409–430.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 599–607.

Dalton, P. (1993). The role of stimulus familiarity in context-dependent recognition. *Memory & Cognition*, *21*(2), 223–234.

Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in the free recall of unmixed lists. *Journal of Memory and Language*, *52*(1), 120–130.

Delaney, P. F., & Sahakyan, L. (2007). Unexpected costs of high working memory capacity following directed forgetting and context change manipulations. *Memory & Cognition*, *35*, 1074–1082.

Delaney, P. F., Sahakyan, L., Kelley, C. M., & Zimmerman, C. A. (2010). Remembering to forget: the amnesic effect of daydreaming. *Psychological Science*, *21*(7), 1036–1042.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452–478.

Depue, B. E. (2012). A neuroanatomical model of prefrontal inhibitory modulation of memory retrieval. *Neuroscience and Biobehavioral Reviews*, *36*(5), 1382–1399.

Depue, B. E., Burgess, G. C., Willcutt, E. G., Ruzic, L., & Banich, M. T. (2010). Inhibitory control of memory retrieval and motor processing associated with the right lateral prefrontal cortex: evidence from deficits in individuals with ADHD. *Neuropsychologia*, *48*, 3909–3917.

Dong, T. (1972). Probe versus free recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(5), 654–661.

Eich, E. (1985). Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 764–770.

Elmes, D. G., Adams, C., & Roediger, H. L. (1970). Cued forgetting in short-term memory: response selection. *Journal of Experimental Psychology*, *86*(1), 103–107.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*(3), 145–154.

Fernandez, A., & Glenberg, A. M. (1985). Changing environmental context does not reliably affect memory. *Memory & Cognition*, *13*(4), 333–345.

Foster, N. L., & Sahakyan, L. (2011). The role of forget-cue salience in list-method directed forgetting. *Memory*, *19*, 110–117.
Foster, N. L., & Sahakyan, L. (2012). Metacognition influences item-method directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1309–1324.
Gaultney, J. F., Kipp, K., Weinstein, J., & McNeill, J. (1999). Inhibition and mental effort in attention deficit hyperactivity disorder. *Journal of Developmental and Physical Disabilities*, *11*(2), 105–114.
Geiselman, R. E., & Bjork, R. A. (1980). Primary versus secondary rehearsal in imagined voices: differential effects on recognition. *Cognitive Psychology*, *12*(2), 188–205.
Geiselman, R. E., Bjork, R. A., & Fishman, D. (1983). Disrupted retrieval in directed forgetting: a link with posthypnotic amnesia. *Journal of Experimental Psychology: General*, *112*, 58–72.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67.
Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, *66*(3), 325–331.
Goernert, P. N., & Larson, M. E. (1994). The initiation and release of retrieval inhibition. *Journal of General Psychology*, *121*, 61–66.
Golding, J. M., & Gottlob, L. R. (2005). Recall order determines the magnitude of directed forgetting in the within-participants list method. *Memory & Cognition*, *33*(4), 588–594.
Golding, J. M., & Keenan, J. M. (1985). Directed forgetting and memory for directions to a destination. *American Journal of Psychology*, *98*, 579–590.
Golding, J. M., Long, D. L., & MacLeod, C. M. (1994). You can't always forget what you want: directed forgetting of related words. *Journal of Memory and Language*, *33*, 493–510.
Golding, J. M., & MacLeod, C. M. (1998). *Intentional forgetting: Interdisciplinary approaches*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.
Gottlob, L. R., & Golding, J. M. (2007). Directed forgetting in the list method affects recognition memory for source. *Quarterly Journal of Experimental Psychology*, *60*(11), 1524–1539.
Hanczakowski, M., Pasek, T., & Zawadska, K. (2012). Context-dependent impairment of recollection in list-method directed forgetting. *Memory*.
Hanslmayr, S., Volberg, G., Wimber, M., Oehler, N., Staudigl, T., HYartmann, T., et al. (2012). Prefrontally driven downregulation of neural synchrony medicated goal-directed forgetting. *Journal of Neuroscience*.
Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review*, *11*(1), 125–130.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.
Hoff, A. L., & Krennan, W. S. (2002). Is there a cognitive phenotype for schizophrenia: the nature and course of the disturbance in cognition? *Current Opinion in Psychiatry*, *15*, 43–48.
Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.
Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 471–480.
Isarida, T., & Morii, Y. (1986). Contextual dependence of the spacing effect in free recall. *Japanese Journal of Psychology*, *57*(1), 20–26.
Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 112–127.
Joormann, J., Levens, S. M., & Gotlib, I. H. (2011). Sticky thoughts: depression and rumination are associated with difficulties manipulating emotional material in working memory. *Psychological Science*, *22*(8), 979–983.

Joormann, J., & Tran, T. B. (2009). Rumination and intentional forgetting of emotional material. *Cognition and Emotion*, *23*(6), 1233–1246.

Joslyn, S. L., & Oakes, M. A. (2005). Directed forgetting of autobiographical events. *Memory & Cognition*, *33*(4), 577–587.

Kimball, D. R., & Bjork, R. A. (2002). Influences of intentional and unintentional forgetting on false memories. *Journal Of Experimental Psychology: General*, *131*, 116–130.

Koppel, R. H., & Storm, B. C. (2012). Unblocking memory through directed forgetting. *Journal of Cognitive Psychology*.

Krafka, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identification. *Journal of Personality and Social Psychology*, *49*(1), 58–69.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 970–988.

Lehman, M., & Malmberg, K. J. (2011a). Overcoming the effects of intentional forgetting. *Memory & Cognition*, *39*(2), 335–347.

Lehman, M., & Malmberg, K. J. (2011b, November). Modeling intentional forgetting in depressed participants. (Poster presentation at the 52nd annual meeting of the Psychonomic Society, Seattle, WA).

Liu, X. (2001). On the dynamics of directed forgetting: facilitation and interference in the updating of human memory. *Dissertation Abstracts International*, *61*(11), 6159B.

Macken, W. J. (2002). Environmental context and recognition: the role of recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 153–161.

MacLeod, C. M. (1975). Long-term recognition and recall following directed forgetting. *Journal of Experimental Psychology: Human Learning and Memory*, *1*, 271–279.

MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Mahwah, NJ: Erlbaum.

MacLeod, C. M. (1999). The item and list methods of directed forgetting: test differences and the role of demand characteristics. *Psychonomic Bulletin & Review*, *6*(1), 123–129.

MacLeod, C. M., Dodd, M. D., Sheard, E. D., Wilson, D. E., & Bibi, U. (2003). In opposition to inhibition. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 43, pp. 163–214). New York, NY US: Elsevier Science.

Macrae, C., Bodenhausen, G.V., Milne, A. B., & Ford, R. L. (1997). On regulation of recollection: the intentional forgetting of stereotypical memories. *Journal of Personality and Social Psychology*, *72*(4), 709–719.

Malmberg, K. J., & Shiffrin, R. M. (2005). The 'One-Shot' hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336.

Malpass, R. S., & Devine, P. G. (1981). Guided memory in eyewitness identification. *Journal of Applied Psychology*, *66*(3), 343–350.

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation*. (Vol. 1). New York: Academic Press.

Marsh, R. L., Meeks, J., Hicks, J. L., Cook, G. I., & Clark-Foos, A. (2006). Concreteness and item-to-list context associations in the free recall of items differing in context variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1424–1430.

McNally, R. J. (2005). Directed forgetting tasks in clinical research. In A. Wenzel & D. C. Rubin (Eds.), *Cognitive methods and their application to clinical research* (pp. 197–212). Washington, DC: American Psychological Association.

McNally, R. J., Metzger, L. J., Lasko, N. B., Clancy, S. A., & Pitman, R. K. (1998). Directed forgetting of trauma cues in adult survivors of childhood sexual abuse with and without posttraumatic stress disorder. *Journal of Abnormal Psychology*, *107*, 596–601.

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455.

Minnema, M. T., & Knowlton, B. J. (2008). Directed forgetting of emotional words. *Emotion*, *8*(5), 643–652.

Mulji, R., & Bodner, G. E. (2010). Wiping out memories: new support for a mental context change account of directed forgetting. *Memory*, *18*, 763–773.

Nelson, D. L., Goodmon, L. B., & Ceo, D. (2007). How does delayed testing reduce effects of implicit memory: context infusion or cuing with context? *Memory & Cognition*, *35*(5), 1014–1023.

Nelson, D., & McEvoy, C. L. (2005). Implicitly activated memories: the missing links of remembering. In C. Izawa & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application: The 4th Tsukuba International Conference on memory* (pp. 177–198). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299–324.

Parker, A., Gellatly, A., & Waterman, M. (1999). The effect of environmental context manipulation on memory: dissociation between perceptual and conceptual implicit tests. *European Journal of Cognitive Psychology*, *11*(4), 555–570.

Pastötter, B., & Bäuml, K.-H. (2007). The crucial role of post-cue encoding in directed forgetting and context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 977–982.

Pastötter, B., & Bäuml, K. (2010). Amount of postcue encoding predicts amount of directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 54–65.

Pastötter, B., Bäuml, K. H., & Hanslmayr, S. (2008). Oscillatory brain activity before and after an internal context change—evidence for a reset of encoding processes. *NeuroImage*, *43*, 173–181.

Pastötter, B., Kliegl, O., & Bäuml, K. T. (2012). List-method directed forgetting: the forget cue improves both encoding and retrieval of postcue information. *Memory & Cognition*, *40*(6), 861–873.

Perlstein, W. M., Carter, C. S., Barch, D. M., & Baird, J. (1998). The Stroop task and attention deficits in schizophrenia: a critical evaluation of card and single-trial Stroop methodologies. *Neuropsychology*, *12*, 414–425.

Postman, L. (1964). Acquisition and retention of consistent associative responses. *Journal of Experimental Psychology*, *67*(2), 183–190.

Power, M. J., Dalgleish, T. T., Claudio, V. V., Tata, P. P., & Kentish, J. J. (2000). The directed forgetting task: Application to emotionally valent material. *Journal Of Affective Disorders*, *57*, 147–157.

Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language*, *68*, 98–122.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Racsmány, M., & Conway, M. A. (2006). Episodic inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 44–57.

Racsmány, M., Conway, M. A., Garab, E. A., Cimmer, C., Janka, Z., Kurimay, T., et al. (2008). Disrupted memory inhibition in schizophrenia. *Schizophrenia Research*, *101*(1–3), 218–224.

Racsmány, M., Conway, M. A., Garab, E. A., & Nagymáté, G. (2008). Memory awareness following episodic inhibition. *Quarterly Journal of Experimental Psychology*, *61*(4), 525–534.

Reitman, W., Malin, J. T., Bjork, R. A., & Higman, B. (1973). Strategy control and directed forgetting. *Journal Of Verbal Learning & Verbal Behavior*, *12*, 140–149.

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition*, *2*(2), 261–269.

Roediger, H., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(1), 91–105.

Russo, R., Ward, G., Geurts, H., & Scheres, A. (1999). When unfamiliarity matters: changing environmental context between study and test affects recognition memory for unfamiliar stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 488–499.

Sahakyan, L. (2004). Destructive effects of 'forget' instructions. *Psychonomic Bulletin & Review*, *11*(3), 555–559.

Sahakyan, L., & Delaney, P. F. (2003). Can encoding differences explain the benefits of directed forgetting in the list-method paradigm? *Journal of Memory and Language*, *48*, 195–201.

Sahakyan, L., & Delaney, P. F. (2005). Directed forgetting in incidental learning and recognition testing: support for a two-factor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 789–801.

Sahakyan, L., & Delaney, P. F. (2010). Item-specific encoding produces an additional benefit of directed forgetting: evidence from intrusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1346–1354.

Sahakyan, L., Delaney, P. F., & Goodmon, L. B. (2008). "Oh, honey, I already forgot that": strategic control of directed forgetting in older and younger adults. *Psychology and Aging*, *23*, 621–633.

Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor in strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review*, *11*, 131–134.

Sahakyan, L., Delaney, P. F., & Waldum, E. R. (2008). Intentional forgetting is easier after two 'shots' than one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 408–414.

Sahakyan, L., & Foster, N. L. (2009). Intentional forgetting of action: comparison of list-method and item-method directed forgetting. *Journal of Memory and Language*, *61*, 134–152.

Sahakyan, L., & Goodmon, L. B. (2007). The influence of directional associations on directed forgetting and interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1035–1049.

Sahakyan, L., & Goodmon, L. B. (2010). Theoretical implications of extralist probes for directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 920–937.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1064–1072.

Sahakyan, L., Waldum, E. R., Benjamin, A. S., & Bickett, S. P. (2009). Where is the forgetting with list-method directed forgetting in recognition? *Memory & Cognition*, *37*(4), 464–476.

Shapiro, S., Lindsey, C., & Krishnan, H. (2006). Intentional forgetting as a facilitator for recalling new product attributes. *Journal of Experimental Psychology: Applied*, *12*(4), 251–263.

Sheard, E. D., & MacLeod, C. M. (2005). List method directed forgetting: return of the selective rehearsal account. In N. Ohta, C. M. MacLeod & B. Uttl (Eds.), *Dynamic cognitive processes* (pp. 219–248). Tokyo: Springer.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 460–471.

Smith, E. E., Barresi, J., & Gross, A. E. (1971). Imaginal versus verbal coding and the primary-secondary memory distinction. *Journal of Verbal Learning and Verbal Behavior*, *10*(6), 597–603.

Smith, S. M., & Vela, E. (1992). Environmental context-dependent eyewitness recognition. *Applied Cognitive Psychology*, *6*, 125–139.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review*, *8*, 203–220.

Soriano, M. F., & Bajo, M. T. (2007). Working memory resources and interference in directed-forgetting. *Psicológica*, *28*, 63–85.

Soriano, M. F., Jiménez, J. F., Román, P., & Bajo, M. T. (2009). Intentional inhibition in memory and hallucinations: directed forgetting and updating. *Neuropsychology*, *23*, 61–70.

Spillers, G. J., & Unsworth, N. (2011). Are the costs of directed forgetting due to failures of sampling or recovery? Exploring the dynamics of recall in list-method directed forgetting. *Memory & Cognition*, *39*(3), 403–411.

Spitzer, B., & Bäuml, K. (2007). Retrieval-induced forgetting in item recognition: evidence for a reduction in general memory strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 863–875.

Spitzer, B., & Bäuml, K. (2009). Retrieval-induced forgetting in a category recognition task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 286–291.

Steyvers, M., & Malmberg, K. J. (2003). The effects of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 760–766.

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition*, *40*(6), 827–843.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford.

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology*, *72*, 145–150.

Unsworth, N., & Engle, R. W. (2008). Speed and accuracy of accessing information in working memory: an individual differences investigation of focus switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 616–630.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012). Dynamics of context-dependent recall: an examination of internal and external context change. *Journal of Memory and Language*, *66*(1), 1–16.

Vadnais, S. A., Behm, A., Laake, L. M., Lopez, N. M., Oddi, K. B., Wu, K., et al. (2012, May). Executive function correlates of symptoms of specific anxiety disorders and major depression. (Poster presented at the 24th Annual Convention of the Association for Psychological Science). Chicago, IL.

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: retrieval-induced inhibition as cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 315–318.

Verde, M. F. (2004). The retrieval practice effect in associative recognition. *Memory & Cognition*, *32*, 1265–1272.

Waldum, E. R., & Sahakyan, L. (2012). Putting congeniality effects into context: investigating the role of context in attitude memory using multiple paradigms. *Journal of Memory and Language*, *66*(4), 717–730.

Wessel, I., & Merckelbach, H. (2006). Forgetting "murder" is not harder than forgetting "circle": listwise-directed forgetting of emotional words. *Cognition and Emotion*, *20*, 129–137.

Whetstone, T., Cross, M. D., & Whetstone, L. M. (1996). Inhibition, contextual segregation, and subject strategies in list method directed forgetting. *Consciousness and Cognition*, *5*, 395–417.

White, H. A., & Marks, W. (2004). Updating memory in list-method directed forgetting: individual differences related to adult attention-deficit/hyperactivity disorder. *Personality and Individual Differences*, *37*, 1453–1462.

Wilson, S., Kipp, K., & Chapman, K. (2003). Limits of the retrieval-inhibition construct: list segregation in directed forgetting. *Journal of General Psychology*, *130*(4), 341–358.

Zellner, M., & Bäuml, K. (2006). Inhibitory deficits in older adults: list-method directed forgetting revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 290–300.

This page intentionally left blank

CHAPTER FIVE

Recollection is Fast and Easy: Pupillometric Studies of Face Memory

Stephen D. Goldinger*,[1], Megan H. Papesh†
*Department of Psychology, Arizona State University, Tempe, AZ, USA
†Department of Psychology, Louisiana State University, Baton Rouge, LA, USA
[1]Corresponding author: E-mail: goldinger@asu.edu

Contents

Abstract

In this chapter, we discuss several distinct conceptualizations of recognition memory, and their treatment of the (putative) processes underlying recollection and familiarity. We focus most closely on the concept of recollection, which many dual-process memory models assume to be a relatively slow, controlled process, during which contextual details from encoding are brought to mind (i.e., retrieval of episodic detail). We then introduce the use of pupillometry—continuously measuring pupil diameter during task performance—as an efficient means to estimate effortful cognitive processes during memory encoding and retrieval. We review evidence from three pupillometric

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00005-8

studies of face memory, providing evidence to suggest that recollection is not slow and effortful, as is often assumed, but is instead fast and easy.

1. INTRODUCTION

Have you ever spotted a familiar colleague who had recently changed hairstyles, or had perhaps shaved off his signature mustache? You may have experienced a slightly unsettled feeling, asking "*What's different about you?*" Similarly, you may feel a seemingly inexplicable familiarity with someone you encounter at the market, avoiding interaction with that person because you cannot determine how, or even if, you know him. The latter situation exemplifies Mandler's (1980) famous "butcher on the bus" phenomenon (wherein you see your local butcher on a city bus, out of context, and cannot determine why he seems familiar), but both situations give rise to a nagging feeling of memory. Despite this subjective experience, both scenarios actually represent *failures* of memory, as you are unable to successfully retrieve information that you know to be stored in your memory. Often, these memory failures go unresolved; you cannot name the person in the market, but are certain that you know him. Contrast that situation with one more easily attributed to the success of memory: you see your butcher at the grocery store, and recognize him instantly and automatically. This experience is by far more common. Throughout the day, you routinely interact with people whom you instantly identify, fluently recollecting their names, recent conversations and personal details. Indeed, a great deal of stored memory is immediately available, with no effort, and no particular "feelings of memory" whatsoever. Although these memories (and memory failures) are experienced as disparate events, their underlying cognitive processes are nonetheless captured by two-component processes that are often said to make up "recognition memory", *recollection* of episodic details, and vague feelings of *familiarity*. This chapter will address some of the characteristics of recollection and familiarity, and will present recent pupillometric data suggesting that certain long-accepted assumptions ought to be reconsidered.

2. RECOGNITION MEMORY

In broadly conceptualizing recognition memory, in particular the experience of recollection, two dominant and opposing views have persisted for decades, single- and dual-process models. Variations upon each model exist, and each has been implemented with varying degrees of success, but

for present purposes, they will be dichotomized by their theoretical treatment of recollection. Dual-process models treat recollection as a threshold process, and they are intuitively appealing: as demonstrated above, we often have the subjective awareness of different memory states corresponding to recollection and familiarity. Dual-process models assume that these processes are separate, and in fact, independent from one another (in both cognitive and neural terms). Whereas familiarity is said to be a fast, automatic process (e.g. you do not control when items "feel" familiar), recollection is said to be slow, and deliberate. Single-process models, on the other hand, assume that recollection is essentially strong familiarity, and that both processes are subsumed within a single, strength-driven, automatic process. Rather than existing as separate entities, they represent a continuum; some memories will yield the subjective experience of recollection; others will elicit feelings of familiarity. In the ongoing debate about the nature of recognition memory, single- and dual-process models are often used to set the stage for strong inference; to accept one is often to reject the other. The various models encapsulated by these discrepant overarching theoretical frameworks all assume that memory decisions are based on the retrieval of some degree of evidence (for consistency, we will call that evidence "strength"; this is for clarity, not a theoretical stance). Interested readers should see Yonelinas & Parks (2007) for a fuller discussion of the evidence for and against each of the models (and others) discussed below.

3. MODELS OF MEMORY

3.1. Threshold-Based Models

Threshold models are often traced back to Fechner's psychophysical research (Boring, 1929), as the assumption inherent in these models is that a single "evidence" threshold must be exceeded before an item can be detected as previously encountered. *High-threshold* model is perhaps the simplest of such views: according to high-threshold model, memory is characterized by two distributions (often visually depicted as square distributions, but this is not a theoretically constrained assumption), one representing target item strength and one representing foil item strength. Each of the item types, targets and foils, is assumed to have some "resting level" of inherent strength; by virtue of the study session, targets accumulate additional strength, which serves to separate the hypothetical target and foil distributions (Figure 5.1). The "threshold", according to this model, is the upper end of the foil distribution; items to the right of this point are said to fall above threshold, and are

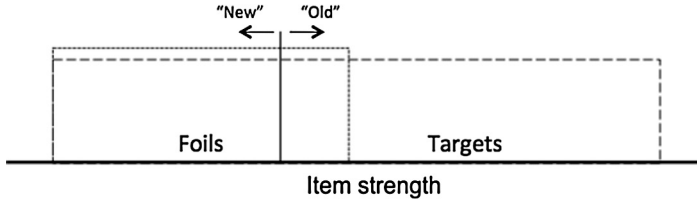


Figure 5.1 Hypothetical distribution of item strength for targets and foils in high-threshold model. Note that, because of their recent exposure, targets are assumed to have increased levels of strength.

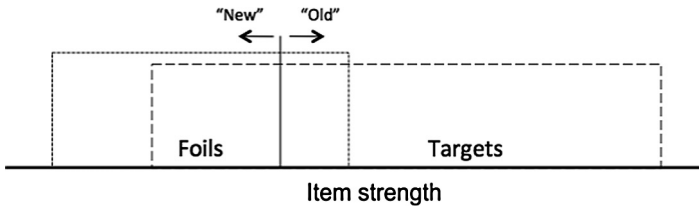


Figure 5.2 Hypothetical distribution of item strength for targets and foils in high-low threshold model. The lower end of the target distribution is the low-threshold and the upper end of the foil distribution is the high-threshold.

recognized in an all-or-none manner. This does not, however, prescribe that observers make memory decisions in reference to the threshold. Rather, decisions are made by adopting a response criterion (shown as the vertical solid line in Figure 5.1); any item with a strength value above that criterion is judged "old". As the criterion is moved to the left, the perceiver is said to respond more liberally; both hits and false alarms increase. Conservative criteria serve to decrease hits and false alarms. A critical assumption, however, is that the observer's criterion, which can move throughout an experiment, dichotomizes the distributions into binary old/new judgments.

Similarly, *high-low threshold model* (sometimes called *two-high threshold model*, see Hilford, Glanzer, Kim, & DeCarlo, 2002; Yonelinas, 2002) assumes that memory is a categorical process, but that two thresholds guide memory decisions, one high-threshold and one low-threshold. Specifically, this model also proposes two overlapping distributions of item strength for targets and foils, but, in contrast to standard high-threshold model, it assumes that the foil distribution extends beyond the lower tail of the target distribution (Figure 5.2), which allows for two thresholds. Like high-threshold models, items with strength values higher than the upper tail of the foil distribution are said to fall above threshold and are recognized as previously encountered. Unlike high-threshold models, high-low threshold models include

a method by which new items can be "recognized" as new (e.g. "I would have remembered the word *waffle* because my dog's name is Eggo."). Any new item with strength values below the lower limit of the target distribution is said to fall below threshold, and can be "recognized" as new. Observers still set a single criterion against which they make memory decisions, just as in high-threshold model, but the cognitive operations are different.

3.2. Continuous Models

In contrast to threshold models, continuous models do not make the assumption that recollection is a threshold, all-or-none process. Rather, they posit a continuous stream of evidence, capable of eliciting a range of recollected details. As summarized by Wixted (2007) and Wixted and Mickes, (2010), continuous models are based on the principles of signal-detection theory, and posit that memory decisions are made by comparing the strength of the retrieved memory signal to a decision criterion. As in standard signal-detection theory, this view proposes the theoretical existence of two Gaussian distributions (Figure 5.3), one reflecting target strength and one reflecting foil strength. As in threshold theories, targets are assumed to have greater strength, owing to their recent presentation. Also consistent with the threshold models, any item that yields item strength exceeding the decision criterion is judged "old", whereas items with lower strengths are judged "new". Performance in such models is typically described by the signal-detection index *d'*, the observer's sensitivity, which corresponds to the difference between the mean of the target and foil distributions. Although equal-variance models (Figure 5.3(a)) usefully illustrate general signal-detection-based models (and, in fact, have been incorporated into hybrid models, see Section 3.3 below), abundant evidence supports *unequal* variance distributions, wherein the target distribution is wider than the

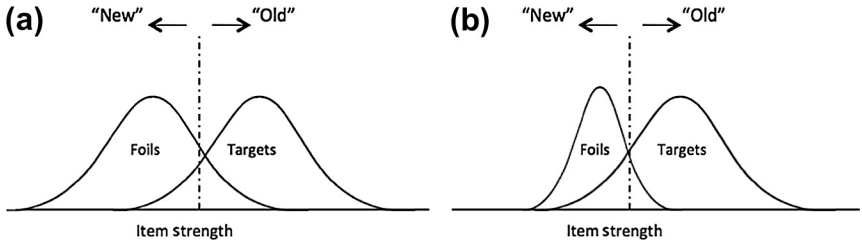


Figure 5.3 Single-process, signal-detection models of recognition memory. (a) Corresponds to equal-variance models and (b) corresponds to unequal-variance models. *(Adapted from Wixted (2007)).*

foil distribution (right panel of Figure 5.1; see, Ratcliff, Sheu, & Gronlund, 1992; Wixted & Stretch, 2004). Critically, regardless of the distributional assumptions, continuous models all propose that recognition decisions are based on a concept of *continuous memory strength* (Wixted & Stretch, 2004), not on a threshold memory process.

3.3. Hybrid Models

Contemporary threshold theories, such as dual-process signal detection theory (DPSD, see Yonelinas, 1994, 2001), incorporate elements from signal-detection theory while maintaining some of the characteristics of earlier threshold models. In similar fashion, recent continuous theories, such as continuous dual-process (CDP) signal-detection theory (Wixted & Mickes, 2010), incorporate aspects of dual-process theory, while maintaining the core assumptions of single-process models.[1]

Hybrid threshold theories, henceforth, "*dual-process theories*", are intuitively appealing. In the "butcher on the bus" example described above, people have the strong sense that they are failing to recollect episodic detail; instead the butcher is merely familiar. This feeling of unresolved knowledge can be easily contrasted with more richly detailed memories, such as not only recognizing your butcher but also remembering that he recently gave you a great recipe for grilled salmon. This intuitive mnemonic dissociation, that richly detailed memories *feel* different than relatively vague memories, captures the nature of dual-process theories: rather than assume that a single memory strength variable explains both types of memory, dual-process theories propose that recognition is served by two distinct, independent processes, *recollection* and *familiarity* (Jacoby, 1991). Whereas recollection is assumed to occur by consciously controlled processing, reflecting a person's ability to recall the specific details of the encoding event, familiarity is said to operate quickly and automatically, reflecting a vague "feeling of knowing" that the encoding event had occurred.

An early version of a hybrid dual-process model was the *two-criterion model*, wherein familiarity-based decisions are made quickly on the basis of

[1] CDP (Wixted & Mickes, 2010) combines select elements of dual-process theory with key aspects of signal-detection theory, to yield a signal-detection-based model capable of explaining subjective feelings of recollection versus familiarity. According to this model, separate (nonindependent) recollection and familiarity components exist, but memory decisions are still based on a continuous stream of evidence, composed of the additive strength from each component. Because of this, the model does not suggest that recollection and familiarity are necessarily quantitatively or qualitatively different, so the predictions are similar to those of Univariate Signal Detection Theory (UVSD) and will not be discussed further.

whether they fall above a high criterion or below a low criterion (Atkinson & Juola, 1973, 1974; as cited in Wixted, 2007). According to this model, recollection is only initiated if the memory strength falls between the two criteria, acting, in essence, as a backup process. Later, Yonelinas (1994, 2001) provided another attempt to combine the two models, with the DPSD model. According to DPSD, recollection is a high-threshold, categorical process: It either occurs or does not, and if it does occur, it will always be associated with high confidence. Familiarity, on the other hand, is generally viewed as a continuous, ahistorical memory-strength variable, capable of ranging in strength from low to high (Mandler, 1980; Yonelinas, 1994). Familiarity, being described by strength of the signal, is therefore compatible with an equal-variance signal-detection model. DPSD differs from the two-criterion model of Atkinson and Juola primarily in its order of operations (and its added quantitative detail). Whereas the two-criterion model assumes that recollection is initiated as a backup for failures of familiarity, DPSD assumes the reverse. That is, if recollection does not occur, responses are based on familiarity. Regardless of the order in which the processes are said to occur (in fact, there are many instantiations of dual-process theory in which the processes are initiated in parallel), dual-process theories almost universally consider recollection and familiarity to be separate, independent processes with distinct quantitative and qualitative characteristics (see Yonelinas, 2002 for a review).

The majority of evidence in favor of separate recollection and familiarity processes comes from *functional process dissociations*, or manipulations that affect the contributions of each system independently. For example, several studies have indicated that responses based on familiarity are faster than those based on recollection (e.g. Hintzman & Caulton, 1997; Hintzman, Caulton, & Levitin, 1998). Furthermore, it has been shown that this "fast familiarity" process increases false alarms immediately after the presentation of a new item, but the false alarm probability drops off with increased response time, reflecting slower recollective retrieval dynamics (e.g. Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; Jacoby, 1999; McElree, Dolan, & Jacoby, 1999). Using event-related potentials (ERPs), many studies have also found distinct electrophysiological correlates for recognition memory responses based on recollection versus familiarity (Curran, 2000; Duarte, Ranganath, Winward, Hayward, & Knight, 2004; Guo, Duan, Li, & Paller, 2006; Klimesch et al., 2001) and some have found opposite effects on recollection and familiarity following hippocampal damage (Sauvage, Fortin, Owens, Yonelinas, & Eichenbaum, 2008). These dissociations (and others)

strongly suggest that two separate neural substrates underlie recognition memory.[2]

If recollection and familiarity are separate processes, then they likely also have distinct temporal properties. In virtually all dual-process conceptualizations of recollection and familiarity, familiarity is agreed to be the "faster", more automatic, process. Evidence supporting this assumption comes from experiments in which processing speed has been manipulated, such that perceivers have more or less time to make memory decisions, or when decision times are measured. For example, Hintzman & Curran (1994) presented participants with lists of singular and plural nouns to study for an upcoming memory test. Test stimuli consisted of old items, completely new items, or distractors that were related to studied items, but had their plurality reversed (e.g. if WAFFLE was studied, then WAFFLES would be a related distractor word). Assuming that participants must *recollect* the studied item in order to reject a related distractor, Hintzman & Curran (1994) used response times to infer the temporal characteristics of recollection. They found that recollection was relatively slow: participants discriminated old from new words approximately 120 ms faster than they could discriminate old words from the related distractors. This finding, that the familiarity process terminates faster than the recollection process, has been observed repeatedly (e.g. Gronlund, Edwards, & Ohrt, 1997; Hintzman et al., 1998), and is the basis for the assumption that familiarity is fast and automatic, while recollection is slow and deliberate. Despite this, other researchers have documented instances in which "old" responses based on familiarity are slower than responses based on recollection (e.g. Dewhurst, Holmes, Brandt, & Dean, 2006; Duarte et al., 2004; Wheeler & Buckner, 2004).

4. ESTIMATING RECOLLECTION AND FAMILIARITY

Although several methods exist for estimating the contributions of recollection and familiarity to recognition memory, the two most popular are the *process-dissociation procedure* (PDP; Jacoby, 1991) and the *remember/know paradigm* (RK; Tulving, 1985). PDP was designed to separate the influences of recollection and familiarity in recognition memory and to elucidate the notion that *task distinctions* do not necessarily map onto *process*

[2] Because the focus of this chapter is on the psychophysiological characteristics of putatively recollection-based memory decisions, a review of the single- versus dual-process debate regarding the neuroscience of is beyond the scope of this chapter. Interested readers are encouraged to consult Squire and Wixted (2011 or Wais, Squire, and Wixted, 2009).

distinctions. Specifically, Jacoby (1991) suggested that researchers should not assume that a task is capable of measuring the exact processes that it is intended to measure, as results from different tasks may reflect the tasks, and not the processes assumed by the researcher. To demonstrate this in recognition memory, Jacoby (1991) developed a procedure that *does* distinguish between processes, while using a constant task, so as to avoid the problem of comparing results from separate tasks.

4.1. The Process-Dissociation Procedure

The major assumption underlying PDP is that recollection and familiarity *independently* contribute to memory (e.g. Mandler, 1980), such that their contributions can be estimated by putting responses based on them in mutual opposition (Jacoby, 1991). When automatic processes (i.e. familiarity) oppose consciously controlled processes (i.e. recollection), researchers can infer which is the more dominant contributor to recognition memory. By analyzing the results of complementary facilitation and interference paradigms, one is able to numerically estimate the extent to which automatic and controlled processes contribute to recognition responses, assuming a process dissociation is observed.

Whereas recollection has been proposed as an intentional (consciously controlled) use of memory (Jacoby, 1991), and is therefore likely a relatively slow process, familiarity contributes automatically (unconsciously), and on a relatively fast time course. In the facilitation paradigm (henceforth referred to as inclusion), participants are typically presented with two lists of stimuli that differ in some concrete way (e.g. List A contains words that are heard and List B contains words that are read).[3] During a subsequent recognition memory test, participants are presented with words from both lists, along with new words. All previously encountered words, regardless of original list, are to be called "old". It is assumed that recollection and familiarity both contribute to "old" judgments during the inclusion test. Based on Jacoby (1991), the probability of issuing an "old" response to words that were heard (h) during study would be $O_{Ih} = R_h + F_h - R_hF_h$, where "I" indicates the inclusion test, and R and F represent the contributions of recollection and familiarity, respectively.

Suppose that participants are then presented with another set of two lists (again heard and read, separately) to study. During test, they are again

[3] Note that Jacoby also modified the paradigm to test associative memory and stem completion (Jacoby & Rhodes, 2006). Only the classic paradigm will be discussed here.

presented with new words, as well as those from the two study lists, and are asked to make recognition responses. However, instead of responding "old" to any previously encountered words, participants are now asked to *exclude* specific items. In this example, assume that they are asked to exclude words that were previously heard, responding "old" only to those words that were read during study. According to Jacoby, the probability of issuing an "old" response to a previously heard word is now given by $O_{Eh} = F_h(1 - R_h)$, where E denotes the exclusion test. Based on simple algebra, this can be represented as $O_{Eh} = F_h - R_hF_h$. In words, the probability of incorrectly responding "old" to a previously heard word during the exclusion test is equal to the familiarity of that word minus the multiplied contributions of recollection and familiarity for heard words. The contribution of recollection can thus be estimated by $R_h = O_{Ih} - O_{Eh}$, which simply indicates that recollection and familiarity both contribute to "old" responses in the inclusion test, but correct responses to excluded items during the exclusion test can only be based on recollection: Participants must recollect an item's presentation modality to reject it. Subtracting the influence of familiarity from the combination of R and F leaves only R. Familiarity can thereby be estimated as $F_h = O_{Eh}/(1 - R_h)$.

Although the PDP method has considerable utility, one of the key criticisms of PDP regards the definition of recollection. Specifically, recollection is a relatively vague term, as people can (arguably) experience varying degrees of recollection. Participants can recall varying degrees of information about List 1 and List 2 items. For example, if spoken words are used, voice can be a critical detail and participants may recall only gender or more detailed, speaker-specific information (as in Dodson, Holland, & Shimamura, 1998). In PDP experiments, both memories would be classified as recollection, despite their varying degrees of specificity. To address this, researchers may differentiate between *diagnostic* and *nondiagnositic* recollection (Gruppuso, Lindsay, & Kelley, 1997; Mulligan & Hirshman, 1997; Yonelinas & Jacoby, 1996), with diagnostic information reflecting memory for the source of test items and nondiagnostic information reflecting other, noncriterial, memorial information. From this, it is clear that the recollection parameter in PDP only measures the overall amount of diagnostic memorial information, ignoring a potentially vast amount of other recollective experiences.

Additional criticisms of PDP regard the underlying assumptions, primarily the assumption that recollection and familiarity independently contribute to recognition memory (Curran & Hintzman, 1995, 1997; Joordens

& Merikle, 1993; Richardson-Klavehn, Gardiner, & Java, 1996; but see Jacoby, 1998; Jacoby, Begg, & Toth, 1997). This issue has been extensively debated and, like the single-/dual-process distinction, each side has interpreted evidence as consistent with their view. For example, Curran and Hintzman (1995, 1997) argued that high correlations between recollection and familiarity at the item level violate the PDP independence assumption, whereas Jacoby et al. (1997) argued that the correlations are "trivial" and likely reflect violations of the boundaries within which PDP must be used. Because the estimates of recollection and familiarity provided by PDP critically rely on their functional independence, equivocal empirical results necessitate investigations via other paradigms.

4.2. The Remember/Know Paradigm

Another approach to dissecting recognition memory is the RK paradigm (Tulving, 1985), which makes use of subjects' subjective feelings of the relative specificity of their memory. Although Tulving (1985) initially intended for the procedure to differentiate between states of awareness associated with subjective experiences of memory, it has more recently been used to support dual-process theories of recognition memory (Wixted & Mickes, 2010). In this paradigm, participants are assumed to appreciate why they make old/new recognition decisions. If an item is judged "old", there are three possible routes to this decision; an item can be "remembered", "known" (henceforth R and K) or simply guessed (although not all studies use a "guess" option). Jacoby, Yonelinas, & Jennings (1997) suggested that remember responses reflect episodic retrieval, and the function of conscious recollection, whereas know responses reflect familiarity, or the recognition of an item's status as "old" without concomitant recollection of its earlier presentation (Yonelinas, 2002).

Results from several studies (e.g. Engelkamp & Dehn, 1997; Gardiner, 1988; Gardiner & Java, 1991; Gardiner, Java, & Richardson-Klavehn, 1996; Gardiner & Parkin, 1990; Rajaram, 1993) indicate that different experimental manipulations selectively enhance or diminish remember/know response frequencies, supporting Gardiner's (1988) reports of functional dissociations between R and K responses. These dissociations imply that recollection and familiarity are distinct, independent processes, and their existence is typically interpreted within a dual-process framework (Gardiner, 2001, as cited by Dunn, 2004).

This interpretation has been criticized, however, as several researchers have proposed that the response types reflect *confidence* in memory more

than they reflect the function of two separate memory processes (Donaldson, 1996; Hirshman, 1998). Instructions to respond "remember" or "know", according to this view, are interpreted by participants as a requirement to adopt a more conservative or liberal response criterion, respectively (Dunn, 2004). Donaldson (1996) approached the RK task from a single-process viewpoint, and suggested that participants complete the task by adopting two-decision criteria, one (high-criterion) for remember responses and one (low-criterion) for know responses. He argued that, although participants were issuing responses that appeared to reflect two different memory systems, they were responding in line with their decision criteria, which reflected memory strength, and not separate processes (cf. Knowlton & Squire, 1995). Meta-analyzes of decades of data and critical tests of recent data have been taken to support both single-process (Donaldson, 1996; Dunn, 2004, 2008; Hirshman & Master, 1997; Wixted, 2007; Wixted & Stretch, 2004) and dual-process models (Conway, Dewhurst, Pearson, & Sapute, 2001; Gardiner, Ramponi, & Richardson-Klavehn, 1998, 2002; Yonelinas, 2002).

The hybrid CDP model (Wixted & Mickes, 2010) was developed in response to the dual-process interpretation of RK data. Although the model is entirely based upon signal-detection theory, it assumes that participants are able to determine whether their memorial experiences are based on the function of recollection or familiarity. In short, CDP assumes that recognition decisions reflect the combined influence of recollection and familiarity, such that decisions are still based on a "strength of evidence" dimension, but that participants have access to the source of the predominant strength. Evidence for this model comes from RK tasks in which participants provide both confidence estimates and source discriminations. During encoding, participants study words presented at either the top or bottom of the screen, in red or blue font. When prompted to recognize a test item (centrally presented, in black), participants provide RK judgments, confidence estimates, and responses regarding original location and color (source discriminations). Wixted and Mickes found that, although "know" responses can be associated with a high degree of confidence, corresponding source accuracy was lower, relative to "remember" responses.

As discussed above, neuropsychological evidence has consistently been interpreted within a dual-process framework, and this is largely because of the widespread use of RK in neuroimaging studies. In many experiments, researchers have observed elevated activity in the hippocampus for R judgments, and almost no activity for K judgments (e.g. Aggleton et al., 2005; Eldridge, Engel, Zeineh, Bookheimer, & Knowlton, 2000; Holdstock,

Mayes, Gong, Roberts, & Kapur, 2005; Moscovitch & McAndrews, 2002; Uncapher & Rugg, 2005; Verfaellie, Rajaram, Fossum, & Williams, 2008; Yonelinas et al., 2002). One difficulty in interpreting these effects, however, is that R judgments are typically associated with higher confidence, relative to K judgments (e.g. Dunn, 2004, 2008; Rotello & Zeng, 2008; Wixted & Mickes, 2010; Wixted & Stretch, 2004), making it impossible to determine whether the effect arises from differences in fundamental processes, or differences in retrieved strength. This "strength confound" was originally pointed out by Wixted (2009), who noted that recollective detail is almost never entirely absent from know judgments (Mickes, Wais, & Wixted, 2009). In fact, when subjective retrieval strength was equated, Wais, Squire, & Wixted (2009) observed similar levels of hippocampal activity during putatively recollection-based and familiarity-based memories (see also Kirwan, Wixted, & Squire, 2008; Wais, 2008; Wais, Wixted, Hopkins, & Squire, 2006). In order to interpret RK data, it seems, one needs to take an additional step and collect overt, metacognitive estimates of confidence.

5. PUPILLOMETRY

5.1. The Neurophysiology of the Pupillary Reflex

To further examine differences in processes that are said to be fast and automatic, versus slow and deliberate, researchers can appeal to neuroimaging evidence, as discussed above, or they can examine a recently rediscovered correlate of cognitive activity, pupil diameter. *Pupillometry*, the measurement of the diameters of the eyes' pupils, has been used for centuries to examine visual and cognitive processing (e.g. Fontana, 1765). Although it is well-known that the pupils dilate in response to changes in ambient lighting, it is less well-known that the pupils also dilate in response to nonvisual stimuli, such as emotions and thoughts (Goldwater, 1972; Janisse, 1974; Loewenfeld, 1958). This distinction characterizes two independent types of pupillary reflex, *tonic* changes, which occur in response to general factors, such as emotional arousal, stress, and anxiety, and *phasic* changes, which occur following the onset of stimuli for cognitive processing (Karatekin, Couperus, & Marcus, 2004). These cognitively evoked pupillary reflexes occur following inhibition of the parasympathetic nervous system's Edinger–Westphal nucleus (Steinhauer, Siegle, Condray, & Pless, 2004), which is controlled by the locus coeruleus–norepinephrine (LC–NE) system. The LC is a subcortical brain system that contains the noradrenergic system, which is the sole source of the neurotransmitter NE. This system plays a critical role in the

control of attention (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; for a review, see Laeng, Sirois, & Gredebäck, 2012). A role for the LC–NE system in memory consolidation has been determined by documenting LC–NE activity during memory retrieval (Sterpenich et al., 2006) and slow-wave sleep (Eschenko & Sara, 2008). Relevant to the experiments we summarize below, the LC–NE system is also critically involved in the pupillary reflex (Koss, 1986). In combined single-cell recording and pupillometry studies with monkeys, researchers have documented a tight correspondence between pupillary reflexes and activity in cells within the LC–NE system (Rajkowski, Kubiak, & Ashton-Jones, 1993; Rajkowski, Majczynski, Clayton, & Ashton-Jones, 2004).

Further neurophysiological evidence for the influence of cognitive processing on pupil size comes from examining the muscles that control pupillary reflexes. Two muscles are known to control pupil dilation and constriction, the dilator and the sphincter; these muscles are differentially affected by activation in the sympathetic and parasympathetic systems (Steinhauer et al., 2004). As noted above, inhibition of the parasympathetic nervous system has been attributed to dilation resulting from cognitive processing. This system also controls the activity of the sphincter muscle; when the parasympathetic system is inhibited, activity on the sphincter muscle decreases (Steinhauer et al., 2004). These autonomic pathways hold reciprocal connections with the central nervous system (CNS), so it has been suggested that they can modulate, or be modulated by, CNS structures related to cognition (Gianaros, Van Der Veen, & Jennings, 2004; Steinhauer et al., 2004). Investigations into the neural mechanisms of successful learning and memory in animals have revealed a close correspondence between accurate performance and the involvement of the autonomic system (Croiset, Nijsen, & Kamphuis, 2000). Such findings are paralleled by recent findings from human experiments, wherein increased autonomic responses (e.g. skin conductance) are positively correlated with memory strength for emotional words (Buchanan, Etzel, Adolphs, & Tranel, 2006). Additionally, stimulation of the vagus nerve (a parasympathetic pathway known to carry signals to the brain) is associated with memory formation and consolidation (Clark, Naritoku, Smith, Browning, & Jensen, 1999).

5.2. Task-Evoked Pupillary Responses

As noted, phasic changes in pupil diameter occur following the onset of cognitive processing. These reflexes are observed independently of tonic changes; in dark-adapted conditions, which inhibit the parasympathetic

system, the pupils reliably dilate in response to cognitive demand (Steinhauer & Hakerem, 1992), leading them to be referred to as *task-evoked pupillary responses* (TEPRs). Although Eckhard Hess is often credited for initiating the psychological study of pupillary reflexes (cf., Hess, 1965; Hess & Polt, 1964; Hess, Seltzer, & Shlien, 1965), his research focused almost exclusively on the pupillary reflex as it reflected "emotionality" (Hess, 1965, p. 46), a tonic response. Since then, the "emotional" component of the pupillary reflex and the "cognitive" component have been clearly dissociated (Partala & Surakka, 2003; Stanners, Coulter, Sweet, & Murphy, 1979). Kahneman and Beatty (1966) and Beatty & Kahneman (1966) are best known for initiating interest in TEPRs, and even suggested that TEPRs reflect a "summed index" of brain activity during cognitive processing. Their early work demonstrated that pupil dilations are time-locked to cognitive processing, and that differences between and within tasks are observable via pupillometry (Kahneman, 1973; Kahneman, Beatty, & Pollack, 1967). For example, in a digit recall task, as participants were given more numbers to retain, their pupils became larger; as the digits were recalled, the pupils constricted with each additional item (Kahneman & Beatty, 1966). Although pupillometry fell out of favor for some time, it has been used to infer cognitive effort in a variety of domains, such as lexical decision (Kuchinke, Võ, Hofmann, & Jacobs, 2007), attention (Kahneman, 1973; Karatekin et al., 2004), word processing (Papesh & Goldinger, 2012), working memory (Granholm, Asarnow, Sarkin, & Dykes, 1996; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004), face perception (Goldinger, He, & Papesh, 2009), general cognitive processing (Granholm & Verney, 2004), and recognition memory (Kafkas & Montaldi, 2011; Papesh, Goldinger, & Hout, 2012).

The appeal of pupillometry lies in its relative independence: phasic pupillary reflexes are largely impervious to attempts to consciously manipulate them (in fact, in our research, most participants assume that the eye-tracking equipment is primarily to track the movement of their eyes across the screen). Further, whereas many neuroimaging methods are time-consuming, expensive, and require the use of tightly controlled environments, pupil measurements can be obtained with relatively basic eye-tracking equipment (e.g. eye-trackers recording at 50 Hz are capable of providing sensitive pupillometric data), with fewer environmental and methodological restrictions. Given the tight correspondences between activity in neuroanatomical areas relevant for cognitive processing and dynamic changes in pupil size, pupillometry represents an ideal, efficient means for estimating neural activity during many cognitive events (Kahneman, 1973).

5.3. Pupillometry and Memory

In the study of memory, pupillometry can be likened to ERP waveforms (Beatty, 1982); enlarged pupils are typically associated with increased cognitive demand (Granholm & Steinhauer, 2004; Porter, Troscianko, & Gilchrist, 2007). Comparing neurophysiological measures across study and test has been used to differentiate the neural activity associated with subsequently remembered versus forgotten information in both functional Magnetic Resonance Imaging (fMRI) (e.g. Ranganath et al., 2004) and ERP investigations (e.g. Cansino & Trejo-Morales, 2008; Duarte et al., 2004; Guo et al., 2006). The logic underlying such studies is that encoding should utilize the same set of processes and neural substrates that are subsequently recruited during successful retrieval. Moreover, the strength and type (e.g. recollection or familiarity) of memory should be observable from different patterns of activation during both encoding and retrieval. In our recent research, pupillometry was used to examine encoding and recognition effort, acting in place of metacognitive confidence estimates.

Although early TEPR investigations were criticized on the grounds that now-standard experimental controls were not implemented, recent work incorporates strict experimental control to eliminate the unwanted influence of tonic reflexes (Võ et al., 2008). Because pupils dilate reflexively to changes in luminance, color, or the spatial frequency composition of the visual input, care must be taken to equate, as much as possible, stimulus characteristics in experimental designs that utilize pupillometry (Porter et al., 2007). Porter and Troscianko (2003) identified several methodological approaches that minimize unwanted pupillary reflexes, including the use of relatively low-stimulus contrast, avoiding colored stimuli, and using relatively long-stimulus exposure durations. Goldinger and Papesh (2012) recently added to this list of constraints by suggesting the use of relatively long (e.g. 1000 ms or more) intertrial intervals (ITIs) and baseline-correction procedures. Both suggestions guard against *carryover effects*, wherein the difficulty of trial n influences the waveform of trial $n + 1$. Recent work has taken such precautions into careful consideration, including several relevant studies on the pupillary reflex and memory.

Early work on the effects of memory on pupil size documented that familiar items elicited increases in pupil diameter, relative to unfamiliar items. For example, Gardner, Mo, & Borrego (1974) demonstrated that familiar consonant–vowel–consonant (CVC) trigrams resulted in enlarged pupils, relative to unfamiliar CVCs; Gardner, Mo, & Krinsky (1974) observed the

same effect with consonant trigrams. More recent research, incorporating the methodological constraints cited above and motivated by the similarity between pupillary waveforms and ERP waveforms, which are known to reflect memorial processes (Dietrich et al., 2000; Johansson, Mecklinger, & Treese, 2004), has also documented effects of long-term memory on pupil dilation. Võ et al. (2008) had participants study a series of positively and negatively valenced words, followed by a speeded old/new recognition test. Although they found effects of emotionality (e.g. better memory for emotional words and smaller pupils to negative words), the most relevant finding was a "pupillary old/new effect", wherein pupils were larger during test trials leading to hits, relative to correct rejections. The authors interpreted this effect within a dual-process framework (as in Yonelinas, 2001, 2002), suggesting that the increased pupil size observed for hits was directly related to the occurrence of recollection.

A similar conclusion was drawn by Kafkas and Montaldi (2011), who investigated the nature of incidental memory for images using a modified RK procedure. In their study, participants viewed images without instruction to remember them, and were later given a surprise memory test in which they distinguished memories based on degrees of familiarity (e.g. F1–F3) or a single recollection response. Although they did not observe an effect differentiating recollection from familiarity, pupil size during encoding differed based on subsequent memory strength: as subsequent memory increased, pupil diameter decreased. These findings are suggestive, but should be interpreted with caution. Kafkas & Montaldi's (2011) procedure eliminated TEPRs during the encoding phase, leaving only tonic changes free to vary.

In an investigation of phasic pupillary changes and memory strength, Papesh et al. (2012) observed the opposite pattern. Participants in their experiment intentionally studied a series of words spoken by multiple talkers; using an intentional encoding process, and feature-matched words, allowed them to eliminate tonic pupillary reflexes, leaving phasic reflexes free to vary. Following a brief break, participants made confidence-based old/new decisions (e.g. 1-very sure new to 7-very sure old) to old and new spoken words. Old words were spoken in one of the three voices—the studied voice, a familiar voice, or a new voice. Examining pupil size by subsequent memory performance revealed a clear relationship between confidence during test and pupil size during encoding: as subsequent confidence increased, so too did pupil size. Or, put another way, pupil dilation during learning predicted future memory performance, on an item-by-item basis.

Further, pupil sizes were also larger when the study and test voices matched, relative to when they did not match. Together, these findings revealed that during intentional encoding of spoken material, the degree of cognitive effort devoted to encoding can accurately predict subsequent recognition memory.

Papesh et al. (2012) interpreted their results within Whittlesea's (1997) Selective Construction and Preservation of Experience (SCAPE) framework for recognition decisions. According to Whittlesea and Leboe (2000) and Whittlesea and Williams (1998, 2001), recognition decisions are made in two stages. The first stage is characterized by the *production* of mental states, such that the perceptual material (e.g. the test item) is identified, elaborated, and any missing information is "filled in" by bringing to mind associated images or ideas (Neisser, 1967). In the second step, the perceiver *evaluates* the fluency of the production process. Rather than directly evaluating the stimulus and comparing memory strength to a decision criterion, the perceiver evaluates the relative harmony of mind; how easily did the production process end? If the production process was characterized by "extra" elaboration, then the expectation of fluent processing is violated, and the perceiver is left with a sense of memory failure (e.g. a sense of familiarity). Because Papesh et al. (2012) essentially observed a pupillary reinstatement of cognitive processing across study and test, they concluded that memory traces are composed of rich, detailed events, and that memory decisions are characterized by an assessment of how easily those details are brought to mind.

Otero, Weekes, & Hutton (2011) obtained a similar result, and proposed that pupillometry can be used to reveal the strength of memories across encoding and retrieval. To more closely examine the roles of recollection and familiarity, they employed the RK procedure, asking participants to provide remember/know judgments for all "old" decisions. Using emotionally neutral words, they obtained a pupillary old/new effect that was stronger when participants reported the subjective experience of recollection: pupils were largest when participants reported remembering, and were smallest when participants correctly rejected items. Responses based on familiarity (know responses) yielded diameters that were reliably smaller than those accompanying remember responses, but they were not significantly different from correct rejections. Although this pattern is consistent with Võ et al.'s (2008) interpretation, that the pupillary old/new effect is driven by the cognitively demanding experience of recollection, when Otero et al. (2011) manipulated levels of processing during study, they observed larger diameters for items processed in deep, relative to shallow, orientations. This finding argues

against the interpretation of recollection in the pupillary old/new effect. Specifically, if recollection is cognitively demanding, which yields enlarged pupils, encoding manipulations known to increase the memorability of items should not also yield *larger* diameters, but they do. Otero et al. (2011) interpreted their findings as reflecting the relative strength of memories. Consistent with a single-process, signal-detection model (and, we argue, consistent with the SCAPE framework), they observed that stronger memories yielded enlarged pupils, and that this effect was not limited to conscious recollection. Papesh (2012) observed a similar result. Across encoding and retrieval of high-frequency and low-frequency words, stronger memories (and subsequently stronger memories) were associated with enlarged pupils (Figure 5.4). Even during encoding, when the pupil measurement was temporally divorced from memory and confidence decisions, pupil size accurately tracked subsequent memory strength.

Interestingly, Otero et al. (2011) also observed a pupillary old/new effect in false alarms; pupils were larger when items were falsely recognized, relative to when they were called "new". Subsequent work by Heaver and Hutton (2011) further suggested that these memory-based effects in the pupillary reflex are automatic, and not capable of influence by cognitive control. In addition to replicating the pupillary old/new effect, they found that this

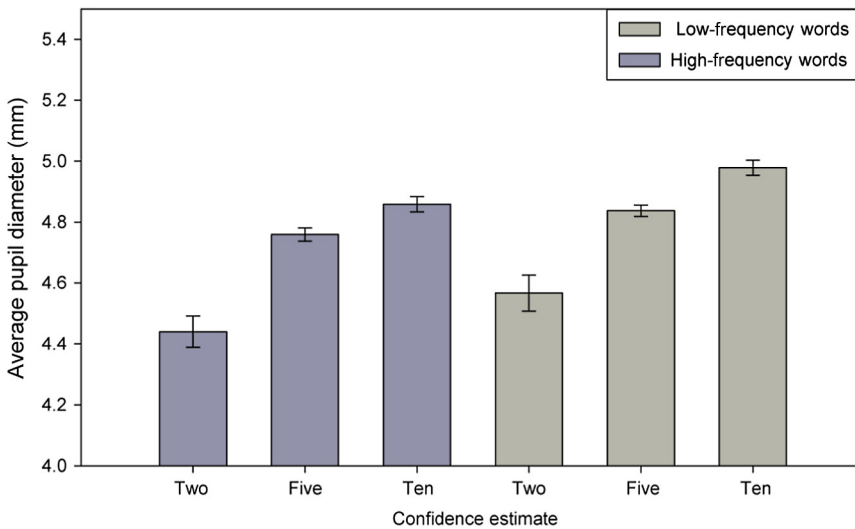


Figure 5.4 Pupil diameter as a function of confidence and word frequency during a recognition test. Error bars represent standard errors of the mean. *(From Papesh (2012)).* (For color version of this figure, the reader is referred to the online version of this book.)

effect was impervious to instructions to feign amnesia or to call every test item "new", regardless of its old/new status. Across three experiments, Heaver and Hutton (2011) observed larger pupil sizes during old, relative to new, test trials, even when participants did not overtly judge the items as "old".

If phasic pupillary changes are positively correlated with cognitive demand, why is retrieving a memory associated with more "demand" than correctly rejecting an experience as novel? This question underscores the differences across studies such as Kafkas and Montaldi (2011), in comparison to those by Otero et al. (2011) and Papesh et al. (2012). Whereas Kafkas and Montaldi (2011) suggested that the experience of future recollection is associated with pupillary constriction, the latter researchers suggested that pupil dilation is related to the strength of memories, and the success of memory retrieval. It is difficult to fully compare these studies, however, due both to differences in stimuli and in methodology. For example, whereas Kafkas and Montaldi (2011) used incidental encoding instructions, Otero et al. (2011) and Papesh et al. (2012) used intentional encoding instructions. This methodological difference alone is capable of explaining the discrepant findings, as incidental encoding does not involve the same degree of demand as intentional encoding, and may in fact yield only tonic reflexes.

6. PSYCHOPHYSIOLOGICAL CORRELATES OF MEMORY FOR FACES

Recently, we have begun to apply converging methods to more fully investigate the pupillary old/new effect, and the possible indices of recollection and/or familiarity in the pupillary reflex. To do this, we have conducted several experiments using the same stimuli, but different methods, each of which is aimed at eliciting measures of recollection versus familiarity. By using the same stimuli, we can begin to more fully appreciate the mnemonic properties of the pupillary reflex.

Goldinger et al. (2009) reported an investigation of the *own-race bias* (ORB) in face recognition, in which we used eye movements and pupillometry to provide insight into a well-studied cognitive phenomenon. Asian and White participants had their eye movements and pupil sizes monitored as they encoded a series of Asian and White faces. Afterward, participants discriminated old from new faces, and were labeled as "high performers" (i.e. participants who were relatively good at recognizing other-race faces) and "low performers" (participants who performed relatively poorly). Goldinger et al. (2009) also examined pupillary waveforms during encoding

as a function of subsequent memory, and observed a strong race effect. Specifically, when encoding own-race faces, participants' pupils dilated much less, relative to when they encoded other-race faces. More interestingly, we also observed an effect of subsequent memory performance. High performers devoted greater effort to encoding other-race faces; participants who subsequently scored poorly to other-race faces had significantly less pupil dilation during encoding. In fact, as the encoding trials progressed, subsequently low-scoring participants gradually "withdrew effort"; they began moving their eyes less vigorously (i.e. they gathered less information) and their pupil sizes peaked at much lower diameters. These results are consistent with those reported by Otero et al. (2011) and Papesh et al. (2012) pupil dilation reveals the experience of memory, even during encoding.

6.1. PDP and Pupillometry

Although these results were suggestive, they do not speak to the experiences of recollection versus familiarity. To investigate this issue, Papesh and Goldinger (2011) conducted two follow-up studies using well-known memory paradigms to separate the contributions of recollection and familiarity, the PDP and RK paradigms described earlier. In both experiments, all participants were White, as recruiting recent Asian immigrants is difficult, and the stimuli had already been verified to result in clear ORBs across both races. In our PDP experiment, 36 participants were divided into two groups, the exclusion group (n = 19) and the inclusion group (n = 17). All participants studied a series of 28 Asian and White faces for 5 s each; each face was bordered in either blue or purple, which was randomly selected on each trial with the constraint that each race has an equal number of each color. During the subsequent recognition test, exclusion participants were instructed to "exclude" (i.e. call "new") certain faces based on their previously seen borders (all test faces were presented without colored borders). Some participants were told to exclude any faces that were originally presented with blue borders; others were instructed to exclude faces that originally had purple borders. Participants in the inclusion group received standard recognition instructions (i.e. call all studied faces "old").

Papesh and Goldinger (2011) replicated the standard ORB, such that participants were better able to recognize own-race faces, relative to other-race faces. More important, we observed clear differences in pupillary waveforms during encoding as a function of subsequent memory performance. Unlike Goldinger et al. (2009), who classified participants based on their overall memory performance, Papesh and Goldinger (2011) divided the waveforms

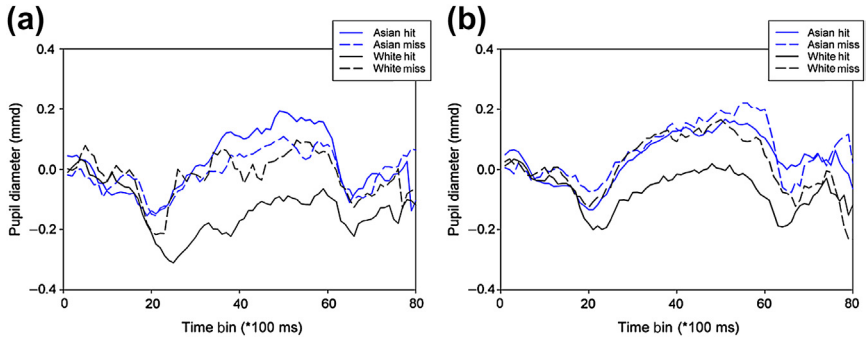


Figure 5.5 Baseline-corrected pupil diameter as a function of face race and subsequent memory performance in the exclusion (a) and inclusion (b) groups. (For color version of this figure, the reader is referred to the online version of this book.)

by memory type. Figure 5.5 shows pupillary results during face encoding. In each panel, the data represent baseline-corrected pupil diameters, shown across an 8 s span. Recall that faces were only visible for 5 s, which is why there are characteristic drops in pupil diameter at the 6 s mark (pupillary changes typically lag by about 500 ms, relative to stimulus changes). Although it is not shown in these figures, we also impose variable ITIs to ensure that pupil diameters return to their resting values before the next face is shown.

In Figure 5.5(a,b), separate functions are shown, representing the race of the studied face (Asian, White) and the eventual recognition outcome (Hit, Miss). Figure 5.5(a) shows data from the exclusion condition; Figure 5.5(b) shows the inclusion condition. In both (a) and (b), similar findings are apparent. First participants devoted greater effort, as operationalized by pupil size, to encoding other-race faces, as the blue (Asian) functions are generally higher than the black (White) functions. Second, and of particular interest, the pupil functions representing future hits show a dramatic asymmetry: when participants studied Asian faces, and would later produce recognition hits, they devoted considerable effort during encoding (this is more easily appreciated in the left-hand panel, but was true in both conditions), replicating Goldinger et al. (2009). When face encoding would lead to future misses, no differences were observed as a function of race. Most intriguing, when participants studied own-race faces, and would later produce recognition hits, their pupils were actually constricted, relative to baseline.

The results in Fig. 5.5 suggest that successful face encoding requires different cognitive resources, depending on the face of the race. However, it remains unclear how these results address the question of future recollection versus familiarity. To more fully appreciate the role of recollection, consider

just the exclusion group, specifically the "excluded" faces. As reviewed earlier, in a correct exclusion trial, the participant must first recognize the face as having been studied. After that, she must also retrieve the original border color. In other words, the participant must recollect the item in order to successfully exclude it. On the other hand, if the participant fails to correctly exclude the face, it can be inferred that there was a failure of recollection: either the participant did not recollect the face, or she may have falsely remembered the wrong border. Given these assumptions, researchers can infer that correct exclusion trials are completed on the basis of recollection, whereas incorrect exclusion trials represent the function of familiarity, in the absence of recollection. As shown in Figure 5.6, we observed significantly larger pupils during encoding trials that led to *incorrect* exclusion. When participants would subsequently experience detailed recollection, their pupil sizes were smaller, relative to when they would experience vague feelings of memory.

From these results, Papesh and Goldinger (2011) suggested that, consistent with anecdotal experiences of easily recognizing well-known acquaintances, recollection is typically fast, automatic, and cognitively easy. Although our earlier work suggested that pupil size nicely tracks future memory strength,

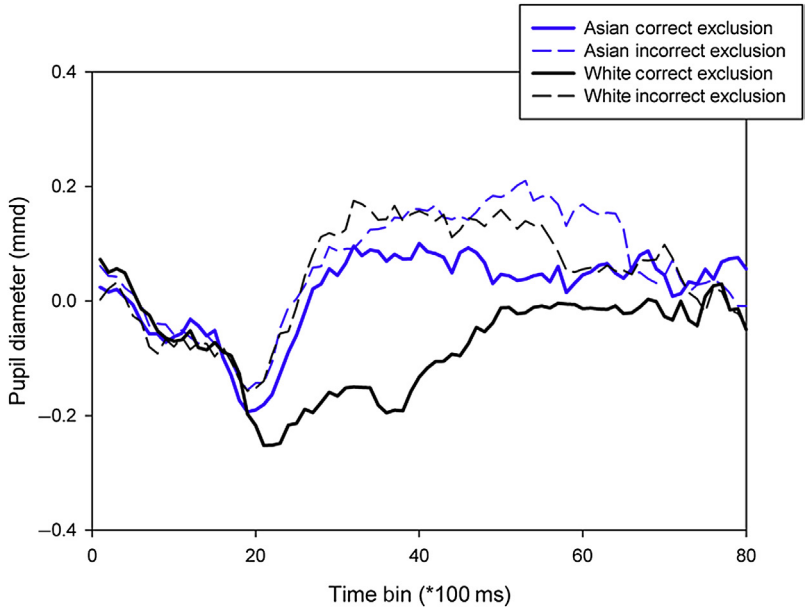


Figure 5.6 Baseline-corrected pupil diameters as a function of face race and subsequent memory performance in the exclusion study trials. (For color version of this figure, the reader is referred to the online version of this book.)

these PDP data suggested that some memories are too strong to elicit this standard pattern; when a memory is retrieved easily, as SCAPE would argue, it involves relatively little cognitive effort.

6.2. RK and Pupillometry

To further assess whether recollection is fast and automatic, Papesh and Goldinger (2011) examined face memory using the RK procedure. As in our PDP study, participants ($N = 30$) encoded a series of 28 Asian and White faces (presented without borders) for an upcoming memory test, with standard RK instructions. As in the previous experiments, pupil size during encoding revealed an ORB: participants devoted greater cognitive resources to encoding other-race faces, relative to own-race faces. We first examined encoding activity on the basis of subsequent memory, conditionalized by whether the participants demonstrated an ORB in their recognition data. Approximately half the participants demonstrated the typical ORB. Despite this, both groups of participants showed that encoding other-race faces demanded more cognitive effort (Figure 5.7). Further, the peaks for subsequent "know" responses were reliably higher than the peaks for subsequent "remember" responses. This finding is consistent with the findings from the PDP experiment. Specifically, when participants are going to have a specific, detailed recollection, they devote *less* effort to the study trial. It is almost as if they might think to themselves "*Well, I'm definitely going to know this later*," and they need not try as hard to commit it to memory. Although the results are not shown in Figure 5.7, it is important to note that pupil diameters were also recorded during the recognition test: we observed a

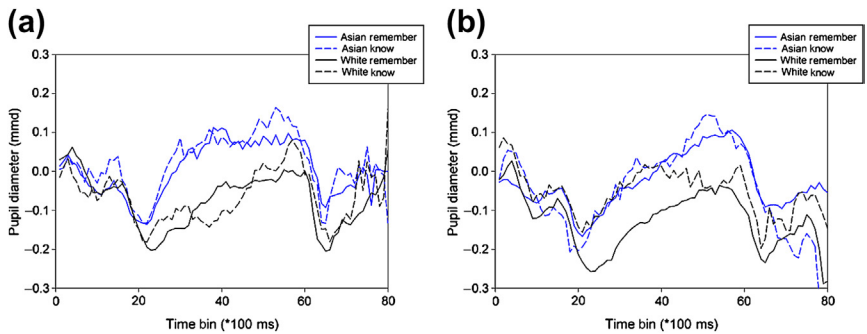


Figure 5.7 Baseline-corrected pupil diameters (during encoding) as a function of face race and subsequent memory performance in the RK experiment. (a) Represents the ORB group and (b) represents the No ORB group. (For color version of this figure, the reader is referred to the online version of this book.)

clear pattern that hits designated as "remember" elicited less pupil dilation than hits designated as "know".

7. GENERAL CONCLUSIONS

In popular conceptualizations of recognition memory, two processes are said to contribute to memorial decisions, recollection and familiarity. All current dual-process models assume that recollection is a slow, deliberate, and demanding cognitive process, whereas familiarity is fast and automatic. Here, we hope to have shown evidence that recollection may not be as slow and cognitively taxing as previously theorized. By examining pupil dilations, as a proxy for cognitive effort, we observe several very clear patterns. During the creation of memories, those items that will eventually lead to the experience of recollection (as determined by the PDP procedure or "R" responses in RK) are characterized by less effort during encoding. In similar fashion, those same items elicit less effort during later recognition testing. Effort during encoding closely predicts effort during test: recollected items have a privileged existence, being easily encoded and easily retrieved.

Indeed, our findings are consistent with evidence that recollection-based responses often occur more quickly than familiarity-based responses (Dewhurst et al., 2006; Duarte et al., 2004; Wheeler & Buckner, 2004). In our research, we use a psychophysiological index of cognitive processing, finding evidence that recollection is fast and relatively automatic. Although pupil dilation is sensitive to the strength and detail of memories (e.g. Otero et al., 2011; Papesh et al., 2012), across two experiments, we have shown that creating and retrieving memories via recollection engenders less pupil dilation, relative to creating and retrieving memories via familiarity. Rather than being slow and deliberate, we suggest that recollection is fast and easy. After all, you do not have to work very hard to recognize your spouse when you get home from work. It might require more effort to recognize your local butcher, if he is there too (or so we assume… it's an empirical question).

REFERENCES

Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., et al. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia*, *12*, 1810–1823. http://dx.doi.org/10.1016/j.neuropsychologia.2005.01.019.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276–292. http://dx.doi.org/10.1037/0033-2909.91.2.276.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.

Boring, E. G. (1929). *A history of experimental psychology*. New York: Century.

Buchanan, T. W., Etzel, J. A., Adolphs, R., & Tranel, D. (2006). The influence of autonomic arousal and semantic relatedness on memory for emotional words. *International Journal of Psychophysiology*, *61*, 26–33. http://dx.doi.org/10.1016/j.ijpsycho.2005.10.022.

Cansino, S., & Trejo-Morales, P. (2008). Neurophysiology of successful encoding and retrieval of source memory. *Cognitive, Affective, and Behavioral Neuroscience*, *8*, 85–98. http://dx.doi.org/10.3758/CABN.8.1.85.

Clark, K. B., Naritoku, D. K., Smith, D. C., Browning, R. A., & Jensen, R. A. (1999). Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nature Neuroscience*, *2*(1), 94–98. http://dx.doi.org/10.1038/4600.

Conway, M. A., Dewhurst, S. A., Pearson, N., & Sapute, A. (2001). The self and recollection reconsidered: how a 'failure to replicate' failed and why trace strength accounts of recollection are untenable. *Applied Cognitive Psychology*, *15*, 673–686. http://dx.doi.org/10.1002/acp.740.

Croiset, G., Nijsen, M. J. M. A., & Kamphuis, P. J. G. H. (2000). Role of corticotropin-releasing factor, vasopressin and the autonomic nervous system in learning and memory. *European Journal of Pharmacology*, *405*, 225–234.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*, 923–938.

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 531–547. http://dx.doi.org/10.1037/0278-7393.21.3.531.

Curran, T., & Hintzman, D. L. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General*, *126*, 228–247. http://dx.doi.org/10.1037/0096-3445.126.3.228.

Dewhurst, S. A., Holmes, S. J., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: remembering is faster than knowing. *Consciousness and Cognition*, *15*, 147–162.

Dietrich, D. E., Emrich, H. M., Waller, C., Wieringa, B. M., Johannes, S., & Münte, T. F. (2000). Emotion/cognition-coupling in word recognition memory of depressive patients: an event-related potential study. *Psychiatry Research*, *96*, 15–29. http://dx.doi.org/10.1016/S0165-1781(00)00187-6.

Dodson, C. S., Holland, P. W., & Shimamura, A. P. (1998). On the recollection of specific- and partial-source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1121–1136. http://dx.doi.org/10.1037/0278-7393.24.5.1121.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*(4), 523–533.

Duarte, A., Ranganath, C., Winward, L., Hayward, D., & Knight, R. T. (2004). Dissociable neural correlates for familiarity and recollection during the encoding and retrieval of pictures. *Cognitive Brain Research*, *18*, 255–272.

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, *111*(2), 524–542. http://dx.doi.org/10.1037/0033-295X.111.2.524.

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological Review*, *115*, 426–446. http://dx.doi.org/10.1037/0033-295X.115.2.426.

Eldridge, L. L., Engel, S. A., Zeineh, M. M., Bookheimer, S. Y., & Knowlton, B. J. (2000). A dissociation of encoding and retrieval processes in the human hippocampus. *Journal of Neuroscience*, *25*, 3280–3286. http://dx.doi.org/10.1523/JNEUROSCI.3420-04.2005.

Engelkamp, J., & Dehn, D. M. (1997). Strategy and consciousness in remembering subject-performed actions. *Sprache & Kognition*, *16*, 94–109.

Eschenko, O., & Sara, S. J. (2008). Learning-dependent, transient increase of activity in noradrenergic neurons of locus coeruleus during slow wave sleep in the rat: brain stem-cortex interplay for memory consolidation? *Cerebral Cortex*, *18*, 2596–2603. http://dx.doi.org/10.1093/cercor/bhn020.

Fontana, F. (1765). *Dei moti dell'iride.* : Lucca: J. Giusti.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*, 309–313.

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition*, *19*, 617–623. http://dx.doi.org/10.3758/BF03197157.

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology*, *50*, 114–122. http://dx.doi.org/10.1037/1196-1961.50.1.114.

Gardiner, J. M., & Parkin, A. J. (1990). Attention and recollective experience in recognition memory. *Memory & Cognition*, *18*, 579–583.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition*, *7*, 285–288. http://dx.doi.org/10.1006/ccog.1997.0321.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: a meta-analysis of remember, know, and guess responses. *Memory*, *10*, 83–98. http://dx.doi.org/10.1080/0965210143000281.

Gardner, R. M., Mo, S. S., & Borrego, R. (1974). Inhibition of pupillary orienting reflex by novelty in conjunction with recognition memory. *Bulletin of the Psychonomic Society*, *3*, 237–238.

Gardner, R. M., Mo, S. S., & Krinsky, R. (1974). Inhibition of pupillary orienting reflex by heteromodal novelty. *Bulletin of the Psychonomic Society*, *4*, 510–512.

Gianaros, P. J., Van Der Veen, F. M., & Jennings, J. R. (2004). Regional cerebral blood flow correlates with heart period and high-frequency heart period variability during working-memory tasks: implications for the cortical and subcortical regulation of cardiac autonomic activity. *Psychophysiology*, *41*(4), 521–530. http://dx.doi.org/10.1111/1469-8986.2004.00179.x.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, and Behavioral Neuroscience*, *10*, 252–269. http://dx.doi.org/10.3758/CABN.10.2.252.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: insights from eye-movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1105–1122. http://dx.doi.org/10.1037/a0016548.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*, 90–95. http://dx.doi.org/10.1177/0963721412436811.

Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, *77*, 340–355. http://dx.doi.org/10.1037/h0032456.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*, 457–461. http://dx.doi.org/10.1111/j.1469-8986.1996.tb01071.x.

Granholm, E., & Steinhauer, S. R. (2004). Introduction: pupillometric measures of cognitive and emotional processing. *International Journal of Psychophysiology*, *52*, 1–6. http://dx.doi.org/10.1016/j.ijpsycho.2003.12.001.

Granholm, E., & Verney, S. P. (2004). Pupillary responses and attentional allocation problems on the backward masking task in schizophrenia. *International Journal of Psychophysiology*, *52*, 37–52. http://dx.doi.org/10.1016/j.ijpsycho.2003.12.004.

Gronlund, S. D., Edwards, M. B., & Ohrt, D. D. (1997). Comparison of the retrieval of item versus spatial position information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1261–1274. http://dx.doi.org/10.1037/0278-7393.23.6.1261.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 846–858. http://dx.doi.org/10.1037/0278-7393.15.5.846.

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 259–278. http://dx.doi.org/10.1037/0278-7393.23.2.259.

Guo, C., Duan, L., Li, W., & Paller, K. A. (2006). Distinguishing source memory and item memory: brain potentials at encoding and retrieval. *Brain Research*, *1118*, 142–154.

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*, 398–405. http://dx.doi.org/10.1080/09658211.2011.575788.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*, 46–54.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192.

Hess, E. H., Seltzer, A. L., & Schlien, J. M. (1965). Pupil response of hetero- and homosexual males to pictures of men and women: a pilot study. *Journal of Abnormal Psychology*, *70*, 165–168.

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L. T. (2002). Regularities of source recognition: ROC analysis. *Journal of Experimental Psychology: General*, *131*, 494–510.

Hintzman, D. L., & Caulton, D. A. (1997). Recognition memory and modality judgments: a comparison of retrieval dynamics. *Journal of Memory and Language*, *37*, 1–23. http://dx.doi.org/10.1006/jmla.1997.2511.

Hintzman, D. L., Caulton, D. A., & Levitin, D. J. (1998). Retrieval dynamics in recognition and list discrimination: further evidence of separate processes of familiarity and recall. *Memory & Cognition*, *26*, 449–462.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18. http://dx.doi.org/10.1006/jmla.1994.1001.

Hirshman, E. (1998). On the logic of testing the independence assumption in the process-dissociation procedure. *Memory & Cognition*, *26*, 857–859. http://dx.doi.org/10.3758/BF03201168.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: reflections on the remember-paradigm. *Memory & Cognition*, *25*, 345–351. http://dx.doi.org/10.3758/BF03211290.

Holdstock, J. S., Mayes, A. R., Gong, Q.Y., Roberts, N., & Kapur, N. (2005). Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. *Hippocampus*, *15*, 203–215. http://dx.doi.org/10.1002/hipo.20046.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541. http://dx.doi.org/10.1016/0749-596X(91)90025-F.

Jacoby, L. L. (1998). Invariance in automatic influences of memory: toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 3–26. http://dx.doi.org/10.1037/0278-7393.24.1.3.

Jacoby, L. L. (1999). Ironic effects of repetition: measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 3–22. http://dx.doi.org/10.1037/0278-7393.25.1.3.

Jacoby, L. L., Begg, I. M., & Toth, J. P. (1997). In defense of functional independence: violations of assumptions underlying the process-dissociation procedure? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 484–495. http://dx.doi.org/10.1037/0278-7393.23.2.484.

Jacoby, L. L., & Rhodes, M. G. (2006). False remembering in the aged. *Current Directions in Psychological Science*, *15*, 49–53. http://dx.doi.org/10.1111/j.0963-7214.2006.00405.x.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: a declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness. Carnegie Mellon symposia on cognition* (pp. 13–47). Hillsdale, NJ: Erlbaum.

Janisse, M. P. (1974). Pupil size, affect and exposure frequency. *Social Behavior and Personality*, *2*, 125–146. http://dx.doi.org/10.2224/sbp.1974.2 2.125.

Johansson, M., Mecklinger, A., & Treese, A. (2004). Recognition memory for emotional and neutral faces: an event-related potential study. *Journal of Cognitive Neuroscience*, *16*, 1840–1853. http://dx.doi.org/10.1162/0898929042947883.

Joordens, S., & Merikle, P. M. (1993). Independence or redundancy? Two models of conscious and unconscious influences. *Journal of Experimental Psychology: General*, *122*, 462–467. http://dx.doi.org/10.1037/0096-3445.122.4.462.

Kafkas, A., & Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *Quarterly Journal of Experimental Psychology*, *64*, 1971–1989. http://dx.doi.org/10.1080/17470218.2011.588335.

Kahneman, D. (1973). *Attention and effort*. New York: Prentice Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585. http://dx.doi.org/10.1126/science.154.3756.1583.

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual deficit during a mental task. *Science*, *157*, 218–219. http://dx.doi.org/10.1126/science.157.3785.218.

Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology*, *41*, 175–185. http://dx.doi.org/10.1111/j.1469-8986.2004.00147.x.

Kirwan, C. B., Wixted, J. T., & Squire, L. R. (2008). Activity in the medial temporal lobe predicts memory strength, whereas activity in the prefrontal cortex predicts recollection. *Journal of Neuroscience*, *28*, 10541–10548. http://dx.doi.org/10.1523/JNEUROSCI.3456-08.2008.

Klimesch, W., Doppelmayr, M., Yonelinas, A., Kroll, N. E.A., Lazzara, M., Röhm, D., et al. (2001). Theta synchronization during episodic retrieval: neural correlates of conscious awareness. *Cognitive Brain Research*, *12*, 33–38. http://dx.doi.org/10.1016/S0926-6410(01)00024-6.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: two difference expression of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 699–710. http://dx.doi.org/10.1037//0278-7393.21.3.699.

Koss, M. (1986). Pupillary dilation as an index of central nervous system α2-adrenoceptor activation. *Journal of Pharmacology Methods*, *15*, 1–19.

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*, 132–140.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: a window to the preconscious? *Perspectives on Psychological Science*, *7*, 18–27. http://dx.doi.org/10.1177/1745691611427305.

Loewenfeld, I. E. (1958). Mechanisms of reflex dilation of the pupil: historical review and experimental analysis. *Documenta Ophthalmologica*, *12*, 185–448.

Mandler, G. (1980). Recognizing: the judgment of previous occurrence. *Psychological Review*, *87*, 252–271. http://dx.doi.org/10.1037/0033-295X.87.3.252.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: a time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 563–582. http://dx.doi.org/10.1037//0278-7393.25.3.563.

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process: implication for dual-process theories of recognition memory. *Psychological Science*, *20*, 509–515. http://dx.doi.org/10.1111/j.1467-9280.2009.02324.x.

Moscovitch, M., & McAndrews, M. P. (2002). Material-specific deficits in "remembering" in patients with unilateral temporal lobe epilepsy and excisions. *Neuropsychologia*, *40*, 1335–1342. doi:PII S0028-3932(01)00213-5.

Mulligan, N. W., & Hirshman, E. (1997). Measure the bases of recognition memory: an investigation of the process-dissociation framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 280–304. http://dx.doi.org/10.1037/0278-7393.23.2.280.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology*, *48*, 1346–1353. http://dx.doi.org/10.1111/j.1469-8986.2011.01217.x.

Papesh, M. H. (2012). *Source memory revealed through eye movements and pupil dilation*. Doctoral dissertation, Arizona State University.

Papesh, M. H., & Goldinger, S. D. (2011). Your effort is showing! Pupil dilation reveals memory heuristics. In P. Higham & J. Leboe (Eds.), *Constructions of remembering and metacognition* (pp. 215–224). : Palgrave Macmillan.

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-blah-metry: cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics*, *74*, 754–765. http://dx.doi.org/10.3758/s13414-011-0263-y.

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*, 56–64. http://dx.doi.org/10.1016/j.ijpsycho.2011.10.002.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-computer Studies*, *59*, 185–198. http://dx.doi.org/10.1016/S1071-5819(03)00017-X.

Porter, G., & Troscianko, T. (2003). Pupillary response to grating stimuli. *Perception*, *32*, 156.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: insights from pupillometry. *Quarterly Journal of Experimental Psychology*, *60*, 211–229.

Rajaram, S. (1993). Remembering and knowing: two means of access to the personal past. *Memory & Cognition*, *21*, 89–102.

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1993). Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society of Neuroscience Abstracts*, *19*, 974.

Rajkowski, J., Majczynski, H., Clayton, E., & Aston-Jones, G. (2004). Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *Journal of Neurophysiology*, *92*, 361–371.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. D. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, *42*, 2–13. http://dx.doi.org/10.1016/j.neuropsychologia.2003.07.006.

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535. http://dx.doi.org/10.1037/0033-295X.99.3.518.

Richardson-Klavehn, A., Gardiner, J. M., & Java, R. I. (1996). Memory: task dissociations, process dissociations, and dissociations of awareness. In G. Underwood (Ed.), *Implicit cognition* (pp. 85–158). Oxford: Oxford University Press.

Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember-know paradigm. *Psychonomic Bulletin & Review*, *15*, 825–832. http://dx.doi.org/10.3758/PBR.15.4.825.

Sauvage, M. M., Fortin, N. J., Owens, C. B., Yonelinas, A. P., & Eichenbaum, H. (2008). Recognition memory: opposite effects of hippocampal damage on recollection and familiarity. *Nature Neuroscience*, *11*, 16–18.

Squire, L. R., & Wixted, J. T. (2011). The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience, 34,* 259–288.

Stanners, R. F., Coulter, M., Sweet, A. W., & Murphy, P. (1979). The pupillary response as an indicator of arousal and cognition. *Motivation and Emotion*, *3*, 319–340. http://dx.doi.org/10.1007/BF00994048.

Steinhauer, S. R., & Hakerem, G. (1992). The pupillary response in cognitive psychophysiology and schizophrenia. *Annals of the New York Academy of Sciences*, *658*, 182–204.

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervations of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, *52*, 77–86. http://dx.doi.org/10.1016/j.ijpsycho.2003.12.005.

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., et al. (2006). The locus ceruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience*, *26*, 7416–7423. http://dx.doi.org/10.1523/JNEUROSCI.1001-06.2006.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist*, *25*, 1–12.

Uncapher, M. R., & Rugg, M. D. (2005). Effects of divided attention on fMRI correlates of memory encoding. *Journal of Cognitive Neuroscience*, *17*, 1923–1935. http://dx.doi.org/10.1162/089892905775008616.

Van Gerven, P. W.M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*, 167–174. http://dx.doi.org/10.1111/j.1469-8986.2003.00148.x.

Verfaellie, M., Rajaram, S., Fossum, K., & Williams, L. (2008). Not all repetition is alike: different benefits of repetition in amnesia and normal memory. *Journal of the International Neuropsychological Society*, *14*, 365–372.

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., et al. (2008). The coupling of emotion and cognition in the eye: introducing the pupil old/new effect. *Psychophysiology*, *45*, 130–140. http://dx.doi.org/10.1111/j.1469-8986.2008.00745.x.

Wais, P. E. (2008). fMRI signals associated with memory strength in the medial temporal lobes: a meta-analysis. *Neuropsychologia*, *46*, 3185–3196. http://dx.doi.org/10.1016/j.neuropsychologia.2008.08.025.

Wais, P. E., Squire, L. R., & Wixted, J. T. (2009). In search of recollection and familiarity signals in the hippocampus. *Journal of Cognitive Neuroscience*, *22*, 109–123. http://dx.doi.org/10.1162/jocn.2009.21190.

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron*, *49*, 459–466. http://dx.doi.org/10.1016/j.neuron.2005.12.020.

Wheeler, M. E., & Buckner, R. L. (2004). Functional-anatomic correlates of remembering and knowing. *NeuroImage*, *21*, 1337–1349.

Whittlesea, B. W. A. (1997). Production, evaluation and preservation of experiences: constructive processing in remembering and performance tasks. In D. Medin (Ed.), *The psychology of learning and motivation*. (Vol. 37). New York: Academic Press.

Whittlesea, B. W. A., & Leboe, J. (2000). The heuristic basis of remembering and classification: fluency, generation and resemblance. *Journal of Experimental Psychology: General*, *129*, 84–106.

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *ACTA Psychologica*, *98*, 141–165.

Whittlesea, B. W. A., & Williams, L. D. (2001). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 3–13.

Wixted, J. T. (2007). Dual-process theory and signal detection theory of recognition memory. *Psychological Review*, *114*, 152–176. http://dx.doi.org/10.1037/0033-295X.114.1.152.

Wixted, J. T. (2009). Remember/know judgments in cognitive neuroscience: an illustration of the underrepresented point of view. *Learning & Memory*, *16*, 406–412. http://dx.doi.org/10.1101/lm.1312809.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*, 1025–1054. http://dx.doi.org/10.1037/a0030874.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*, 616–641.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1341–1354. http://dx.doi.org/10.1037//0278-7393.20.6.1341.

Yonelinas, A. P. (2001). Components of episodic memory: the contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *356*, 1363–1374.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. http://dx.doi.org/10.1006/jmla.2002.2864.

Yonelinas, A. P., & Jacoby, L. L. (1996). Response bias and the process-dissociation procedure. *Journal of Experimental Psychology: General*, *125*, 422–434. http://dx.doi.org/10.1037/0096-3445.125.4.422.

Yonelinas, A. P., Kroll, N. E. A., Quamme, J. R., Lazzara, M. M., Sauve, M. J., Widaman, K. F., et al. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, *5*, 1236–1241. http://dx.doi.org/10.1038/nn961.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, *133*, 800–832. http://dx.doi.org/10.1037/0033-2909.133.5.800.

CHAPTER SIX

A Mechanistic Approach to Individual Differences in Spatial Learning, Memory, and Navigation

Amy L. Shelton[1], Steven A. Marchette, Andrew J. Furman
Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA
[1]Corresponding author: E-mail: ashelton@jhu.edu

Contents

Abstract

Navigation is a complex task that depends on the processes of perception, learning, memory, and reasoning to be successful. Given this complexity, it is not surprising that humans (and other species) vary dramatically in their approach and success at navigation. This wide range of abilities has been of great interest to the field of human spatial cognition. In addition, spatial navigation is a cross-species phenomenon that can speak

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00006-X

to a variety of learning and memory processes. Therefore, understanding individual differences in this domain can offer a wide range of insights that affect many behaviors in the real world. A cognitive framework that gives precedent to the flexible use of spatial information and explicit or declarative learning processes has driven much of the work on individual differences in navigation in humans. However, animal models of basic learning mechanisms may also offer substantial insight into individual differences in both how well people navigate their surroundings and in the strategies or styles that they bring to bear on the navigational problems. This mechanistic approach may offer a stronger foundation for not only how individual differences might emerge but also how they interact with differences in the environments and goals that drive our need to learn, remember, and navigate in the world.

1. INTRODUCTION

Whether considering how ancestral humans learned to hunt for food and shelter in the wilderness or the more mundane issues of how you got to work today, there is little question that navigation is an important aspect of the human experience. Many of our everyday tasks—finding the car in a parking lot, getting to the checkout line at the grocery store—involve spatial navigation. As humans, we accomplish this navigation in various ways, and everyone has an opinion about his or her own navigational abilities. Ask any random group of people, and you will get answers ranging from "I am great with directions" to "I get lost all the time". This massive range of individual differences in perceived spatial prowess can also be seen in behavior. Most people can identify those friends or family members who they prefer to ride with on trips or those who give good or bad directions. We also know people who are dependent on their Global Positioning System (GPS) units for even the most basic of travel tasks. This variability in navigational ability likely reflects the fact that navigation is a complex task that requires individuals to encode the immediate surroundings, interrogate memory (working memory and/or long-term memory), determine how to proceed, and execute the relevant movements through space. As such, understanding individual differences in navigation requires a comprehensive approach to understanding how individuals might differ in the way they learn and remember information about the spatial organization of their environments.

Individual differences in spatial learning and navigation have enjoyed a long history in the literature. Here, we review that literature in light of the extensive work on nonhuman animal models of spatial learning mechanisms. In particular, we explore how using these learning mechanisms to

ground individual differences in human spatial learning and navigation might offer a framework for understanding how the wide range of differences emerge. In addition, navigational tasks have been used to understand more general processes of learning and memory in human and nonhuman animals (for brief review, see Bird & Burgess, 2008). In humans, much of the focus has been on the role of explicit or declarative kinds of processes, consistent with the description of navigation as a cognitively demanding task of perceiving, planning, and executing. By couching learning and navigational differences in basic learning mechanisms, we can more readily investigate the contributions of other implicit or habit-based mechanisms to complex behaviors. Moreover, this will allow extension to other sources of variability such as aging and hormones. Taken together with the extant literature, this framework has substantial potential to enrich a wide range of domains.

The following sections provide a context for thinking about individual differences in spatial learning and navigation beginning with ideas about what it means to measure navigational ability. Subsequent sections describe the dominant framework for thinking about individual differences in strategies and elaborate on a neurobiological framework coming out of rodent learning that might offer a broader and more comprehensive approach. The final sections highlight the advantages of this new framework for understanding many sources of variability.

2. WHAT DOES IT MEAN TO MEASURE SPATIAL LEARNING AND NAVIGATIONAL ABILITY?

Despite the remarkable variability in reports of navigational ability, humans (and other animals) are quite successful on the daily task of navigating. That is, few of us fail to reach our office or make it home from the grocery store. What, then, characterizes individual differences in this ability? In this section, we discuss the two broad facets of performance that comprise individual differences, success rate and strategies/style.

2.1. Differences in Navigational Success Rates

The first dimension on which we tend to classify people is on the dimension of success—or good/bad navigation ability. Interestingly, when we consider whether people are good or bad navigators, we often emphasize how well they perform when faced with a new learning or a novel navigational situation. For example, if we are traveling with friends and

want to figure out how to get to a new restaurant, we often appeal to the person or people in the group who we think will be most likely to find this new location. Similarly, we consider who is going to be best at giving directions. This anecdotal emphasis on these novel or flexible uses of spatial information suggests that good/bad *navigators* may actually be good/bad *explorers*.

This emphasis on new learning and flexibility carries through to empirical work on individual differences in navigational abilities. In most experimental studies of spatial memory and navigation, participants learn novel environments (for exceptions, see Marchette, Yerramsetti, Burns, & Shelton, 2011; Yerramsetti, Marchette, & Shelton, 2012). Individual differences are largely characterized in terms of differences in accuracy and success rate at navigation to make inferences about the quality of spatial learning and the factors associated with those abilities (e.g. Allen, Kirasic, Dobson, Long, & Beck, 1996; Fields & Shelton, 2006; Ishikawa & Montello, 2006). In many cases, the assessment emphasizes the explicit knowledge of spatial relationships. For example, Ishikawa and Montello (2006) used a test of straight-line distances to assess how well people learned under various conditions. Similarly, Fields and Shelton used judgments of relative direction after different encoding conditions. In both these cases, the measure of learning/navigation was how accurately people judged the distances or directions. Critically, these tests require that a person understand the global structure and use the space flexibly. This emphasis on flexibility in learning is also reflected in the Santa Barbara Sense of Direction questionnaire (SBSOD; Hegarty, Richardson, Montello, Lovelace, & Subbiah, 2002), one of the most commonly used measures of spatial navigational ability. In this self-report measure, participants report on their general ability and comfort level with different kinds of spatial scenarios. Like many of the measures, the SBSOD emphasizes the ability to use spatial information flexibly, marking flexibility in perceptual and cognitive processing as a primary factor in navigational success (see Wolbers & Hegarty, 2010).

2.2. Differences in Strategies/Styles: Landmark–Route–Survey Framework

Whether couched in terms of successful navigation or successful exploration, success rate is not the only dimension along which humans differ in navigation. There is also substantial anecdotal and self-report data to support the claim that humans differ dramatically in the *way* they successfully navigate. For example, imagine asking a room full of healthy adults how

they navigate to work each day. Some individuals will report using the same set route over and over again, whereas others will report navigating by just keeping a desired direction in mind. Still others might report changing their route as a function of time of day, observed patterns of traffic, or even weather conditions. Any comprehensive understanding of individual differences will require an appreciation of these different styles or strategies for navigating and how those different approaches might be more or less effective for different individuals and under different navigational conditions.

Like success rate, navigational styles and strategies have a history in the research literature. A common framework for thinking about such individual differences has stemmed from ideas about the progression of knowledge (Siegel & White, 1975). In this popular framework, one first learns the landmarks in an environment. Landmarks are defined as features or objects that can demarcate locations or directions. As landmarks become linked, one develops route knowledge, which is generally defined as the knowledge of a set of ordered paths. Finally, this information begins to form a representation of the global context as survey knowledge, generally defined as knowledge of the organization of locations in a more global framework. In its strongest instantiation, this framework suggests that learning an environment goes through each of these stages, each building upon the previously acquired knowledge, with survey knowledge as the ultimate representation of an environment.

The landmark–route–survey framework has influenced the direction of a wide range of studies in spatial cognition, including approaches to individual differences. For example, one tool used to assess spatial navigational strategies is the Questionnaire on Spatial Representation (QSR; Pazzaglia & De Beni, 2001). In this self-report survey, questions are directed at assessing the degree to which people use different types of information, and scoring allows assessment of landmark use, route-based strategies, and survey-based strategies. Similar efforts to classify navigational behaviors have taken a similar approach, with a strong emphasis on the route-based and survey-based distinctions (e.g. Lawton, 1994, 1996).

2.3. Challenges to the Landmark–Route–Survey Framework

This classification scheme has been useful in advancing the ideas that navigational ability includes both success rate and strategic differences, but grounding individual differences in a stage-based theory promotes the idea that some forms of knowledge or types of strategies are inherently better

than others. Most notably, this framework has given top status to survey knowledge or the use of spatial information in a flexible manner based on knowledge of the global organization. This is reflected in many empirical studies that assess ability based on flexibility and use of detours. However, there are important counterpoints to consider when considering how to classify navigational style and its relationship to ability.

The first counterpoint is the evidence challenging the successive nature of the stages of knowledge. Although there is little doubt that the distinction between route-based and survey-based information is important to understanding the kinds of strategies people use and the cues that might be involved in spatial learning, it is not clear that the progression from one type of knowledge to another is a critical feature (e.g. Ishikawa & Montello, 2006; Taylor, Naylor, & Chechile, 1999). For example, Taylor, Naylor, and Chechile gave people specific goals that emphasized the need for either route-based knowledge or survey-based knowledge. They then gave tests that tapped each type of knowledge. By the stage theory, individuals who show survey-based knowledge should also have the route-based knowledge from which the survey knowledge was built. By contrast, their results showed that the specific goals drove the type of representation such that people who developed survey-based representations did not necessarily develop the route-based representations as well. This suggests that different types of representations can be built up somewhat independently (or at least without a strict dependence on each other's presence).

Second, the sense that survey-based strategies convey a clear advantage over other strategies does not account for the potential interaction between the navigational situation and strategies. Emerging frameworks have started to acknowledge that the success of navigational wayfinding likely depends on the combination of strategies (e.g. Ishikawa & Montello, 2006) and/or the interplay of skills or strategies and the specific environment or navigational challenges (e.g. Carlson, Hölscher, Shipley, & Dalton, 2010; Wolbers & Hegarty, 2010). For example, Carlson et al. put forth a conceptual framework for indoor navigation where navigational success depends on the combination of the building features, the cognitive representation of the space that has been formed, and the strategies and skills that are brought to bear. The latter distinction between the mental representation and the strategies and skills suggests that there is not a one-to-one link between a particular navigational strategy and the underlying representation. This opens the door for thinking about strategies (or styles of navigation) as a more flexible tool that interacts with the information that is represented.

Finally, both anecdotal and empirical evidence suggests that the route-based and survey-based dichotomy (or landmark–route–survey trichotomy) may be insufficient to account for the variety of behaviors people actually utilize in navigation. There have been recent proposals for additional categories of processing such as "graph knowledge" (for review, see Chrastil, 2012). However, even within a category, there may be additional distinctions that need to be considered (e.g. Allen et al., 1996; Chan, Baumann, Bellgrove, & Mattingly, 2012; Waller & Lippa, 2007). For example, Allen et al. tested environmental learning with tasks that tapped both the forward and the backward sequences of the routes. Participants could vary in both of these depending on the degree to which they had strong explicit knowledge of the route or more associative knowledge of the sequence in a stimulus–response fashion. Similarly, people who report using landmarks could be using them as cues to a specific action (e.g. turn left at the church) or simply as beacons detached from the specific action required (e.g. look for the church and head in that direction). These two cases would likely result in similar classification as some combination of landmark and route-based strategies, but they are serving different purposes and require different levels of explicit knowledge in the act of navigating. Waller and Lippa demonstrated how differences in performance depended on which way the landmarks were used, suggesting that these distinctions may be important for understanding what people can or cannot do. This represents just one case where the landmark–route–survey framework may be incomplete. Once we break away from this framework, the potential to identify and characterize these additional navigational strategies will enhance our appreciation for individual differences.

3. DUAL SYSTEMS FOR SPATIAL LEARNING IN RODENTS

One framework that has been emerging as a major candidate for explaining aspects of human navigational behavior is based on the dual-system model in rodents. This model stems from several decades of work on the learning mechanisms or systems used by rodents to learn an environment. In the basic navigational paradigm, rodents are placed in a simple cross maze with one arm blocked (i.e. dual-solution T-maze; Figure 6.1(A)). The rat is placed in the same start arm relative to both the environment and the location of a food reward over many trials. After learning, the rat is given a probe trial on which the start location is rotated with

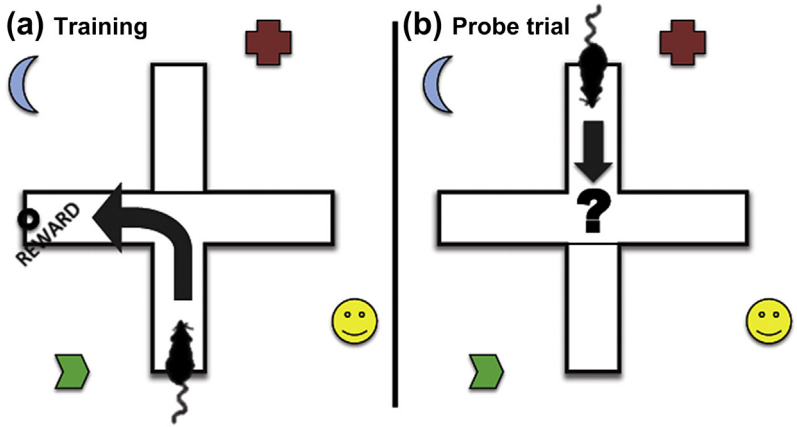


Figure 6.1 Diagram of the dual-solution T-maze used to explore place and response learning in rodents (e.g. Packard & McGaugh, 1996; Restle, 1957; Tolman et al., 1947). (A) Encoding environment with stable cues and path; (B) probe trial from novel start arm. Shapes represent cues visible in the surrounding environment. (For color version of this figure, the reader is referred to the online version of this book.)

respect to the cues in the larger environment (Figure 6.1(B)). On probe trials, rats can exhibit one of two behaviors. If the rat turns down the arm that is the same location as the reward relative to the cues in the environment (right turn in Figure 6.1(B)), this is taken as evidence of *place learning*. Alternatively, if the rat makes the same physical turn as during the learning phase (left turn in Figure 6.1(B)), this is taken as evidence of *response learning*. This distinction is intended to reflect the fact that during repeated learning in a stable environment, the rat has the potential to learn the location in (at least two ways), relative to the environmental cues or as a series of required responses (e.g. Restle, 1957; Tolman, 1948; Tolman, Ritchie, & Kalish, 1947).

Place and response learning have been distinguished in the rodent based on a number of different properties in both the dual-solution T-maze and other preparations. For example, place learning tends to emerge earlier in learning before sufficient repetition has occurred to establish the response (e.g. Packard & McGaugh, 1996). Similarly, conditions that require place learning tend to be more rapidly learned than conditions that require response learning (e.g. Ritchie, Aeschlimm, & Pierce, 1950; Tolman, Ritchie, & Kalish, 1946). Place learning has also been viewed as more cognitively demanding than response learning as rats placed under conditions of distraction or stressors tend to show more response-learning behaviors (e.g. Elliot & Packard, 2008; Schwabe, Schächinger, de Kloe, & Oitzl, 2010).

In addition to behavioral differences, these two learning mechanisms have been associated with different underlying brain systems in the rat. Using temporary deactivation of brain regions via lidocaine injection, Packard and McGaugh (1996) had rats learning a T-maze preparation over time. In different groups of rats, they temporarily deactivated either the caudate nucleus or the hippocampus. These deactivations had both temporally and behaviorally specific results. In control rats (saline administration), place-learning behavior (using the environmental cues) was observed in earlier training blocks and response-learning behavior (engaging the same left or right turn) became more prevalent in later training blocks. This pattern was unaffected by the site or timing of saline injection. By contrast, lidocaine injections to the hippocampus had pronounced effects on performance early in training with a reduction in place-learning behaviors, whereas lidocaine injections to the caudate nucleus affected later performance with a reduction in response-learning behaviors. This and subsequent work has established the hippocampus and caudate nucleus as markers for the place- and response-learning systems, respectively, in the rodent brain.

Notably, the results not only supported the use of two systems but also suggested that they are simultaneously acquired. In particular, rats that received lidocaine injections to the caudate late in learning not only showed a reduction in their ability to manifest response-learning behaviors but also shifted to place-learning behaviors. This unveiling of place-learning behavior suggests that those animals had the necessary place-learned knowledge to compensate when the response-learning system was deactivated. That is, although their typical behavior was the expression of response learning, the capacity to show the place-learning behavior was intact. There was no symmetrical boost in response-learning behavior during early inactivation of the hippocampus because there had presumably been insufficient repetition to acquire the response. Taken together with the behavioral characterization, this work offers both behavioral and neurobiological markers of place and response learning that can be useful for thinking about how humans develop their capacity to remember and navigate in the world.

An added advantage to incorporating nonhuman animal models is the potential to address a wider range of factors. The dual-system model has been used to investigate the influences of hormones, neurotransmitter systems, and aging on rodent learning and memory. Ultimately, the use of this framework as a foundation for individual differences will invite these critical extensions as well. We return to these ideas in the latter sections. First, we consider how place and response learning have been characterized in humans.

4. PLACE AND RESPONSE LEARNING IN HUMANS

The growing popularity of both desktop and immersive virtual reality has broadened the range of tools and types of questions about spatial learning and navigation that can be addressed in humans. One important consequence of these new tools has been the motivation to adapt rodent paradigms for use in humans. Although this adaptation comes with a variety of challenges (Box 6.1), there have been variations on most of the standard paradigms. Studies using variations on the virtual water maze (e.g. Cornwell, Johnson, Holroyd, Carver, & Grillon, 2008; Hamilton, Driscoll, & Sutherland, 2002; Newman & Kraszniak, 2000) and the virtual radial arm maze (Bohbot, Iaria, & Petrides, 2004; Iaria, Petrides, Dagher, Pike, & Bohbot, 2003) employ close approximations of the rodent preparations, whereas other more ecological tasks such as virtual towns (e.g. Hartley, Maguire, Spiers, & Burgess, 2003), mazes (e.g. Brown, Ross, Tobyne, & Stern, 2012), and dual-solution mazes (e.g. Marchette, Bakker, & Shelton, 2011) offer more conceptual adaptations. Together, these tools have begun to offer both general distinctions in human navigational behavior and insights into the potential mechanisms of individual differences.

4.1. Virtual Water Maze

The Morris water maze (MWM; Morris, 1981, 1984) is a circular pool of cloudy water with a hidden platform at a goal location (Figure 6.2). Rats are placed in the maze and must swim to the platform. Early in learning, rats must explore the environment to locate the platform, but most animals become quite adept at finding the platform. Manipulations of the environment have demonstrated that rats tend to use stable environmental cues to guide navigation (place learning), and performance on the mnemonic aspects of this task breaks down with damage to the hippocampus (e.g. Morris, 1984; Morris, Garrud, Rawlins, & O'Keefe, 1982; Stewart & Morris, 1993). As such, the MWM has been used in a variety of rodent studies as a classic test of spatial memory associated with the hippocampal place-learning system.

Variations of the MWM for humans (Cornwell et al., 2008; Driscoll, Hamilton, Yeo, Brooks, & Sutherland, 2005; Hamilton et al., 2002; Newman & Kaszniak, 2000; Skelton, Ross, Nerad, & Livingstone, 2006) typically involve placing individuals inside a virtual drum or multisided room (for an actual physical hidden sensor version, see Bohbot et al., 1998). As in the

Box 6.1 Challenges to Adapting Rodent Paradigms for Humans

One key to making connections between human navigation and the dual-system model of place and response learning in rodents has been the adaptation of the rodent paradigms for use in humans. This adaptation brings with it some obvious and not-so-obvious challenges. Here, we summarize some of those challenges and the key questions that one might need to address in developing new paradigms.

Scalability

Virtual reality has made it easy to create environments of nearly any size and shape, but one can ask whether a physically scalable environment is still appropriate for asking the same questions in humans. The virtual radial arm maze (e.g. Bohbot et al., 2004; Iaria et al., 2003) may represent the most direct analog to a rodent paradigm in terms of scalability because both the human versions can essentially replicate the physical structure of the rodent mazes. The virtual variants on the MWM for humans (Cornwell et al., 2008; Hamilton et al., 2002; Newman & Kaszniak, 2000; Skelton et al., 2006) are often similar in structure to the rodent paradigm, in that they tend to use circular (or multisided) drums scaled up for humans; however, humans are not dropped into murky water to hide the locations and instead navigate in "dry" versions, which may have additional implications for other factors noted below.

In contrast to the radial arm maze and MWM, some rodent paradigms are physically scalable in virtual space but likely operate very differently for humans. For example, the dual-solution T-maze has been a long-standing tool for eliciting place- and response-learned behaviors in rodents (e.g. Packard & McGaugh, 1996; Restle, 1957; Tolman, 1948), but they are not particularly useful in (adult) humans because humans reduce them to single-trial learning. In pilot work, healthy adult humans very quickly learned that they had two things to remember and often asked after just one training trial whether they were supposed to just make a specific turn or go toward a specific arm in the room. As such, these mazes did not offer us the same window into place and response systems in humans that they did in rodents.

Motivation

A second important issue is to consider whether the motivation for navigation matters. For example, when a rat is dropped into the MWM, it is arguably swimming for survival, whereas a human placed in a virtual water maze has little or no threat to survival if he/she decides to stand still. At worst, the human participant faces boredom (or perhaps reduced compensation in incentive-based paradigms), and it would be unethical to introduce truly comparable manipulations. The cases that use reward in rodents may allow for more comparable motivational states if "getting out", and/or monetary compensation in humans is sufficiently rewarding. However, given that navigation and memory systems are affected by such things as stress (e.g. Kim & Yoon, 1998), it is important to at least consider how the motivational factors contribute.

(Continued)

Box 6.1 Challenges to Adapting Rodent Paradigms for Humans—(cont'd)

Comparable Ecological Validity

Finally, it is useful to consider whether the ecological validity is similar across species. For both rats and humans, the water maze-type tasks may be equally unusual navigational situations, but what about T-mazes and radial arm mazes? Rats outside the lab tend to gravitate to tight places—underground, in walls and foundations, etc., whereas humans largely evolved in larger open spaces and have extensive experience in highly structure buildings and rich developed environments. One could imagine that something like the radial arm maze is more "natural" for rats than for humans. Critically, the ecological validity in general is important for making claims about real-world navigational ability, but it is equally important to make sure that the ecological validity is at least comparable in the rodents and humans if the goal is to allow for cross-species insights. Although this is not evidence based, it represents an issue worth considering when adapting paradigms across species and understanding the limitations of their interpretation. Given these challenges, paradigm development should begin with careful consideration of what conceptual elements one is trying to capture to offer the strongest analogs between humans and rodents. For example, when creating an analog to the dual-solution T-maze (see Marchette, Bakker, et al., 2011), it became abundantly clear that we needed an alternative environment, and we considered what elements of the paradigm were critical. We identified several, including (1) encoding conditions had to afford learning a stable environment and learning a repeated behavior; (2) retrieval had to offer a way of distinguishing which sources of learning was driving behavior; and (3) the retrieval had to be sufficiently ambiguous so that individuals would be free to use either behavior. We then considered additional elements such as whether to use incentives for successful completions, whether to impose stressors or time limits, or whether the stability of the environment had to be based solely on extra-maze distal cues. The intended result was a paradigm that looked very different on the surface but (we hope) shared a deeper conceptual structure with its rodent analog. Moreover, this careful analysis also offered a large number of potential manipulations that could be useful directions for future research.

rodent MWM, there is a predefined goal location that is typically hidden from view (or shares the same visual features as other items in the environment). Humans are given repeated opportunities to learn the location of the hidden goal using the information in and around the virtual maze. Much of this work has shown that, as in rodents, performance depends on the integrity of the hippocampus and surrounding medial temporal lobe (e.g. Astur, Taylor, Mamelak, Philpott, & Sutherland, 2002; Bohbot et al., 1998), supporting the parallel between the rodent and human work.

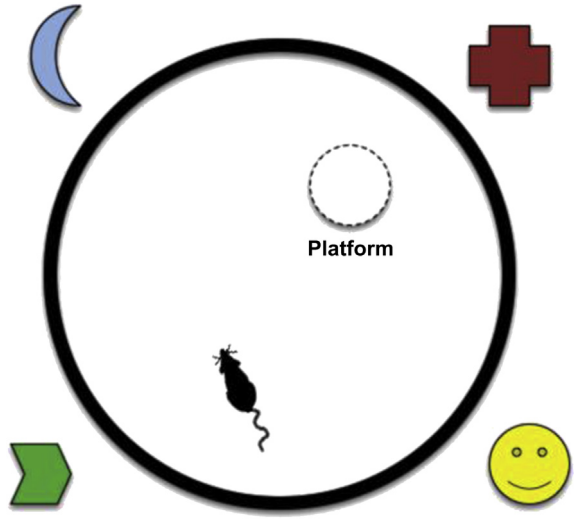


Figure 6.2 Diagram of a typical water maze preparation (e.g. Morris, 1981). Shapes represent cues visible in the surrounding environment or on the walls of the drum. The platform is not visible to the animal during testing. (For color version of this figure, the reader is referred to the online version of this book.)

4.2. Human Radial Arm Mazes

Radial arm mazes in rodents (e.g. Olton, Collison, & Werz, 1977) require rats to explore a series of visually identical arms arranged in around a central start location (Figure 6.3). In the simplest version, each arm is baited with a reward, and the rat is required to remember which arms he has visited. Returning to a previously visited arm is recorded as an error, and one can infer how the rat is using various cues based on the pattern of navigation and errors. The human virtual radial arm maze has been used in healthy controls and patients with medial temporal lobe damage to investigate differences between the use of "spatial" strategies that depended on flexible understanding of the spatial relations akin to place learning and "nonspatial" strategies that depend on more associative behavioral responses (e.g. Bohbot et al., 2004; Iaria et al., 2003). Iaria et al. observed that healthy individuals were evenly split on their spontaneous use of the spatial (place-like) and nonspatial (response-like) strategies early in the experiment. Over time, many of those who began with a spatial strategy shifted to the nonspatial strategy after many repetitions. Bohbot et al. used this same preparation to demonstrate that activation in the hippocampus and caudate was associated with the spatial and nonspatial strategies, respectively, suggesting that the putative place and response systems were supporting the different strategies, in a manner consistent with what has been observed in rats.

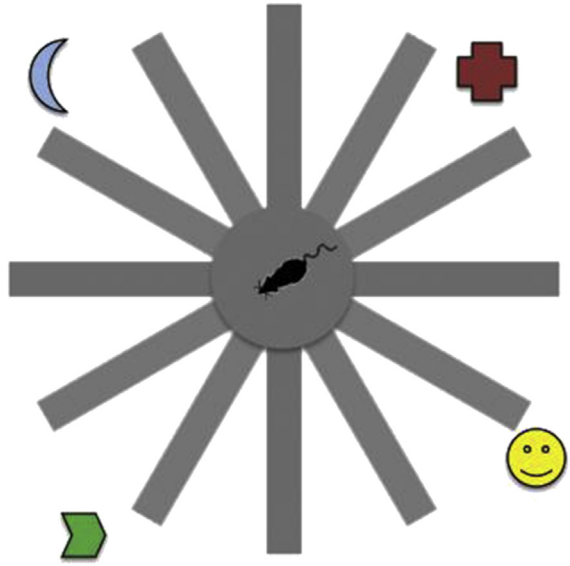


Figure 6.3 Diagram of a typical eight-arm radial arm maze, similar to those used in rats (e.g. Olton et al., 1977) and humans (e.g. Bohbot et al., 2004; Iaria et al., 2003). Shapes represent cues visible in the surrounding environment. (For color version of this figure, the reader is referred to the online version of this book.)

4.3. Ecological Paradigms

An alternative to using paradigms identical to those used in rats has been to develop paradigms for humans that capture the critical features of the rodent paradigms in tasks more suited to human spatial learning and navigation. These tasks offer ecologically grounded tasks and environments but also include elements that mirror or approximate the features of rodent paradigms known to evoke the place/response distinction.

Maguire and colleagues have examined both real and virtual navigation ability (e.g. Hartley et al. 2003; Maguire, Frith, Burgess, Donnett, & O'Keefe, 1998; Spiers & Maguire, 2007). Maguire et al. showed that activation in the right hippocampus was associated with learning and flexibly navigating in a virtual environment, consistent with building up and utilizing a place-learned representation (see also Shelton & Gabrieli, 2002, 2004). Conversely, the degree to which participants demonstrated rapid travel through the environment was associated with activation in the right caudate nucleus. Similarly, Hartley et al. varied the type of learning task and observed a similar distinction. Activation in the hippocampus and caudate was more closely associated with exploratory wayfinding and route following, respectively.

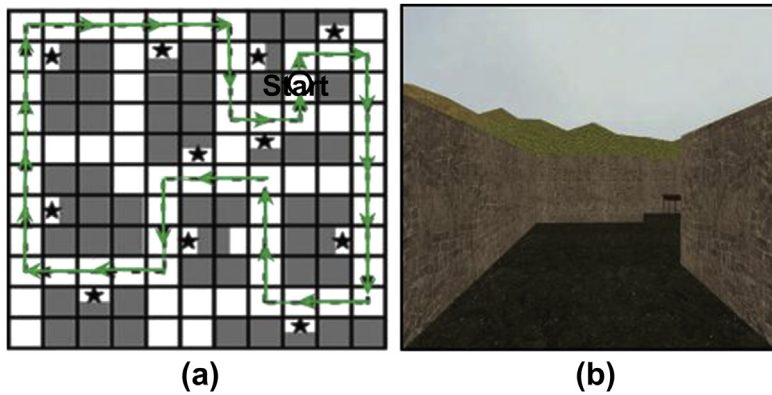


Figure 6.4 The human DSP: (A) schematic of an environment showing the familiar path learned during encoding and locations of the 12 target objects. *(Adapted from Figure 1 of Marchette, Bakker, et al. (2011))*. (B) Image of the actual environment as seen by participants during initial encoding and subsequent navigation. (For color version of this figure, the reader is referred to the online version of this book.)

Exploratory wayfinding offers more flexible exploration, promoting place-like learning, whereas route following offers the repetition required for response-like learning. Activation in these regions was also associated with measures of performance on the respective tasks, suggesting that they might support individual difference in strategies.

More recently, we have introduced the dual-solution paradigm for humans (DSP; Marchette, Bakker, et al., 2011) as a human analog to the dual-solution T-maze. In this paradigm, participants learn the locations of 12 objects in a virtual maze by viewing the same circuitous path through the environment repeatedly. The path is constructed such that one experiences the entire environment (all possible pathways are visible at some point in the path), but the direct experience also allows learning a single specific path. As such, this is an environment that should allow both place-like and response-like learning of the locations (Figure 6.4). At test, participants are placed along the learned path and asked to navigate to a given target. On any given trial, the participant is free to use whatever path he/she prefers; on critical trials, there is always a novel path that is shorter than the familiar learned path. If people are trying to navigate efficiently, as they are instructed to do, then they should opt for shortcuts if they are able to recognize them. That is, when there are two solutions available, they should arguably take the shorter option. The primary behavioral measure is the propensity to use either the shortcut or the familiar path (i.e. solution index).

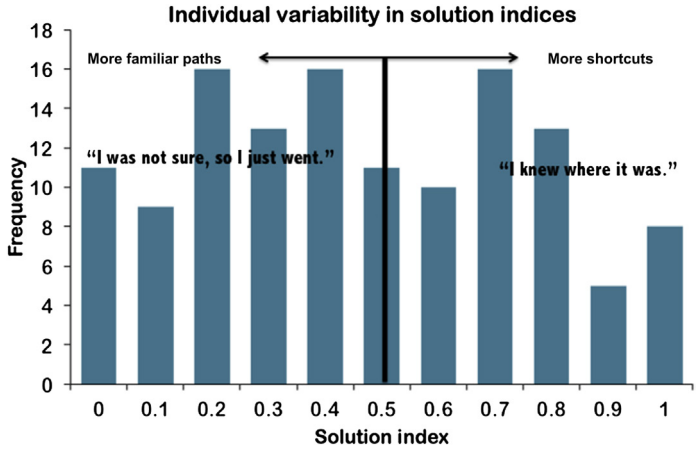


Figure 6.5 Histogram of individual differences in the solution index in the DSP. *(Adapted from Marchette, Bakker, et al. (2011), Figure 2).* (For color version of this figure, the reader is referred to the online version of this book.)

The DSP has been useful in characterizing individual differences because the solution index has revealed substantial variability (Figure 6.5; Marchette, Bakker, et al., 2011) and appears to have good stability in the basic paradigm (Furman, Marchette, & Shelton, 2013). More importantly, people are not strictly using one type of solution or the other. Instead, participants fall along a continuum of performance from people who always use familiar paths to people who always use shortcuts when available, with many individuals using a mixture of both familiar paths and novel shortcuts. In addition, brain activation in the markers for the putative place and response systems is associated with the solution index. In Marchette, Bakker, and Shelton, 20 participants learned the DSP environment during functional magnetic resonance imaging (fMRI) scanning and then performed the subsequent retrieval. Brain activation in the hippocampus and caudate nucleus was measured, and a normed ratio of activation was calculated. This normed ratio of activation during encoding was significantly correlated with the subsequent solution index (Figure 6.6), indicating that the balance of these systems *during encoding* predicts how one chooses to interrogate and utilize the memory during navigation. This predictive value of encoding activation suggests that the interplay of these two systems may offer important clues to why individual differ in their preferred solutions to navigational challenges.

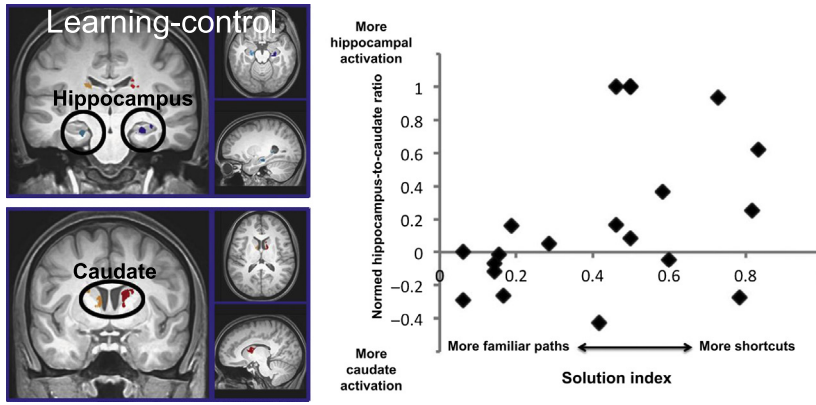


Figure 6.6 Regions of interest analysis for linking brain activation during encoding to navigational solutions. Left panel shows the regions identified in an overall contrast of learning to control. Right panel shows the correlation between the normed ratio of activation in these regions and the subsequent solution index that was observed. *(Adapted from Marchette, Bakker, et al. (2011), Figure 4).* (For color version of this figure, the reader is referred to the online version of this book.)

Taken together, these human analogs to rodent studies provide a clear case both for the existence of place-like and response-like mechanisms and for individual differences in the engagement of these systems. In the following section, we take this a step further to consider how framing individual differences in terms of the contributions of these well-described learning mechanisms might strengthen the current framework and offer novel insights into both the existence and the emergence of different strategies and styles both within and across individuals.

5. THE PLACE/RESPONSE FRAMEWORK FOR INDIVIDUAL DIFFERENCES

Given the wealth of evidence for a distinction akin to place and response learning in humans, it seems plausible that this framework could be considered in parallel with the classic landmark–route–survey framework that has dominated the literature on individual differences in navigational strategies/styles. In this section, we describe how this framework might be utilized and what questions still need to be addressed to fully appreciate a new classification scheme that begins with underlying mechanisms and builds up to how people talk about their own navigational strategies in their daily lives.

5.1. Balance of Systems as the Foundation for Individual Differences

At the core of this framework is the idea that any given individual has both place-like and response-like learning mechanisms available and that these mechanisms underlie most, if not all, human spatial learning. However, even at this mechanistic level, individuals differ in the engagement of these systems during spatial learning. This framework proposes that it is these low-level differences that lay the foundation for the wide-ranging variability in navigational success and strategy across individuals. The classic landmark–route–survey framework and the associated measures start at the end state of what people have learned or prefer to learn and at least implicitly assumes that this classification will generalize to a broad range of navigational challenges. By incorporating a place/response framework, we can begin to address what motivates an individual to report or express a certain end-state strategy. That is, preferences and strategies emerge as skills and experiences that are built upon differences in the bias associated with the relative engagement of the place and response systems. Those experiences and skills will then feed back on the place and response systems to either maintain or shift that relative engagement, allowing both stabilization of preferences and the potential for developing a range of strategies.

The framework does not posit a one-to-one mapping from place and response to specific strategies (e.g. place = survey) but rather offers a more basic level of analysis for thinking about how strategies and preferences emerge. Place learning, as described above, has been associated with terms like *cognitive map* and *explicit learning* of space. It is generally defined as developing a mental representation of locations with respect to the structure and cues in the environment such that it allows flexible interrogation and inferences about the space. Arguably, most people exhibit some form of this type of learning (although many people would claim otherwise), and this is consistent with the extensive observation of activation in the "navigation network" in the brain (e.g. Maguire et al., 1998; Marchette, Bakker, et al., 2011; Shelton & Gabrieli, 2002). However, stronger engagement of this system likely offers access to more survey-like strategies (e.g. Chocolatea is north of campus) as well as more explicit uses of route information (e.g. the Starbuck's is coming up on the right). That is, it is the system that allows one to learn and represent relationships among locations. By contrast, response learning has largely been defined in terms of habits or triggers for specific actions in space. A system that offers ready

access to associated actions would be well suited for serving route-based strategies either by priming particular reactions at decision points in the environment or in coordination with place learning to support more explicit routes.

An important difference between this framework and the more classic classification is the ability to consider how different strategies and preferences might have different profiles of advantages and disadvantages rather than whether those strategies are generally good or bad. When making broad classifications, the interaction of ones classification and the context of navigation is largely limited to whether one is likely to be successful or not given the conditions. For example, an individual classified as a route learner will be expected to succeed or fail as a function of whether the environment affords learning clear paths that can be recreated later. However, if individual differences emerge from biases in lower level learning mechanisms, we can begin to consider what different strategies might bring to any given navigational situation. An additional consequence of this framework is that it affords thinking about a given individuals *range of available strategies* and how that range might interact with the specific environment or context. These advantages are elaborated in the following sections.

5.2. Classification vs Predicting Solutions

Although one important goal in studying individual differences is to offer clear classification, as noted above, any general classification of navigational strategy has the potential to ignore the role of strategy shifts in response to changing navigational challenges. One advantage of the learning mechanisms framework is that it does not reduce individual differences to single categories but offers ideas about how having different biases might alter the set of available solutions and affect an individual's ability to select from within that set. When measuring strategies and solutions with the DSP, we find that most participants varied not in *whether* they used shortcuts or familiar paths but in *how often* they used each solution (Marchette, Bakker, et al., 2011). This variability in degree of use suggests that an important part of characterizing individual differences needs to address when and why a given individual might select one solution in one case and a different solution in another case.

By considering learning mechanisms, we can characterize individuals according to their bias toward place- or response-learning mechanisms as a starting point for developing strategies and predicting which solution(s)

might be employed. For example, if an individual shows a strong caudate/response-learning bias (in brain activation and/or in tendency to use familiar paths), he/she may be primed to use that type of strategy in many situations and may have difficulty engaging alternative solutions. Alternatively, if an individual has a more balanced initial bias, he/she may be more readily able to shift between solutions. By introducing a variety of navigational challenges that favor one solution or another, we can begin to develop a model that not only considers one's initial bias but also explores the ability to shift.

To make some early strides toward this effort, in a recent series of experiments, participants in the DSP were first classified with a solution index. In a subsequent series of trials, participants were told when shortcuts were available. This manipulation was intended to motivate people to look for a shortcut. The results suggested that people who were already using some shortcuts often shifted to using them more frequently once they were cued, whereas some people who were using very limited shortcuts ignored the cues (Furman et al., 2013). Even in this simple case of reduced ambiguity, an individual's initial score on the solution index was a reasonable indicator of how he/she responded to the cues. Ultimately, this will still offer ways to classify individuals, but the classification will be more continuous and offer insights into the range of strategies a given individual might be able to employ.

5.3. Advantages and Disadvantages for Different Solutions

Inherent in the above discussion is the notion that different solutions/strategies may be more or less appropriate for different navigational conditions and challenges. This is in direct contrast to the landmark–route–survey progression, which suggests that flexible survey knowledge is the gold standard. Given that the learning mechanisms framework actually stresses the coexistence of (at least) two systems, appealing to learning mechanisms raises the important question of why these two systems coexist for navigation. One obvious answer is that each system may have an important part to play in the wide range of spatial learning and navigational challenges that humans (and other animals) face. Bringing this to bear on individual differences opens the door for thinking about how different solutions and strategies might offer distinct advantages and disadvantages to allow successful navigation under different conditions.

First, there are known advantages to the place- and response-learning systems in rodents. Whereas place learning affords flexible use of the environmental cues in the face of detours or changes in start location, response learning offers a form of "automaticity" that can be engaged under

conditions of stress or distraction (e.g. Kim, Lee, Han, & Packard, 2001; Packard & Wingard, 2004). Human navigators also face a wide range of conditions, including both unexpected detours and concurrent distractions while navigating. Individual differences in the bias to engage the place and response systems likely affect the reaction to these changing conditions. In a study of distraction, we asked participants to complete the DSP navigational task with and without verbal shadowing (Clark, Marchette, Furman, & Shelton, in preparation). We measured their propensity to use shortcuts or familiar paths (solution index) and observed that people shifted to using the familiar path more often when distracted by verbal shadowing than when not distracted. We also found that performance was worse under distraction. In addition, those who maintained the use of shortcuts in the face of distraction also appeared to take the largest hit on accuracy, suggesting that these individuals were not switching to the more optimal familiar route solutions.

To consider more broadly how different solutions and different learning mechanisms might convey different advantages, it is useful to think in terms of the interaction of (1) the environmental structure and available cues, (2) the skills that the observer brings to bear, and (3) the goals and demands of the navigational challenge. This interaction of factors is consistent with recent models (e.g. Carlson et al., 2010), and the learning mechanisms framework offers a starting point for thinking about how some aspects of this coordination might occur.

In the simplest form, the goal of navigation is to get to one's destination. To that simple goal, we can add different features (one-way streets, hills, multiple stoplights) or different constraints (running late, avoiding bad neighborhoods) that may affect which solutions or strategies might be optimal. For example, in a city with a lot of one-way streets, establishing known routes may be the safest and most efficient way to be successful. This might convey an advantage for someone traditionally classified as a route learner or someone in the place/response framework who has a stronger tendency toward familiar paths. However, this advantage may disappear in that same environment when walking instead of driving. Similarly, we can imagine an environment rich with salient landmarks. In such an environment, encoding locations with respect to landmark proximity might offer an advantage over established sets of routes.[1]

[1] This example reminds me of an anecdote from a visit to a colleague's campus. A graduate student was guiding us to lunch and seemed to be taking a somewhat circuitous route. Upon our arrival at the building, he admitted that he was nervous about getting there correctly because he had been using a crane to beacon to that location for the past 2 years. The crane had recently been removed, and his previously efficient strategy was no longer available.

In addition, there is the internal motivation of an individual that might drive what he/she is willing to do. For example, if the immediate goal is to get somewhere as quickly as possible without error, then it is critical that the individual uses a strategy that will be error free. In this case, any uncertainty about a potential path would reduce the likelihood of using that path. In our work on cuing shortcuts (Furman et al., 2013), we observed this phenomenon in several people. Individuals who opted for the familiar path when uncued would sometimes switch to using a shortcut once they had the very basic knowledge that there was a shortcut available. Many of these individuals reported that they often knew that there might be a shortcut but indicated that they were only willing to try it when the cue offered them certainty. Notably, we also had individuals who found the cues frustrating because they simply had no idea which paths might be the shortcuts; these individuals did not shift their behavior in response to the cues. In the landmark–route–survey framework, both types of individuals would likely be labeled as route learners, but the latter group clearly has a different set of available solutions and must employ them differently.

Finally, by thinking in terms of advantages and disadvantages, we can also begin to think about a broader range of strategies that might emerge from the interaction of these learning mechanisms with experiences, environments, and goals. Even within the place- and response-learning networks, there have been suggestions for additional distinctions. For example, within the striatal system that is believed to support response-like learning, different portions of the neuroanatomy have been associated with different kinds of landmark use: beaconing has been associated with activation in the ventral and medial aspects of the striatum (Devan & White, 1999), whereas the use of landmarks as associative cues has been associated with more dorsal and lateral regions (Featherstone & McDonald, 2004, 2005; see also Chang & Gold, 2004). Understanding how such distinctions in the underlying neural systems manifest themselves in complex behaviors will offer richer ideas about how individual differences emerge. Taken together, thinking about different solutions and strategies as having distinct advantages also raises important questions about what it means to be a good or bad navigator.

5.4. Separating Preference from Prowess

As noted earlier, the classification of spatial ability in terms of success has strong anecdotal and empirical bases, and there has been an emphasis on certain strategies being better or worse than others. However, there has also

been clear recognition that "success" has been defined in specific contexts that may favor one solution or strategy over another. One of the important features of the DSP, driven by the place/response framework, is that different solutions are equally successful. That is, whether someone sticks to the familiar route or opts for a novel path, the trial will be successful as long as they arrive within the time limit. This ability to use different solutions to solve the same problem is consistent with the real-world situation of navigating familiar environments. Although we do encounter the occasional roadblock, our environments generally afford finding and using a particular route or knowing the environmental structure well enough to use multiple routes.

The first important question then is whether there is a clear relationship between success rate and the types of solutions people use. In earlier studies, it is clear that sense of direction as measured by SBSOD was related to an individual's ability to complete certain navigational tasks (for review, see Wolbers & Hegarty, 2010); however, as we have suggested, both the SBSOD and the empirical tests of navigation tend to emphasize the flexible use of space and should therefore by related. In the DSP, where success is determined by taking a clean path to the destination within a reasonable time limit, we can compare solution index (rate of use of shortcuts or familiar paths) to the success rate (percent trials completed). As shown in Figure 6.7, in a large sample, the ability to successfully complete the trials was not dependent on which solution the individual preferred. Critically, the relationship trends in the opposite direction of that predicted from the idea that flexible use of space is best; that is, people using more shortcuts (more place learning) appear more variable and slightly less successful on average than those using more familiar paths. More recent work (as yet unpublished) has replicated this work and substantiated it by showing that self-report measures of strategic preferences are also unrelated to success rate on the DSP, even with more stringent time limits.

Predicting success also carries with it the important question of generalization. Taking the individual's most frequent solution or most pronounced solution may miscalculate their actual ability to respond to certain situations. In other words, just because someone *prefers* to use familiar routes does not necessarily mean that he/she does not have the knowledge to use shortcuts. The use of a more continuous measure of preference (as in the solution index and its variants) may offer more insight into how to measure success rates. Substantial efforts are underway to try to make use of more quantitative approaches to classification, and the place/response framework may offer one of the clearest starting points from which to work.

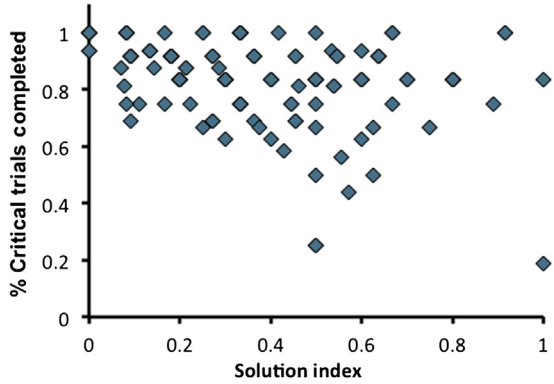


Figure 6.7 Data from the Marchette, Bakker, et al. (2011) studies showing no relationship between solution index and rate of success. (For color version of this figure, the reader is referred to the online version of this book.)

Finally, returning to the question of what it means to be a good or bad navigator, we offer a more adaptive approach. In the simplest form, being good means getting where you need to go. In this sense, the ability to apply an appropriate strategy or solution, given the constraints of the environment, the skills of the observer, and the current goals, should be the mark of good navigation. Our work on shifting strategies under instruction (Furman et al., 2013) and conditions of distraction (Clark et al., in preparation) begin to weigh in on this issue by assessing who can shift solutions and how much they shift as a function of a wide variety of spatial skills and preferences. Ongoing work in this domain is likely to offer remarkable new insights and new ways to think about individual differences in navigational ability.

6. CONNECTIONS TO OTHER SOURCES OF VARIABILITY

In addition to providing a general framework for thinking about individual differences, the appeal to a place/response model has implications for other sources of variability. A large body of research has identified factors, such as aging, sex differences, and hormonal balance that can modulate navigational success and strategy. Although post hoc descriptions of how these factors impact navigation have been marshaled, there is currently no framework for explaining how these modulators may shape differences among individuals However, a direct benefit of considering neurobiologically grounded learning mechanisms is that it allows us to build a connection

between these factors and how they may shape individual differences in the way people use strategies and solutions.

6.1. Aging

Aging is associated with decline in a number of cognitive domains; among them are the processes of spatial learning and navigation. When compared with their younger counterparts, older adults are impaired in their ability to learn and remember ways of finding their way to important destinations (for review, see Moffat, 2009), the physical layout of the environment (Iaria, Palermo, Committeri, & Barton, 2009; Kirasic & Mathes, 1990; Kirasic, 2000), as well as their position within it (Iaria et al., 2009; Mahmood, Adamo, Briceno, & Moffat, 2009). This suggests one important area where an appeal to learning mechanisms might offer insights into the effects of aging.

In rodents, there has been substantial effort to understand how such changes in behavior with aging might be related to differential changes in the underlying spatial learning mechanisms (e.g. Gallagher & Pelleymounter, 1988; Rapp, Rosenberg, & Gallagher, 1987; Rosenzweig & Barnes, 2003). For example, when allowed to solve a task using place- or response-related information, aged rats rely on more response-based solutions than younger animals (Barnes, Nadel, & Honig, 1980). In humans, a similar case appears to be building. Notably, the kinds of tasks that demonstrate the spatial memory decline focus on explicit route learning or the ability to make flexible inference using spatial representations of the environment. For example, in variants of the MWM adapted for humans, the common observation is that older adults take consistently longer to locate the platform, spend less time near its former location on probe trials, and are less able to explicitly indicate where the platform is located compared with younger adults (e.g. Driscoll et al., 2005; Moffat, 2009; Moffat & Resnick, 2002). In addition, these age-related deficits are accompanied by functional impairment within the medial temporal lobe (e.g. De Leon et al., 1997; Jack et al., 1998) and often the broader navigation network (parahippocampal gyrus and retrosplenial cortex; see e.g. Meulenbroek, Petersson, Voermans, Weber, & Fernández, 2004; Moffat, Elkins, & Resnick, 2006). Taken together, these results in rodents and humans are consistent with a relatively selective impairment in the place-learning system.

In the framework we have outlined, these apparent deficits in place learning with aging would suggest the loss or impairment with some strategies and only in some situations, allowing specific predictions about patterns of impairment and sparing. First, individuals who favor solutions and

strategies that depend more heavily on a response-like system might show less degradation in their overall navigational experiences. Similarly, individuals who can readily shift strategies should have the capacity to compensate the loss of place-learning proficiency by utilizing other solutions. Indeed, on versions of the virtual MWM that allow more response-based strategies, caudate nucleus volume has been positively correlated with successful performance, suggesting that increased dependence on the response-learning system might counter at least some of the deficits observed with aging (Moffat, Kennedy, Rodrigue, & Raz, 2007). These predictions suggest that the way we test for spatial learning and memory deficits in aging may need to include tests for alternative solutions and styles.

A final key fact about aging is that the impairments are graded in nature, with some individual rats even seeming to escape impairment (Gallagher, Burwell, & Burchinal, 1993), and similar sparing in human aging is associated with a reduced loss of hippocampal volume (Driscoll et al., 2003). These results suggest that age-related impairments in spatial learning may be tied to the degree place-learning impairment, an observation in line with our results showing that use of strategies was associated with place learning falls along a continuum. Although it currently remains unexplored, an expectation of our model is that these graded deficits in place learning should be met with similarly graded increases in reliance on response learning. In addition, our work on shifting strategies suggests that some individuals may fail to show this shift toward response learning. These predictions offer another potential avenue for examining why some individuals show more marked decline then others, which may feed ideas for methods for helping aging individuals to adapt their strategies more readily.[2]

6.2. Sex Differences and Hormonal Influences

The domain of navigation has traditionally been a fertile ground for studying sex differences because there have been many reports of differences between males and females in their ability to learn and use particular sorts of spatial information. Most commonly, males tend to prefer using knowledge of their position with respect to locations within the environment, or survey knowledge, whereas females tend to prefer more route-based forms of information (e.g. Lawton, 1994, 1996). In line with these self-reports,

[2] Much of this discussion is also relevant to some of the navigational deficits observed in such neurological disorders as mild cognitive impairment and Alzheimer's disease, where impairments have been associated with hippocampal damage or dysfunction (e.g. Bellassen, Iglói, de SouzaDubois, & Rondi-Reig, 2012).

sex differences tend to be revealed in tasks that require participants to make judgments over the larger scale global structure (e.g. Lawton, Charleston, & Zieles, 1996) and in situations that require navigation using cardinal directions rather than landmark information (Saucier et al., 2002). These results suggest that men and women may rely on and learn about different sources of information in the environment (see also Kelly, McNamara, Bodenheimer, Carr, & Rieser, 2009; Sandstrom, Kaufman & Huettel, 1998).

Sex differences can also be linked to different hormonal profiles, another potential sources of variability that may affect navigational behaviors. Moreover, place- and response-learning mechanisms have already been associated with different hormonal profiles in both rodents (e.g. Davis, Jacobson, Aliakbari, & Mizumori, 2005; Gold & Korol, 2010; Korol, 2004; Naghdi, Majlessi, & Bozorgmehr, 2005; Roof & Havens, 1992; Warren & Juraska, 1997) and humans (Bell & Saucier, 2004; Driscoll et al., 2005). For example, work by Korol and colleagues has shown that experimentally manipulated levels of estrogen in female rats modulate the preferential use of place and response learning. In particular, they found that estrogen infusions into the hippocampus, but not striatum, enhanced place learning, whereas estrogen infusions into the striatum, but not hippocampus, impaired response learning. Although results with testosterone have been more varied (e.g. Clark, Mitre, & Brinck-Johnsen, 1995; Naghdi et al., 2005; Roof & Havens, 1992; Sandstrom, Kim, & Wasserman, 2006), there have been suggestions that testosterone has a relationship to MWM performance that may be associated with sex difference. For example, Roof and Havens observed that the hippocampal formation in female rats given neonatal testosterone was more similar to male rats than to other females that did not receive testosterone; the treated females also showed more male-like performance than nontreated rats. These hormonal influences in rodents offer a potential explanation for sex differences as a function of differential hormonal effects on place and response learning.

In humans, circulating testosterone levels appear to be a predictor of both spatial reasoning and the ability to acquire spatial knowledge (Bell & Saucier, 2004; Driscoll et al., 2005). Critically, lower levels of testosterone were associated with better performance in men, whereas higher levels of testosterone are associated with better performance in women, suggesting that too much or too little circulating hormone affected performance on explicit kinds of spatial learning tasks. Our framework, in conjunction with the observations in rodents, would suggest that these associations with apparent place learning might predict changes in response-learning rates as well. In particular, it

predicts that conditions that reduce place learning might enhance response learning, which would, in turn, change the balance of strategies that individuals utilize. Again, as in other areas, this would take our thinking away from ideas about how hormones are good or bad for navigation and toward ideas about what people might engage under different hormonal influences.

Hormones are a particularly interesting factor to consider because, in addition to capturing features about the specific individual (e.g. baseline levels of circulating testosterone), hormones are sensitive to acute state changes. For example, stress-related hormones may play a similar role in changing the bias between the memory systems associated with place and response learning (for review, see Packard, 2009; Brain Research). In particular, chronic stress appears to reduce the hippocampal plasticity that is thought to be a core component of the place-learning system (e.g. Kim & Yoon, 1998). What's more, systemic administration of corticosteroids, a class of hormones released in response to stress, biases rodents away from place learning and toward response learning in a maze that allowed both solutions (Sadowski, Jackson, Wieczorek, & Gold, 2009; Schwabe et al., 2010). This may offer interesting insights into human navigational styles because high spatial anxiety tends to coincide with more frequent use or reliance on familiar routes (Hund & Minarik, 2006; Lawton, 1994). Thinking in terms of biasing systems allows us to make specific predictions about the interplay of stress hormones, acute or chronic stress/anxiety, and navigational styles. This is another example of the theoretical richness a neurobiologically grounded framework of individual differences might provide.

7. COMPETITION OR INTERACTION OF SYSTEMS

Before drawing some final conclusions about the role of a place/response framework in thinking about individual differences, it is important to address the relationship between the putative place and response systems. In the literature on rodents and humans, there has long been a debate about the degree to which these systems are competing with each other for resources or interacting with each other to give rise to behavior (e.g. Poldrack & Packard, 2003). Although a thorough discussion of this issue is beyond the scope of this chapter, there are some important points to make with respect to how the debate affects the proposed framework for individual differences.

In some sense, the basic place and response characterization based on rodent behavior evokes a sense of competition by necessity. That is, only

one type of behavior can be produced, and therefore, only one system can be manifested at any given time. For example, when Tolman and colleagues placed rats in a T-maze from a novel start position, the rat had to either turn toward the arm consistent with the target location relative to the environmental cues or turn toward the arm consistent with the repeated action during the acquisition (e.g. Packard & McGaugh, 1996; Restle, 1957; Tolman et al., 1946, 1947). Therefore, in terms of being able to drive behavior, there must at least be some form of competition that allows one system to dominate at any given time.

Beyond the competition to drive behavior, one can ask whether there is a necessary trade-off in resources between the two neural systems that support the different learning mechanisms. The first important point is that the two systems are operating in parallel during encoding, as evidenced by the extensive work showing that rodents can successfully learn under conditions that require either a place or a response approach. Moreover, shutting down one system at the time of retrieval allows the other system to control behavior (Packard & McGaugh, 1996), further supporting the ideas that learning was occurring simultaneously. This coincidental learning is consistent with an account that allows for either independence or coordination of these systems, with the "competition" largely at the behavioral output end.

In humans, the story with respect to competition or cooperation is less clear. Marchette, Bakker, et al. (2011) found that the activation in the hippocampus and caudate nucleus during encoding predicted the solutions that were employed. Critically, although the individual regions predicted performance as expected, it was the relative activation of the two regions that offered the strongest correlation. This appeal to the balance of activation suggests that both systems may be operating at higher or lower levels such that when they are in balance, people are likely to use a mixture of solutions and that mixture shifts as the balance of engagement shifts.

In an alternative approach to the distinction between hippocampal and striatal contributions, Brown et al. (2012) investigated the concurrent activation in hippocampus and caudate, as well as prefrontal regions, in participants learning and navigating in virtual mazes. Participants learned to navigate in a series of small mazes and were then asked to navigate those learned environments during fMRI scans. The activation was then interrogated for mazes that shared overlapping features compared with those that did not. The results revealed greater functional connectivity between the

hippocampus and the caudate in conditions that required making behavioral distinctions in environments with overlapping features compared with conditions in which the environments did not overlap. This finding suggests that the underlying place-like and response-like systems may operate in the disambiguation of different navigational context.

Overall, the question of whether these two systems are competing, independent, or interactive may depend on the particular situation in which an individual is acting. In the framework proposed here, the critical concept is that these two systems are present and show individual variability in their engagement. As both the questions of competition vs coordination are addressed and the issues surrounding individual differences become clarified, the proposed place/response framework will become elaborated and enriched.

8. CONCLUSIONS

Our understanding of individual differences in navigation has long been shaped by models and approaches that emphasize how people engage novel spaces. The standard landmark–route–survey framework proposes that, in an attempt to maximize their later success, people try to focus on learning the information they think will be important to remember later, such as the location of a particularly salient landmark, a sequence of directions along a route, or the spatial relationships between important places. People then ultimately meet with different amounts of success that reflect how much their strategy has allowed them to accurately learn about the environment, with good navigators identified by their ability to form more holistic spatial representations. This framework stresses what people explicitly try to do, and much of our understanding of these complex strategies necessarily comes from people's self-report. These descriptions of what people report using are then in turn used to classify navigators into different discrete categories.

The work summarized here on human place- and response-learning mechanisms paints a complementary picture to the landmark–route–survey model but may also challenge some of its norms for defining individual differences. The principle challenge stems from the observation that when we tested participants on the DSP, we did not see two clear clusters of people who preferred place-like or response-like solutions. Instead, participants ranged continuously across the spectrum of possible preferences. Consistent with the rodent models, this continuous range suggests that both

learning mechanisms are available to any individual and that each individual has some bias of variable strength toward one learning mechanism or the other. In practice, this also means that an individual is using fundamentally different solutions on different trials. This observation may push us out of old norms for thinking about distinct categories for classifying individuals based on what they report to be their dominant strategy.

Once we allow the possibility that the use of navigational strategy, at least at the level of the deployment of learning mechanisms, is continuous rather than discrete, we have to develop ideas about *why* strategies are deployed continuously. One possibility is that the strategy of the moment may be shaped by the particular circumstances of navigation, rather than being determined by what participants try to do in general. This is a concept that heralds a change from thinking about there being strategies that are "better" and "worse" overall: normatively, under a better/worse scheme, an individual should principally use the best strategy that he or she can!

An alternative to the standard viewpoint, and one that we subscribe to, is that different strategies may convey advantages and disadvantages, which might be suitable to particular situations or optimize different navigational criteria, such as weighing speed of travel against certainty of arrival. Having access to both learning mechanisms is then adaptive because it allows an individual to select the strategy that will best match their own navigational goals and constraints. This framing may redirect our attention to how strategy selection may reflect the *constraints* that an individual navigator places on defining a successful trip.

Taken together, these different profiles of strengths and weaknesses mean that some situations will call for place learning whereas others may call for response learning. Navigational success, then, is not simply associated with which solution one may prefer in the DSP. However, a fine point in this argument is that we are not arguing that one's predisposition for place and response learning is unrelated to navigational success in the real world. The DSP was designed such that both place and response learning can support successful navigation, but in the real world, this may not always be the case. Our suggestion is that, while prowess with a chosen strategy matters, navigational success on a particular trip will be based primarily on how well the individual's strategy fits the particular situation. Our traditional predisposition as researchers to call place-like strategies "good" consequently results from the fact that they are widely applicable, rather than universally more successful than response-like strategies. Under this view, a truly successful

navigator is not someone with a commitment to a particularly "good" strategy, but rather one who is able to recognize the right strategy at the right time and have the flexibility, and ability, to apply it.

The learning mechanisms framework addresses a contradiction that arises in the study of individual differences in navigation. A common trope that virtually every paper on navigation uses to motivate their study is that navigation is a core skill for mobile beings and is crucial to survival. Despite this claim, a notable portion of the population reports being quite bad at navigation without experiencing daily threats to their survival. In our view, the resolution of this contradiction must be to accept that most healthy individuals have adapted to the navigational needs of their time and environment and found solutions that work. We do not question the value of research into how people learn and adapt to novel environments. Indeed, it is precisely these kinds of studies that will enlighten our understanding of individual differences. However, we suggest that studying learning mechanisms will open up a rich largely unstudied territory in navigational research: what are the challenges and criteria of daily navigation in the real world and how might an individual's strategies and aptitudes be mustered to meet them? How flexibly can these be shifted as circumstances change?

In conclusion, the distinction between place and response learning is both historically important and incredibly promising for enlightening our understanding of spatial learning mechanisms. However, it is not immediately obvious how place- and response-learning mechanisms translate to the many complex strategies that people report for navigating the world. There do seem to be some analogies between place learning and survey or "map-like" strategies and response learning and route-based navigation; however, we suggest here that this superficial similarity masks more interesting relationships. Rather than simple one-to-one correspondences, the concepts of place and response learning form the foundations for understanding how different learning mechanisms with different intrinsic biases can cooperate or compete to give rise to complex strategies that might be appropriate for the challenges of navigation. Here, we have argued that the study of these learning mechanisms provides a framework for understanding individual differences in navigation—not a fully specified model. This is not an oversight on our part, but instead reflects the multidimensional nature of navigation and the need to expand our focus to the range of navigational circumstances, criteria, and constraints that individuals face on a daily basis.

REFERENCES

Allen, G. L., Kirasic, K. C., Dobson, S. H., Long, R. G., & Beck, S. (1996). Predicting environmental learning from spatial abilities: an indirect route. *Intelligence*, *22*(3), 327–355.

Astur, R. S., Taylor, L. B., Mamelak, A. N., Philpott, L., & Sutherland, R. J. (2002). Humans with hippocampus damage display severe spatial memory impairments in a virtual Morris water task. *Behavioural Brain Research*, *132*(1), 77–84.

Barnes, C. A., Nadel, L., & Honig, W. K. (1980). Spatial memory deficit in senescent rats. *Canadian Journal of Psychology*, *34*(1), 29.

Bellassen, V., Iglói, K., de Souza, L. C., Dubois, B., & Rondi-Reig, L. (2012). Temporal order memory assessed during spatiotemporal navigation as a behavioral cognitive marker for differential Alzheimer's disease diagnosis. *Journal of Neuroscience*, *32*(6), 1942–1952.

Bell, S., & Saucier, D. (2004). Relationship among environmental pointing accuracy, mental rotation, sex, and hormones. *Environment and Behavior*, *36*(2), 251–265.

Bird, C. M., & Burgess, N. (2008). The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, *9*(3), 182–194.

Bohbot, V. D., Iaria, G., & Petrides, M. (2004). Hippocampal function and spatial memory: evidence from functional neuroimaging in healthy participants and performance of patients with medial temporal lobe resections. *Neuropsychology*, *18*(3), 418.

Bohbot, V. D., Kalina, M., Stepankova, K., Spackova, N., Petrides, M., & Nadel, L. Y. N. N. (1998). Spatial memory deficits in patients with lesions to the right hippocampus and to the right parahippocampal cortex. *Neuropsychologia*, *36*(11), 1217–1238.

Brown, T. I., Ross, R. S., Tobyne, S. M., & Stern, C. E. (2012). Cooperative interactions between hippocampal and striatal systems support flexible navigation. *NeuroImage*, *60*(2), 1316.

Carlson, L. A., Hölscher, C., Shipley, T. F., & Dalton, R. C. (2010). Getting lost in buildings. *Current Directions in Psychological Science*, *19*(5), 284–289.

Chan, E., Baumann, O., Bellgrove, M. A., & Mattingley, J. B. (2012). From objects to landmarks: the function of visual location information in spatial navigation. *Frontiers in Psychology*, *3*, 304.

Chang, Q., & Gold, P. E. (2004). Inactivation of dorsolateral striatum impairs acquisition of response learning in cue-deficient, but not cue-available, conditions. *Behavioral Neuroscience*, *118*(2), 383.

Chrastil, E. R. (2012). Neural evidence supports a novel framework for spatial navigation. *Psychonomic Bulletin & Review*.

Clark, S., Marchette, S. A., Furman, A. J., & Shelton, A. L. *Talking while walking: investigating the effects of verbal distraction on human navigation performance and strategy selection*, in preparation.

Clark, A. S., Mitre, M. C., & Brinck-Johnsen, T. (1995). Anabolic-androgenic steroid and adrenal steroid effects on hippocampal plasticity. *Brain Research*, *679*(1), 64–71.

Cornwell, B. R., Johnson, L. L., Holroyd, T., Carver, F. W., & Grillon, C. (2008). Human hippocampal and parahippocampal theta during goal-directed spatial navigation predicts performance on a virtual Morris water maze. *Journal of Neuroscience*, *28*(23), 5983–5990.

Davis, D. M., Jacobson, T. K., Aliakbari, S., & Mizumori, S. J. Y. (2005). Differential effects of estrogen on hippocampal-and striatal-dependent learning. *Neurobiology of Learning and Memory*, *84*(2), 132–137.

De Leon, M. J., George, A. E., Golomb, J., Tarshish, C., Convit, A., Kluger, A., et al. (1997). Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease. *Neurobiology of Aging*, *18*(1), 1–11.

Devan, B. D., & White, N. M. (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. *Journal of Neuroscience*, *19*(7), 2789–2798.

Driscoll, I., Hamilton, D. A., Petropoulos, H., Yeo, R. A., Brooks, W. M., Baumgartner, R. N., et al. (2003). The aging hippocampus: cognitive, biochemical and structural findings. *Cerebral Cortex*, *13*(12), 1344–1351.

Driscoll, I., Hamilton, D. A., Yeo, R. A., Brooks, W. M., & Sutherland, R. J. (2005). Virtual navigation in humans: the impact of age, sex, and hormones on place learning. *Hormones and Behavior*, *47*(3), 326–335.

Elliott, A. E., & Packard, M. G. (2008). Intra-amygdala anxiogenic drug infusion prior to retrieval biases rats towards the use of habit memory. *Neurobiology of Learning and Memory*, *90*(4), 616–623.

Featherstone, R. E., & McDonald, R. J. (2004). Dorsal striatum and stimulus–response learning: lesions of the dorsolateral, but not dorsomedial, striatum impair acquisition of a simple discrimination task. *Behavioural Brain Research*, *150*(1), 15–23.

Featherstone, R. E., & McDonald, R. J. (2005). Lesions of the dorsolateral or dorsomedial striatum impair performance of a previously acquired simple discrimination task. *Neurobiology of Learning and Memory*, *84*(3), 159–167.

Fields, A. W., & Shelton, A. L. (2006). Individual skill differences and large-scale environmental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 506.

Furman, A. J., , Marchette, S. A., & Shelton, A. L. (2013). *Individual differences in navigational style and its relationship to volitional strategy shifting*. Manuscript submitted for publication.

Gallagher, M., Burwell, R., & Burchinal, M. R. (1993). Severity of spatial learning impairment in aging: development of a learning index for performance in the Morris water maze. *Behavioral Neuroscience*, *107*(4), 618.

Gallagher, M., & Pelleymounter, M. A. (1988). Spatial learning deficits in old rats: a model for memory decline in the aged. *Neurobiology of Aging*, *9*, 549–556.

Gold, P. E., & Korol, D. L. (2010). Hormones and memory. In G. Koob, M. Le Moal & R. F. Thompson (Eds.), *Encyclopedia of behavioral neuroscience* (Vol. 2, pp. 57–64). Oxford: Academic Press.

Hamilton, D. A., Driscoll, I., & Sutherland, R. J. (2002). Human place learning in a virtual Morris water task: some important constraints on the flexibility of place navigation. *Behavioural Brain Research*, *129*(1), 159–170.

Hartley, T., Maguire, E. A., Spiers, H. J., & Burgess, N. (2003). The well-worn route and the path less traveled: distinct neural bases of route following and wayfinding in humans. *Neuron*, *37*(5), 877–888.

Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, *30*(5), 425–447.

Hund, A. M., & Minarik, J. L. (2006). Getting from here to there: spatial anxiety, wayfinding strategies, direction type, and wayfinding efficiency. *Spatial Cognition and Computation*, *6*(3), 179–201.

Iaria, G., Palermo, L., Committeri, G., & Barton, J. J. (2009). Age differences in the formation and use of cognitive maps. *Behavioural Brain Research*, *196*(2), 187–191.

Iaria, G., Petrides, M., Dagher, A., Pike, B., & Bohbot, V. D. (2003). Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: variability and change with practice. *Journal of Neuroscience*, *23*(13), 5945–5952.

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129.

Jack, C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., et al. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*, *51*(4), 993–999.

Kelly, J. W., McNamara, T. P., Bodenheimer, B., Carr, T. H., & Rieser, J. J. (2009). Individual differences in using geometric and featural cues to maintain spatial orientation: cue quantity and cue ambiguity are more important than cue type. *Psychonomic Bulletin & Review*, *16*(1), 176–181.

Kim, J. J., Lee, H. J., Han, J. S., & Packard, M. G. (2001). Amygdala is critical for stress-induced modulation of hippocampal long-term potentiation and learning. *Journal of Neuroscience*, *21*(14), 5222–5228.

Kim, J. J., & Yoon, K. S. (1998). Stress: metaplastic effects in the hippocampus. *Trends in Neurosciences*, *21*(12), 505–509.
Kirasic, K. C. (2000). Age differences in adults' spatial abilities, learning environmental layout, and wayfinding behavior. *Spatial Cognition and Computation*, *2*(2), 117–134.
Kirasic, K. C., & Mathes, E. A. (1990). Effects of different means for conveying environmental information on elderly adults' spatial cognition and behavior. *Environment and Behavior*, *22*(5), 591–607.
Korol, D. L. (2004). Role of estrogen in balancing contributions from multiple memory systems. *Neurobiology of Learning and Memory*, *82*(3), 309–323.
Lawton, C. A. (1994). Gender differences in way-finding strategies: relationship to spatial ability and spatial anxiety. *Sex Roles*, *30*(11), 765–779.
Lawton, C. A. (1996). Strategies for indoor wayfinding: the role of orientation. *Journal of Environmental Psychology*, *16*(2), 137–145.
Lawton, C. A., Charleston, S. I., & Zieles, A. S. (1996). Individual-and gender-related differences in indoor wayfinding. *Environment and Behavior*, *28*(2), 204–219.
Maguire, E. A., Burgess, N., Donnett, J. G., Frackowiak, R. S., Frith, C. D., & O'Keefe, J. (1998). Knowing where and getting there: a human navigation network. *Science*, *280*(5365), 921–924.
Mahmood, O., Adamo, D., Briceno, E., & Moffat, S. D. (2009). Age differences in visual path integration. *Behavioural Brain Research*, *205*(1), 88–95.
Marchette, S. A., Bakker, A., & Shelton, A. L. (2011). Cognitive mappers to creatures of habit: differential engagement of place and response learning mechanisms predicts human navigational behavior. *Journal of Neuroscience*, *31*(43), 15264–15268.
Marchette, S. A., Yerramsetti, A., Burns, T. J., & Shelton, A. L. (2011). Spatial memory in the real world: long-term representations of everyday environments. *Memory & Cognition*, *39*(8), 1401–1408.
Meulenbroek, O., Petersson, K. M., Voermans, N., Weber, B., & Fernández, G. (2004). Age differences in neural correlates of route encoding and route recognition. *NeuroImage*, *22*(4), 1503–1514.
Moffat, S. D. (2009). Aging and spatial navigation: what do we know and where do we go? *Neuropsychology Review*, *19*(4), 478–489.
Moffat, S. D., Elkins, W., & Resnick, S. M. (2006). Age differences in the neural systems supporting human allocentric spatial navigation. *Neurobiology of Aging*, *27*(7), 965–972.
Moffat, S. D., Kennedy, K. M., Rodrigue, K. M., & Raz, N. (2007). Extrahippocampal contributions to age differences in human spatial navigation. *Cerebral Cortex*, *17*(6), 1274–1282.
Moffat, S. D., & Resnick, S. M. (2002). Effects of age on virtual environment place navigation and allocentric cognitive mapping. *Behavioral Neuroscience*, *116*(5), 851.
Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation*, *12*(2), 239–260.
Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*(1), 47–60.
Morris, R. G. M., Garrud, P., Rawlins, J. N. P., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*(5868), 681–683.
Naghdi, N., Majlessi, N., & Bozorgmehr, T. (2005). The effect of intrahippocampal injection of testosterone enanthate (an androgen receptor agonist) and anisomycin (protein synthesis inhibitor) on spatial learning and memory in adult, male rats. *Behavioural Brain Research*, *156*(2), 263–268.
Newman, M. C., & Kaszniak, A. W. (2000). Spatial memory and aging: performance on a human analog of the Morris water maze. *Aging, Neuropsychology, and Cognition*, 7(2), 86–93.
Olton, D. S., Collison, C., & Werz, M. A. (1977). Spatial memory and radial arm maze performance of rats. *Learning and Motivation*, *8*(3), 289–314.

Packard, M. G. (2009). Anxiety, cognition, and habit: a multiple memory systems perspective. *Brain Research*, *1293*, 121–128.
Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, *65*(1), 65–72.
Packard, M. G., & Wingard, J. C. (2004). Amygdala and "emotional" modulation of the relative use of multiple memory systems. *Neurobiology of Learning and Memory*, *82*(3), 243–252.
Pazzaglia, F., & De Beni, R. (2001). Strategies of processing spatial information in survey and landmark-centred individuals. *European Journal of Cognitive Psychology*, *13*(4), 493–508.
Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251.
Rapp, P. R., Rosenberg, R. A., & Gallagher, M. (1987). An evaluation of spatial information processing in aged rats. *Behavioral Neuroscience*, *101*(1), 3.
Restle, F. (1957). Discrimination of cues in mazes: a resolution of the "place-vs.-response" question. *Psychological Review*, *64*(4), 217.
Ritchie, B. F., Aeschliman, B., & Pierce, P. (1950). Studies in spatial learning. VIII. Place performance and the acquisition of place dispositions. *Journal of Comparative and Physiological Psychology*, *43*(2), 73.
Roof, R. L., & Havens, M. D. (1992). Testosterone improves maze performance and induces development of a male hippocampus in females. *Brain Research*, *572*(1–2), 310–313.
Rosenzweig, E. S., & Barnes, C. A. (2003). Impact of aging on hippocampal function: plasticity, network dynamics, and cognition. *Progress in Neurobiology*, *69*(3), 143–179.
Sadowski, R. N., Jackson, G. R., Wieczorek, L., & Gold, P. E. (2009). Effects of stress, corticosterone, and epinephrine administration on learning in place and response tasks. *Behavioural Brain Research*, *205*(1), 19–25.
Sandstrom, N. J., Kaufman, J., & Huettel, S. A. (1998). Males and females use different distal cues in a virtual environment navigation task. *Cognitive Brain Research*, *6*(4), 351–360.
Sandstrom, N. J., Kim, J. H., & Wasserman, M. A. (2006). Testosterone modulates performance on a spatial working memory task in male rats. *Hormones and Behavior*, *50*(1), 18–26.
Saucier, D. M., Green, S. M., Leason, J., MacFadden, A., Bell, S., & Elias, L. J. (2002). Are sex differences in navigation caused by sexually dimorphic strategies or by differences in the ability to use the strategies? *Behavioral Neuroscience*, *116*(3), 403.
Schwabe, L., Schächinger, H., de Kloet, E. R., & Oitzl, M. S. (2010). Corticosteroids operate as a switch between memory systems. *Journal of Cognitive Neuroscience*, *22*(7), 1362–1372.
Shelton, A. L., & Gabrieli, J. D. (2002). Neural correlates of encoding space from route and survey perspectives. *Journal of Neuroscience*, *22*(7), 2711–2717.
Shelton, A. L., & Gabrieli, J. D. (2004). Neural correlates of individual differences in spatial learning strategies. *Neuropsychology*, *18*(3), 442.
Siegel, A. W., & White, S. H. (1975). The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, *10*, 9–55.
Skelton, R. W., Ross, S. P., Nerad, L., & Livingstone, S. A. (2006). Human spatial navigation deficits after traumatic brain injury shown in the arena maze, a virtual Morris water maze. *Brain Injury*, *20*(2), 189–203.
Spiers, H. J., & Maguire, E. A. (2007). A navigational guidance system in the human brain. *Hippocampus*, *17*(8), 618–626.
Stewart, C. A., & Morris, R. G. M. (1993). The watermaze. In A. Saghal (Ed.), *Behavioral neuroscience. A practical approach* (Vol. 1, pp. 107–122). Oxford: IRL Press.
Taylor, H. A., Naylor, S. J., & Chechile, N. A. (1999). Goal-specific influences on the representation of spatial perspective. *Memory & Cognition*, *27*(2), 309–319.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. II. Place learning versus response learning. *Journal of Experimental Psychology*, *36*(3), 221.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1947). Studies in spatial learning. V. Response learning vs. place learning by the non-correction method. *Journal of Experimental Psychology*, *37*(4), 285.

Waller, D., & Lippa, Y. (2007). Landmarks as beacons and associative cues: their role in route learning. *Memory & Cognition*, *35*(5), 910–924.

Warren, S. G., & Juraska, J. M. (1997). Spatial and nonspatial learning across the rat estrous cycle. *Behavioral Neuroscience*, *111*(2), 259.

Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences*, *14*(3), 138–146.

Yerramsetti, A., Marchette, S. A., & Shelton, A. L. (2012). Accessibility versus accuracy in retrieving spatial memory: evidence for suboptimal assumed headings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/a0030905. Advance online publication.

This page intentionally left blank

CHAPTER SEVEN

When Do the Effects of Distractors Provide a Measure of Distractibility?

Alejandro Lleras*,[1], Simona Buetti*, J. Toby Mordkoff†
*Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA
†Department of Psychology, University of Iowa, Iowa City, IA, USA
[1]Corresponding author: E-mail: alejandrolleras@gmail.com

Contents

Psychology of Learning and Motivation, Volume 59
ISSN 0079-7421, http://dx.doi.org/10.1016/B978-0-12-407187-2.00007-1

Abstract

We discuss how, at the present time, there is a large deal of confusion in the attention literature regarding the use of the label "distractor" and what may be inferred from experiments using distractors. In particular, investigators seem to use the concepts of distractor interference and distractibility almost interchangeably. In contrast, we argue at both the theoretical and empirical levels that these two concepts are not only different, but in fact mutually exclusive. To that end, a brief review of several subliteratures is presented, in which we identify some examples of the misuse of these terms. We also propose a new paradigm for the study of distraction, as well as present a contemporary general theory of visual attention that provides a better framework for understanding distractor-interference effects, as well as instances of true distraction.

1. INTRODUCTION

What is "distraction"? To borrow some from James' famous definition of attention, one could say that distraction is the taking possession of the mind in clear and vivid form by a thought or stimulus *that one never intended to process in the first place*. It is a thought or stimulus that takes us "out" of our intended task and drives our attention onto new thoughts or sensations. This definition seems quite close to our intuitions of what being distracted is like, and matches the vernacular definition of distraction (e.g. "the drawing away [of the mind or thoughts] from one point or course to another; diversion of the mind or attention", Oxford English Dictionary). It is also a definition that is implied by work in applied psychology, such as the recent studies on distracted driving (e.g. Cooper, Vladisavljevic, Medeiros-Ward, Martin, & Strayer, 2009; Strayer, Watson, & Drews, 2011). It certainly is an important concept in our everyday life, so it is to no surprise that we, in the attention field, may want to draw inferences about this concept from our investigations. One of the primary goals of this chapter is to point out where this effort has faltered and identify one reason why this has happened: the inappropriate assumption that the effects of distractors are always a measure of distraction.

The history of the term "distractor" is as old as the discipline of psychology itself. One can easily find papers using this term as far back as the nineteenth century (e.g. Darlington & Talbot, 1898). In the early days, distractors were understood to be that type of stimulus that can distract us from our main tasks and, thereby, disturb our performance. Darlington & Talbot (1898) were interested, for example, in testing whether music could be considered a distractor and used a weight-lifting task to assess the impact of music on performance. (Results suggested that, no, music

was not a "distractor".) Another example was Tinker (1925) who studied whether the presence of bells ringing intermittently during an intelligence test would affect performance of the test, a marker that distraction hinders performance. (On average, it did not.) Yet, from the start, the early intuition that distraction ought to hinder performance was there and a significant debate arose as to what ought to be called a distractor (see Dulsky, 1932, *Psychological Review*). In fact, early efforts to study attention using distractors often failed "because it was impossible to obtain a satisfactory distractor" (Dulsky, 1932, p. 590). That said, what did come from those early years was a new way of talking about distractors: "distractor" is not a label to be used in relation to attention (or as a means to study *that* abstract concept), but rather in relation to performance. The rationale is simple: given that we don't know what attention is, but we do know how to measure performance, it is best to simply define distractors as those stimuli that hinder performance (Dulsky, 1932).

The goal of this chapter is to challenge this conception of "distractor," which is closely related to the modern use of this term. Our hope is to highlight the problems with this term, the confusion that it engenders and, hopefully, provide a way out of this state of affairs by creating a more modern context in which we can (a) redefine the study of distraction; (b) clarify the terminology that is used to study attention *and* distraction; and (c) identify current sources of confusion in the literature that are very much related to the confusability between the labels "distractor" and "distraction". We will begin by presenting a literature where the term distractor is heavily used (viz, visual search) in spite of the fact that the goal of this work is not to study distraction at all (but in fact attention). Along the way, we will argue that not all "distractors" are created equal and, thus, one must be careful when trying to extend the findings from laboratory visual search to real-world search tasks. Next, we will quickly discuss (and dismiss) an example of the use of the term "distractor" within the literature on redundancy gains. Then we will describe a relatively more nuanced case study regarding the perils of the use of the label "distractor": the Flanker Effect literature, a domain in which this term has gained a very prominent role. From this section, we will conclude that it is in fact *inappropriate* to make claims about the phenomenon of "distraction" from experiments using modern-day "distractors," in spite of our field's tendency to want to make such claims. One of our goals, too, is to differentiate the concept of "distraction" from that of "distractor interference," which are actually quite different phenomena.

We will next discuss the inattentional blindness paradigm, which might be a better way of studying and conceptualizing distraction than the previous paradigms. We will also describe data from an experiment using this methodology to highlight why it is so critical to acknowledge the difference between distractor interference and actual distraction. We will argue that distractor-interference effects reflect, at a basic level, an important degree of *success* by attentional selection (rather than a failure, as it frequently claimed). Even more, we will argue that the presence of a "Flanker Effect" should be taken as evidence in favor of early selection models of attention, rather than as evidence for late selection models, as it is commonly done.

The next section of this chapter will focus on a proposed new task to study distraction and distractibility using an eye-tracking methodology. We will present data from our laboratory showing that heightened cognitive focus reduces distractibility. And we will conclude by presenting, in rough strokes, the skeleton of a model of attention that incorporates the properties of attention highlighted in this chapter, as well as other factors like motivation and the need for inner focus when completing complex cognitive tasks.

2. WHEN DO "DISTRACTORS" CAUSE DISTRACTION?

2.1. Visual Search

2.1.1. Effect of Distractors on Search Performance

To answer the question of *when do distractors cause distraction?* we will begin with the most popular paradigm that uses the first of these terms: namely, visual search. Another reason for doing this is that attention is often thought of as being the opposite or the antidote to distraction, and the concept of attention has played a major role in our understanding of visual search from the very beginning.

In the context of visual search, the term "distractor" is used to refer to a majority of the elements within the display, and was first used by Shiffrin and Gardner (1972). The collection of all distractors (plus the target item, if there is one) is referred to as *set size*. Visual search is typically studied by measuring response time (RT and/or proportion correct) as a function of set size. In other words, distractors are a set of a priori possible targets that one must inspect in order to find the target in the display. We draw inferences about the workings of visual attention by inspecting the reaction time × set size function (e.g. Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). Notice that the distractors are stimuli created by the investigator to probe how attention works. Therefore, in most laboratory tasks

concerning search, there is a certain degree of (purposeful) overlap between the distractors and target stimuli (e.g. all items share a common shape, or vary within a small set of colors, orientations, etc.), because the goal is to induce a *measurable* effect on performance, and performance, itself, is defined as a function of set size (i.e. the number of distractors). So, if one defines "distraction" as either "attention is being allocated to a nontarget stimulus" or "an effect on performance that is due to distractors," then, yes, visual search would provide data that are relevant to the study of distraction. But, as we will argue, what people really mean by "attention" in the context of visual search is very different from the kind of attention that is needed to avoid being distracted. And what people really mean by "distraction" is very different from the kinds of effects that distractors have on response time in visual search. But before discussing what can be learned about distraction from cases of visual search where distractors do have an impact on performance, we will address a related, but separate issue, regarding how the absence of such effects have been interpreted, and how we believe that they ought to be interpreted, in their relation to the concept of attention.

2.1.2. *When* Distractors *Do Not Impact Search Performance*

Most theories on attention that rely on visual search experiments indeed focus most of their explanatory efforts on accounting for performance when set size matters (e.g. Bundesen, 1990; Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). In fact, whenever the number of distractors is found to have no effect on RT (i.e. so-called flat search slopes), the experimental results are brushed under the attention rug: clearly, if performance was unaffected by set size, then attention must have not been involved in the processing of the display and the detection of the target. In fact, whenever we find a case where performance is not affected by set size, we refer to it as being the result of "preattentive" processing. Color pop-out is a fine example of this: if asked to look for a red apple among green ones, the number of green apples does not influence RT to find the red target: the red apple "pops out" from the green ones and we introspectively believe that our attention automatically moved to the red target apple. But did attention really have nothing to do with us finding that red apple? Evidence from at least two smaller literature within the attention field would argue that this conclusion is misguided: the contingent-capture literature (e.g. Folk, Leber, & Egeth, 2008; Folk, Remington, & Wright, 1994) and the priming of pop-out literature (e.g. Fecteau, 2007; Maljkovic & Nakayama, 1994, 1996). For now, simply consider the following data. In Wan and Lleras

(2010), subjects were asked to detect the presence of a color oddball item (i.e. the one item that differed from all others in color). Two different sets of color pairs were used. In one block of trials, participants were tested with red and green items, while, in a different block, the items were pink and fuchsia (two colors closely matched in luminance). Both color pairs produced visual pop-out: in both cases, RTs were unaffected by set size (3, 8, or 12), with search slopes of −0.09 ms/item for red among green and 1.19 ms/item for pink among fuchsia, neither of which was different from zero. So far so good: the standard interpretation would be that these pop-out discriminations were computed in preattentive fashion by the visual system. Yet, there is a kink with this interpretation. Overall, responses were much slower with the pink and fuchsia stimuli (mean RT = 698 ms) than with the red and green stimuli (565 ms). How are we to interpret this 133 ms increase in mean RT? At what point(s) in processing was this rather large delay added? And what were observers doing during this time? Were they waiting for preattentive vision to finish processing, without attending to the display? This seems implausible given other performance benchmarks of the attention system. For instance, saccades to sudden onsets that match our attentional template can be triggered extremely quickly, in the range of 130–150 ms (Hollingworth, Matsukura, & Luck, in press; Kirchner & Thorpe, 2006). Not only it is difficult to believe that participants would have an added processing difficulty in "preattentive" vision of a magnitude of 133 ms (during which attention would not be involved), but the overall magnitude of the RTs is also equally difficult to interpret. Preattentive vision is supposed to be "fast and automatic;" the sort of stuff that gets done "before we know it." How can this type of processing produce *detection* RTs on the order of 700 ms? Are we to infer that preattentive vision can sometimes be that slow? No, we would argue, even though this answer is at odds with the prevailing view in the field on how to interpret results from visual search experiments (e.g. Duncan & Humphreys, 1989). Attention is involved in finding a pop-out target, and some pop-outs are, in fact, rather hard to find.

Going back to the central issue of what distractors are and how they are used, it is clear that we (as vision scientists) have a very poor model for what goes on with attention when distractors fail to have a measurable effect on performance. This area of potential research has gone unrecognized for the most part, and is still in its infancy (with some exceptions, Purcell et al., 2010; Schall, Purcell, Heitz, Logan, & Palmeri, 2011; Townsend, 1972; Tseng, Glaser, Caddigan, & Lleras, submitted for publication; Wenger and Townsend (2000), who earlier recognized the perceived need to use

different tools to analyze presumably parallel forms of processing). Yet, this is not surprising given that one of our principal tools to index attention *at work* has been the manipulation of set size. For instance, a very influential model of visual search, Similarity Theory (Duncan & Humphreys, 1989), is predicated on the idea that preattentive vision precedes attentive vision in time. Preattentive vision is parallel and capable of producing null effects of set size, but, almost by definition, is uninteresting for studying attention per se. The main thrust of the theory is in interpreting performance when distractors do affect RTs, because one can infer attentional processing times: for instance, the more similar targets are to distractors, the more complex the attentional involvement is in the search task (e.g. larger inspection times per item). So, the logic goes, the more that distractors impact RTs, the more attention is involved in determining performance. This assumption regarding the normal processing steps in vision (encoding, in some parallel, preattentive form, followed by identification, in a more serial, attentive form) is common to most theories of attention, both formal (e.g. Treisman & Gelade, 1980; Wolfe, 1994; Wolfe, Vo, Evans, & Greene, 2011) and computational ones (e.g. Itti & Koch, 2001; Navalpakkam & Itti, 2007; Peters, Iyer, Itti, & Koch, 2005; Townsend, 1972; Wenger & Townsend, 2000). But the lack of a set-size effect does not imply a lack of attentional processing. If a set-size effect implies attentional processing (A ergo B), it is an error in logic to argue that the absence of a set-size effect implies an absence of attentional processing (not A ergo not B). Those are not logically equivalent propositions. What we can be sure of is that, if attention is not involved, then one can never expect a set-size effect to emerge (not B ergo not A). That would be correct, but it is not the argument being made in these cases.

2.1.3. The End of Preattentive Vision

How can we understand attention when our principal tool of inquiry produces no results? Understanding attentional processing in pop-out situations will require newer methodologies (new computational models and new brain imaging techniques). But it is important to realize that it is a logical flaw to argue that because our tool (set-size manipulation) failed to influence performance, that attention was not involved in producing that performance. Thus, we cannot brush aside all forms of "parallel" searches under the preattentive rug. In fact, there is overwhelming evidence that *preattentive vision is dead*. And we are not the first ones to say so (e.g. Di Lollo, Kawahara, Zuvic, & Visser, 2001, "The pre-attentive emperor has no clothes", *JEP:General*). The general idea that there is a temporal order to visual processing whereby

some processing always occurs prior to the allocation of attention is likely wrong. Biased competition theory (Desimone & Duncan, 1995), for example, suggests that attention can bias processing to favor one type of stimulus (or property) over another at an extremely early point. And, thus far, the evidence in favor of this proposal is quite strong. In fact, evidence suggests that top–down attention can bias neural processing before the onset of a visual stimulus, which can result in extremely fast eye movements (Hollingworth et al., in press; Kirchner & Thorpe, 2006), as well as produce parallel feature-based selection over the entire visual field (Bichot, Rossi, & Desimone, 2005; Serences & Boynton, 2007). Even more, thanks to fMRI, we now know that every possible layer of visual processing in the brain can be modulated by attention, not only at the cortical level (reference V1 and V2, which are now routinely shown to be modulated by attention), but even subcortical structures like the LGNshow modulations by attention (e.g. Kastner et al., 2004; O'Connor, Fukui, Pinsk, & Kastner, 2002), that entail as complex a computation as increasing the precision encoding of attended objects while completely filtering unattended objects (Fischer & Whitney, 2012).

Thus, given our understanding of attention, not only within the framework of biased competition, it stands to reason that (other than the processing in the retina) there is likely no strict temporal order to visual processing such that some degree of visual processing always occurs prior to the allocation of attention. Even the initial levels of visual processing are likely modulated by attention, under the right circumstances (e.g. Hillyard & Munte, 1984). That is not to say that there is not a lot of processing that goes on outside of the focus of attention. Jeremy Wolfe's most recent model of vision now includes a "nonselective" pathway in vision: a parallel route of visual processing that proceeds independently from attention and that is responsible for things like computing the sensation that the visual world is more than the object being attended, as well as extracting visual regularities from the world (Chong & Treisman, 2003, 2005a, 2005b; Choo & Franconeri, 2010; Haberman & Whitney, 2011). Whether there is such a pathway is a matter for debate, but what is clear is that there is a lot of *unattended vision* that goes on at any given point in time, and proposing that this form of processing is taking place simultaneously to *attended vision* is also likely correct.

The distinction between unattended and attended vision is, actually, critical for us to begin to form an understanding of "distraction". We can use a definition of distraction that follows our common intuition: distraction is when our current train of thought gets interrupted by a secondary train of thought, against our wish. One can study *cognitive* distractions: the occurrence

of distracting thoughts, which can happen in a benign manner (as in the task-unrelated thoughts literature, Giambra, 1989; McVay & Kane, 2009, 2012; Smallwood, Obonsawin, & Heim, 2003) or even pathological manner (as in Obsessive compulsive disorder, Depression and Anxiety). However, we are here interested in perceptual distraction: when the occurrence of an event in the world comes to derail our train of thought or interfere with our task at hand. More specifically, we will use the domain of vision to understand this concept. In our terminology, *distraction* can be defined as times when a visual event that was initially processed by unattended vision forces the attention system to orient to it. This reorienting must occur at the cost of disengaging attention from whatever other object (in vision or thought) that was previously being attended. We will return to this definition after presenting theory and methodology that is currently used to assess "distractibility" (the study of distraction) in our field. But first we must dispense with visual search.

2.1.4. Nothing about Distraction Can be Learned from Visual Search Experiments

If we return to the question of what the use of distractors can tell us about distractibility in the domain of visual search, the answer is clearly *nothing*. This conclusion is, hopefully, not very controversial. Distractors are a group of stimuli that are, by design, meant to be inspected by attention because of their resemblance to the target stimulus. That is, it would be awkward to argue that seeing a red hat in my closet is a distraction from my search of a red scarf in said closet. If one has set a priority for detecting red items in the closet, seeing and evaluating a red item is not an episode of distraction, but is an integral step in achieving my goal. I want to inspect *all* red items with the hope that one of them is my scarf. Thus, visual search, in the traditional form, is not a methodology suited for studying or understanding distraction. In fact, it has been very difficult to translate what we know about visual search from our vision laboratories to visual search in the real world, precisely because in the real world, not all objects in a scene are a priori meant to be possible search targets. If we are looking for a fork in a kitchen, it is irrelevant how many appliances are in that kitchen. No one placed the refrigerator, oven or dishwasher in the room with the hope that people looking for forks would look at them. These items are not targets for the attention system, thus cannot be counted as distractors in the laboratory sense; however, if asked to identify all objects in a kitchen, most observers would likely count all appliances therein as being objects in the kitchen scene. Labeling all objects in a real-world scene as a priori targets

of the search (i.e. as distractors) will always produce an overestimate of set size and may lead to the awkward conclusion that real-world search is more efficient than laboratory search. If one has a black-and-white styled kitchen, with all black-and-white utensils, appliances, and counters, searching for a red kitchen towel would produce a very efficient search (unaffected by the number of elements in the scene), whereas searching for a specific black utensil will likely be highly dependent on the number of objects in the room. And what we have learned about guidance in the laboratory is absolutely applicable to this real-world scene. It is merely complicated because every type of search target in a real-world scene determines the set size in that scene. In other words, the set size in a scene cannot be (and should never be) determined in the absence of a definition about what the search target in that scene will be. Once such a definition is provided (e.g. "look for a couch"), one can determine the subset of elements in the scene that will be a priori interesting to the attention system (e.g. all furniture items). We propose that we should use the term "candidates" to refer to this specific variety of distractors: the subset of items in a scene that might be a target. Again, candidates can only be defined once the target of search is known. Once a target of the type *furniture* has been defined, the number of windows, paintings, and toys on the floor are immediately discounted by that definition. Those objects could never be targets and will not be inspected in the search, so they are not what we here called "candidates". Very promising work in this arena has been recently conducted by Zelinsky et al. (Alexander & Zelinsky, 2011; Neider & Zelinsky, 2011) as well as Wolfe et al. (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). Note that, by design, most of laboratory visual search tasks have situations where the set of candidates is the number of objects in the scene (what people have referred to as set size), whereas in the real world, the number of candidates is often substantially smaller than the number of objects in the visual scene.[1]

[1] A notable exception is the literature on the phenomenon known as Visual Marking (Watson & Humphreys, 1997). In this phenomenon, it is typically found that previewing half of the search items (say all green stimuli) for some time before the second set of search items (say, the blue items) is presented produces search efficiencies identical to those that are obtained when participants only see and search through the second set of search items (i.e. only blue items). Interestingly, this improved search efficiency in the "preview" condition is always accompanied by a substantial main effect, in the order of 50–150 ms. This visual marking effect is critically dependent on attention being allocated to the previewed items (Watson & Humphreys, 1997), and we would argue that it reflects the process of candidate selection (selecting all blue items as search candidates, while ignoring green items) and is central to this paradigm. In spite of its importance, most of the research on visual marking has focused overwhelmingly in the analysis of slopes, ignoring all intercept effects.

Definition

Candidates: A candidate is a stimulus that contains at least some of the defining attributes of a target. In both visual search and many selective-attention tasks, a candidate is an object that closely resembles the target in terms of its event-like descriptive characteristics. Does it belong to the same visual category of items than the target? Is it visually similar to the target? Does it have the same temporal characteristics as the target (i.e. it appears/disappears at the same time as a target)? Does it appear nearby or at a possible target location? As a result, which elements of a display are candidates can only be determined once a task is defined and the target of the attention system is determined. In visual search, the number of candidates is what determines the *functional* set size for a scene. Candidates are always highly relevant to the search task because of their status as "potential targets". Candidates do not have stimulus–response associations in a task, but they can nonetheless inform responses. For example, in a present/absent visual search task, every inspected and discarded distractor is further evidence that the response on the trial is more likely to be "absent". See Bundesen (1990) for a similar definition regarding relevant distractors: those that are similar to the target along the defining characteristics of the target.

2.2. Divided Attention

A very similar point can be made for the task that is often used to study divided attention—viz, the redundant-targets detection task (e.g. Miller, 1982; Mordkoff & Yantis, 1991). In these experiments, subjects are asked to respond to prespecified targets (usually letters), just as they are when doing visual search, but displays are always quite small (e.g. two letters at most) and some displays contain multiple instances of the same target. The key result from these tasks is the response-time advantage for trials with two targets, instead of only one, which is known as a "redundancy gain" (Raab, 1962).

What makes this task interesting here is how some researchers have interpreted the finding that a single target accompanied by a distractor can produce slower responses than a single target presented alone (e.g. Grice, Canham, & Boroughs, 1984; Grice & Gwynne, 1987). The slowing in the former case was said to be due to a "distraction decrement;" that is, the nontarget in the second display location was said to distract the observer by pulling resources away from the target. As it turned out, however, the slowing was actually caused by the nontarget being correlated with the absence

of targets (Mordkoff & Yantis, 1991); when these correlations were eliminated, single-target trials produced exactly the same RTs whether or not a nontarget was also included within the display. In other words, there was no actual decrement in performance due to the nontarget item; the slowing of RT was caused by the information carried by this item, instead. Even more, we would here argue that this nontarget should not really be thought of as being a distractor, because it was a candidate, due to being in one of the very few locations in the entire display at which targets ever appear.

The second way in which this work might be relevant to the study of distraction is in how a distinction was made between nontargets that sometimes appeared with targets, which were called "distractors," and nontargets that never appeared with targets, which were called "noise". The main reason for keeping these two types of nontarget stimuli separate was to be able to calculate separate correlations between their presence and the correct response (Mordkoff & Yantis, 1991); these calculations were crucial to the demonstration that the so-called "distraction decrement" does not actually exist. But the use of two different labels for items that, on the surface, play identical (non-)roles in the task also shows that the idea that some nontargets are more equal than others is not completely new. When one adds in that these nontargets can carry different amounts of information—in the technical sense of the word—a few doubts should be raised about the use of the label "task-irrelevant" to refer to nontargets. But this is probably better explained in the context of the task to be discussed next.

2.3. Flanker Effect

2.3.1. Traditional View: The Flanker Effect Implies Late Selection

One can also study attention in the absence of location uncertainty, that is, when one knows where the target is going to be (e.g. Eriksen & Eriksen, 1974), yet there is uncertainty about the identity of the upcoming target. This experimental procedure has come to be known as the *Flanker Paradigm* and is typically a variation of the following set up: target position is defined a priori and participants are asked to report as quickly as possible the identity of the target, almost always a letter. Just as in visual search, the target is presented among "noise" elements that may be related to the target in some way. For example, in the original experiment, the noise elements could resemble the target in terms of their low-level features (target and noise letters looked alike) or in terms of response associations (noise letters were letters picked from the possible set of target letters). As in the visual search literature (Estes & Taylor, 1966), over the years, in this paradigm,

the term "noise" element was replaced by "distractor" (or sometimes just "flanker" because the distractors tend to be presented on both sides of the target letter, flanking it). In other words, in this context, one uses the word "distractor" to refer to target-like stimuli (most often, exact copies of the possible targets themselves) that are presented at locations that are near to the target. The goal of this paradigm is to measure the degree to which attention can focus solely on the target item itself (a test for what is referred to as "early selection", Neisser & Becklen, 1975; Sperling, 1960; Treisman & Geffen, 1967). If performance is unaffected by the presence or identity of the distractors, then one can conclude that selective attention successfully "filtered out" the "irrelevant" information from the display. In most cases, however, it is found that response times are strongly modulated by the identity of these flanking distractors (e.g. evidence for so-called "late selection", Deutsch & Deutsch, 1963; Norman, 1968), an effect known as the Flanker Effect. This result is interpreted as evidence that the distractors have been processed to a sufficiently deep level of processing (at least to the level of stimulus identification) such that they, too, can activate response selection processes. In fact, the Flanker Effect is most often measured as the difference in RTs between two conditions: one in which the distractors activate the same response as the target for the trial (*compatible* or *congruent* condition) and one in which the distractors activate a different (incorrect) response from the target (*incompatible* or *incongruent* condition), with responses on incongruent trials typically being anywhere between 30 and 100 ms slower than on congruent trials. When a Flanker Effect is obtained, it is generally concluded that selective attention has failed in some way, allowing task-irrelevant items to be processed.[2]

2.3.2. The Flanker Effect is Not a Measure of Distraction

The Flanker Effect is an example of *distractor interference*, and it is not the only one. The Simon effect (Simon & Rudell, 1967) and the Stroop effect (Jaensch, 1929; Stroop, 1935) are also well-known examples of distractor interference. What these effects have in common is that one can consistently find evidence that some aspect of the distractor stimulus (or the distractor attribute) influences behavior. But is distractor interference a measure of distractibility? Hardly so. If we follow the logic presented above

[2] It should be noted that, with regard to the presence or absence of a Flanker Effect, the same logical error is being made as is for visual search. While the presence of a (significant) Flanker Effect is evidence that the flanking distractors were, indeed, processed, the absence of a Flanker Effect does not necessarily imply that the flankers were not processed. This point was made rather forcefully by Driver & Tipper (1989).

in the context of visual search, the Flanker Effect is no different. Distractors in this task are selected by the experimenter with the hope that, if they are processed, they can have a measureable effect on performance. To achieve a measureable effect on performance, the critical distractors in the flankers task are, in the vast majority of studies, exact replicas of one of the possible target stimuli. The rationale is simple: given that participants have a response associated with every target, we can hijack those stimulus–response associations to measure distractor processing. So, all we need to do is use targets (or target-like stimuli) as the distractors. Participants will know that they are not the target on any given trial because targets and distractors are always presented at different locations. Yet, if participants process those distractors, they will activate those stimulus–response associations and compete (with the information coming from the actual target) to determine the response on the trial. In sum, congruent and incongruent distractors in a flankers task are *by design* stimuli that fit the participants' task set. For example, if the participants' task is to identify whether a central letter is an X or an O, the critical distractors in most designs of a flankers task would be either Xs or Os. Thus, if one assumes that participants have a task set that stresses the importance of Xs and Os, it would be difficult to come up with more task-relevant stimuli than congruent and incongruent distractors (i.e. Xs and Os). Therefore, given that the critical distractors in this task are actually very much task-relevant, they cannot tell us *anything* about distractibility.

2.3.3. The Information Processing Tradition and the Flanker Task

Overall, this simple analysis of the flankers task reveals that most of the distractors used in these studies are by design task-relevant and they should be understood as such. However, there has been a tradition in this literature, going back more than 25 years, to describe the Flanker distractors as being *task-irrelevant*, instead (e.g. Lavie, Hirst, de Fockert, & Viding, 2004; Miller, 1987; see also Figure 7.4). To understand why, one must understand the theoretical tradition within which the initial studies on this topic were conducted. The dominant methodology in the mid-twentieth century experimental psychology was the detection of target signals embedded in noise (often interpreted in the terms of signal-detection theory). In fact, early studies on visual search refer to the distractor elements in those displays as "noise" elements (e.g. Estes & Taylor, 1966), just as Eriksen and Eriksen described them in their first report of the Flanker Effect. Furthermore, information theory was the theoretical backdrop to

cognitive psychology. One viewed and studied the mind as an information processing problem, with an input signal that underwent several discrete transformations (or processing stages) inside the black box that was the mind before producing an overt response. Within this worldview, the "noise" stimuli used in tasks like the Flankers paradigm were irrelevant to the observer because they carried no information regarding which response should be made. Thus the term was born of distractors being task-irrelevant: the word "task" being a short-hand for "which response should be made" and the word "irrelevant" indicating that the distractor carried no information with regard to the task. In sum, the identity (or even presence) of distractors in a visual search or flanker task changed nothing to the fact that only the target signal determined the appropriate response on each trial. If, for example, the target on a given trial was the letter X, the response on that trial was to press the button associated with that specific letter, irrespective of any and all other noise letters in the display. The noise elements were irrelevant to the actions of the participant because *they carried no information about the required response on any given trial*. In that sense, perhaps a better label to qualify the distractors in a visual search or flankers task would be to say that distractors are *response uninformative*.[3]

A great deal of confusion would be avoided if a more precise term like this one was used, instead of the more misleading label task-irrelevant that has now come to be the dominant term for describing flankers. Another alternative would be to simply drop this qualifying term altogether from this literature. Distractors would simply be referred to as distractors, with no mention of their task relevancy, given that it is clear that they are task-relevant, in the way that we now use these terms. Here, however, we propose the use of the term *foil* to refer to congruent and incongruent distractors. It implies, as it should, that they were put in place in the experiment with the goal of somehow tricking or interfering with the participants' responses: in a Flanker task, foils are a close match to what participants are looking for. An exact copy, but just off in some minor way (like their location).

[3] Note that not all stimuli need be response informative to be considered task-relevant. Candidates are one such example: in visual search tasks where participants must report the identity of the target, candidates themselves are uninformative to the trial's response. A spatial cue in a Posner-type cueing paradigm is also an example of a task-relevant but response-uninformative stimulus: it tells observers where they might find items that contain response-related information. But the cue itself is not informative to the response.

Definition

Foil: A *foil* is a type of candidate that can cause a participant to produce the wrong response in a trial because foils have links to potentiated responses in a task. They often contain the same response-defining attributes as the set of targets. In a selective attention task, a foil is often an object that is identical to one of the possible targets but is not the target itself because of a priori reasons: it differs from the target along one (or several) defining characteristics. For instance, the target might be determined by its unique location, or the foil itself might have a unique prespecified location. A foil may also appear at a different point in time in a trial than the target (as in priming experiments). Foils are automatically selected by the attention system, to the extent that they are perceptually well represented. Selecting and rejecting foils is a critical part of what attention must do in order to properly respond on a given trial.

2.3.4. Why are Flankers Referred to as "Task-Irrelevant"?

But, why did the authors in this literature feel that they had to specify that, in their design, distractors were indeed task-irrelevant? That is because not all distractors in a flanker task need to be response-uninformative, though most are. Several papers have documented the "correlated Flanker Effect" (e.g. Miller, 1987; Mordkoff & Halterman, 2008): that is, one can, by design, pick a combination of target–distractor co-occurrence frequencies such that the identity of distractors can indeed tell the participant something about the response. For instance, say that the task is to identify a target letter (an A, B, Y, or Z, with A and B mapped to one response and Y and Z mapped to the opposite response), and that on 75% of the trials when the flankers are Ms, an A or B is the target, and on 75% of the trials when the flankers are Ns, a Y or Z is the target. The distractors in this task have no direct association with either response and they do not resemble the targets. Yet, their identity carries information about the likely response. As one can imagine, responding A/B on an A trial is faster and more accurate when the distractors are Ms than when they are Ns. This is the correlated Flanker Effect and it can also be observed in more subtle ways whenever there are contingencies in the design of the flanker task such that the identity of the distractors provides useful information of any sort (Mordkoff, 1996). In such cases, the distractors are response-informative and, as such, can be characterized as "task-relevant" by the older definition of task relevancy. So, it is easy to

see why there was a need in the literature to differentiate between these two forms of Flanker Effects: the one that arises when distractors carry some information about the response (by design or by accident), which is the correlated-flankers effect, and the one that arises when the distractors carry no information about the response, but have an effect because they are the same stimulus as one of the actual targets, which is the traditional Flanker Effect.[4]

In a somewhat unfortunate turn of events, contemporary cognitive psychology has continued to use the original "task-irrelevant" modifier to describe distractors in a flanker task. It is unfortunate because most of us have moved beyond that use of the term "task relevancy" and use a more contemporary definition of relevance that implies a very different understanding of the role of distractors in the flankers task and their relationship to the study of distraction. The newer use of the term "task-relevant" arises from research on phenomena like attentional capture, task switching, and executive control. In attentional capture, for instance, people study what aspects of a stimulus make the stimulus more or less "attractive" to the attention system. Folk et al. (1994) demonstrated the phenomenon of *contingent* attentional capture: that a visually salient element captures attention to the extent that some of its defining characteristics match those characteristics that define the target stimulus. A moving stimulus draws attention to itself to the extent that the attention system is looking for moving-like things (e.g. targets that are moving or have sudden onsets, or some other stimulus property that also characterizes motion). These *attentional control settings* are determined by the task demands, and, some would argue also by the *behavioral urgency of certain stimuli* (e.g. Franconeri, Alvarez, & Enns, 2007; Lin, Franconeri, & Enns, 2008; von Mühlenen & Lleras, 2007).

In sum, we now use the term "task-relevant" to describe the level of overlap between distractors and targets, either in terms of their visual characteristics or their functional description with respect to the task characteristics (Are they both sudden onsets? Do they appear in tandem? Does one precede the other? At the same or different locations? etc. See also Yantis & Egeth, 1999).

[4] It is worth noting that, by the information theory-based definition of task relevance, all the nontargets in nearly every published visual search study have actually been highly relevant, since each item that is identified as not being a target increases the likelihood that any given not-yet-identified item is a target. Likewise, the total number of items in the display is equally relevant, because, as more items are included, the odds of any particular item being a target are reduced. In other words, reverting to the old definition of task relevance is not going to change our earlier argument that the nontargets in a visual search display are not a source of distraction, even when they cause responses to be slower.

To sum up, in our terminology (and in agreement with the attention capture literature), candidates are always task-relevant. So are foils. In a later section, we will expand on the theoretical and applied consequences of using the term "task-irrelevant distractors" in a flankers task, but for now, let us finish on the following point. Congruent and incongruent distractors in a flanker task are not task-irrelevant by the current definition of task relevancy because they are exact replicas of the target and resemble the target in every possible functional way, save for the fact that they appear at a different location than the target. Thus, as with visual search, nothing about distraction and distractibility can be learned from studies on the Flanker Effect. Of course, by the same logic, one can also argue that little about distractibility can be learned from studies on attentional capture, because, in that paradigm, distractor stimuli most often fall clearly within the "task-relevant" category of stimuli. Therefore, we are going to have to turn to a very different kind of task if we hope to learn anything about actual distraction.

2.4. A Different Form of Distractor: The Inattentional Blindness' "Critical Stimulus"

2.4.1. The Unexpected Event Paradigms

Most often, the phenomenon known as "inattentional blindness (IB)" is studied with a paradigm that is a variant of the original experiments by Mack and Rock (1998): participants are asked to do a fairly engaging visual discrimination (e.g. deciding which of the two arms of a cross is longer, when the two arms are actually almost identical in length) in a brief period of time. After the participants have completed a few trials of this task, experimenters present an unexpected "critical stimulus" within the display: an unexpected visual object that is completely unrelated to the task at hand. It can be, for instance, a small square near fixation. After participants complete the response on this critical trial (also known as the "inattention trial"), experimenters ask the subjects whether they saw anything unexpected on that last trial. Henceforth, we will refer to this overall experimental framework as the *unexpected event paradigm*. The term "inattentional blindness" refers to the finding that quite often, people fail to report the unexpected stimulus.[5] Importantly, after the inattention trial, participants complete a few additional trials, including a final trial in which they are told to completely

[5] The label "inattentional blindness" was applied prematurely and is rather unfortunate, in that it strongly suggests that the unexpected stimulus is not seen. Subsequent work (e.g. Moore & Egeth, 1997) has indicated that the unexpected stimulus is, indeed, perceived, but never reaches awareness and/or is quickly forgotten. But we are stuck with the name for the paradigm.

ignore the stimulus for their primary task (i.e. the cross in our example) and just focus on trying to detect any other objects in the display. In this last trial (referred to as the full attention trial), participants are typically at ceiling at detecting the critical stimulus. This is important because it guarantees that the critical stimulus is actually perfectly visible. Furthermore, it allows the experimenters to conclude that the reason why participants failed to be aware of the critical stimulus was that they were completely immersed in processing a different object (the cross) at the time it appeared.

This is not the only paradigm in which inattentional blindness has been studied. Neisser had earlier pioneered the use of video to study the same phenomenon (Neisser, 1979; Neisser & Becklen, 1975) asking participants to attend to one of the two superimposed movies. Participants were, as one would suspect, quite unaware of unexpected events in the unattended film. The video paradigm was also popularized recently by Simons et al. who demonstrated that one does not need two superimposed videos to create inattentional blindness: grabbing a cue from Mack and Rock (1998) and Simons and Chabris (1999) showed that simply asking participants to be intensely engaged in some attentive task (like counting the number of bounces and passes of a basketball) would trigger inattentional blindness to unexpected events occurring in that same video. The added benefit of this technique is that unexpected events can be prolonged in time (the time it takes a gorilla to walk across a scene, and pump its chest a few times), as long as they start and finish while participants are engaged in the attentive task. Follow-up experiments also showed that the likelihood participants will become aware of the critical stimulus increases the more the critical stimulus resembles the events in the attentive task. For instance, people counting bounces and passes in a team of people wearing black t-shirts will be more likely to detect the appearance of a (black) gorilla, than those counting passes in a team wearing white t-shirts (for a similar demonstration, see Most et al., 2001).

In this sense, unexpected event paradigms do hold promise for the study of distractibility. Why? Because they allow one to ask the question: will an observer's train of thought be interrupted by a secondary event, against the observer's own goals to concentrate on a primary task. In fact, the unexpected event paradigm has inspired a number of applied investigations into issues relating to distractibility. In the driving domain, Strayer et al. (e.g. Strayer & Drews, 2007) have proposed that talking on a cell phone produces a form of inattentional blindness (to the extent that people overly concentrate on the cell phone conversation) such that participants will be

less likely to remember or identify objects that they actually fixated during the driving task, if they were talking on the cell phone when those objects appeared. In other words, the work of Strayer et al. suggests that the additional cognitive load involved in having a cell phone conversation creates a situation where participants are less likely to attentively experience the objects that their eyes fixate during driving. This is perfectly in line with Mack and Rock's initial account of inattentional blindness but extends that account to attentionally engaging events outside of the realm of vision. That is to say, whereas one could have interpreted the initial findings of inattentional blindness as reflecting a limitation of *visual* attention, the findings in the applied literature suggest that they are more general than that. We do not fail to see the critical stimulus in the unexpected event paradigm (or the gorilla in Simons & Chabris' video) because our visual attention cannot be focused on two visual aspects of the scene at once. Rather, we fail to see the critical stimulus because our attention system (irrespective of modality) has a difficult time doing two things at once: processing the information required in the demanding primary task, as well as encoding and responding to the unexpected stimulus.

2.4.2. Recruitment of "Central" Resources and the Ensuing Blindness

Elegant evidence in support of the idea that a "central" limited resource is responsible for inattentional blindness was found by Fougnie and Marois (2007). They asked participants to do a task where they had to either rehearse in memory a set of letters (the maintain condition) or reorganize them in alphabetical order (the manipulate condition). Shortly after the letters were presented to the participants, an unexpected critical stimulus was presented in the display. Only 35% of participants failed to detect the unexpected stimulus in the maintain condition, whereas 68% of participants in the manipulate condition missed it. There was no difference between groups in the full attention trial and participants were near ceiling at detecting the critical stimulus on that trial. This evidence suggests that the degree of involvement of central executive resources in a primary task determines the degree to which participants will become aware of unexpected and a priori unattended events in the world. In a clever follow-up experiment, Fougnie and Marois presented the critical stimulus toward the end of the retention interval, at a time where participants in the manipulate condition would have already finished reorganizing the letters in memory. In this experiment, there were no differences in the degree of inattentional blindness across groups, presumably because the central executive was no longer

actively engaged by the reorganization task (i.e. by the time the Critical Stimulus appeared, participants were just rehearsing the ordered letters in memory).

Overall, this brief (and admittedly selective) glance at some of the inattentional blindness literature suggests that distractibility may depend on the degree to which we are currently engaged in a centrally demanding task: the more engaged we are in that task, the less *distractible* we are by unexpected events (see also Macdonald & Lavie, 2008, 2011). To touch back briefly on an issue raised in the section on visual search, all evidence suggests that visual processing proceeds normally even in the absence of attention. Visual processes like perceptual grouping and surface completion (and others) are all performed even in the absence of attention, as operationalized by the occurrence of inattentional blindness (e.g. Moore & Egeth, 1997; Moore, Grosjean, & Lleras, 2003). This is important because it underscores the need for revisiting some visual attention theories. We can be fairly confident that (1) vision unfolds with or without attention, which is the reason we have a rich phenomenal experience of the visual world even though we cannot attend to it in its entirety at any given moment; (2) attention can access any level of visual processing to bias processing in a particular direction; and (3) therefore, there seems to be no need to assume or propose some form of uniquely preattentive processing that must precede attention in time. There are no proto-objects (Rensink, 2000); there is no massive parallel preattentive vision (Treisman & Gelade, 1980). We believe that any model of attention that assumes so is likely misguided.

Returning to the issue of distraction, the question is: why do we experience inattentional blindness in the first place? What factors determine whether we do? Is inattentional blindness a failure of some sort or might it reflect an efficient, calculated tradeoff, instead? These are complex question and we will attempt to offer a possible answer to them in Section 3 of this chapter. But first, let us reexamine the issue of task relevance in the stimulus, using a simple inattentional blindness paradigm.

2.5. Empirical Study: Comparing the Salience vs the Relevance of a Distractor

There has already been evidence in the inattentional blindness literature that the degree of visual similarity between the critical stimulus and the stimulus in the attentionally engaging task determines the degree of inattentional blindness to the critical stimulus (Most et al., 2001; Most, Scholl, Clifford, & Simons, 2005). Here, we just want to reiterate this point in a less subtle

variation. Rather than comparing stimuli that vary along their color, we will simply compare two different critical stimuli: one will be clearly task-relevant (it will be a foil), the second one will not. However, the irrelevant stimuli will be significantly more luminant than the first. Here then, we are pitting the salience of a stimulus (in terms of its overall luminance) against the relevance of a stimulus (in terms of the task set). The question is: can we create inattentional blindness when the critical stimulus is a foil (i.e. task-relevant)? If not, then this is evidence that task-relevant stimuli (like the distractors used in the flankers paradigm) are not good stimuli to use when one wants to draw conclusions about distractibility. That is because such evidence would show that task-relevant stimuli are never *unattended* in the first place. To answer this question, we will use two possible critical stimuli (CS) a bright-white square and a white letter. The main task of the participants will be (as in a flankers task) to identify the letter in the center of the display. The task-relevant critical stimulus will be simply a traditional flanker (i.e. a foil, one of the possible target letters, presented slightly above, below, to the right or left of the target). We will compare performance in this condition to performance in a condition where, rather than a letter, we present a white square that is 1.5 times larger than the letter. Because the square is completely filled-in with white, and the letter is only drawn with a small portion of the pixels used to fill up the square, we can be certain that the square is substantially more luminant than the letter. Thus, if we are correct that stimuli which are highly relevant to the task are almost always selected by our attention system (as something that attention *should* select from the display), one would expect very small levels of inattentional blindness to the foil version of the CS. In contrast, given that the square is completely irrelevant to the letter-identification task, one would expect a much larger degree of inattentional blindness to the square, in spite of its much larger visual presence in the display.

Eighty four subjects were asked to complete a brief five-trial study, using a traditional unexpected event paradigm. Participants were told that their task was to identify a briefly presented letter and report its identity. The letter could be an A or an E, in lower or upper case, although the letter case was irrelevant to the response. To make matters challenging, the letter (about 1° of visual angle in extent) was preceded by a ~100 ms long random-dot mask (about 2° of visual angle in extent), which was immediately followed by the letter (presented only for ~70 ms), which was in turn immediately followed by a new random-dot pattern mask that stayed on the screen until participants had recorded a response. In the first three

trials of the experiment, those were the only visual events in the display, and participants completed the task with a good degree of accuracy. Average accuracy in the letter identification task in the first trial was 74%, and rose to 90% and 92% on the second and third trials, respectively. By the third trial, participants had a mean response time of 1470 ms (st. dev. 316 ms), which is not unusually long given that response time was not stressed in the task.

Then came the unexpected event trial: the critical stimulus was presented on the screen at the same time as the first random-dot mask and remained on the display until the target letter was replaced by the second random-dot mask. In all, the critical stimulus stayed on the screen for 160 ms, more than twice the duration of the target. The critical stimulus was presented 2° of visual angle away from the mask and could appear above, below, to the right or left of it.

Let us begin by establishing whether or not our main task and stimuli can, indeed, produce inattentional blindness by looking at performance in the group of subjects that was assigned to the square critical stimulus (which was approximately 0.75° of visual angle), a stimulus that was very much like the critical stimulus in many of Mack & Rock's original studies. How many participants reported seeing the unexpected square? Only 18 out of 42 participants (43%). When forced to guess the location of the unexpected stimulus, 17 out of 42 correctly reported its location. In other words, we produced a substantial degree of inattentional blindness to the square. Did the onset of the unexpected square interfere with the participants' effort in the primary task? Trial 4 accuracy was 97%, so, at first sight, one would argue that it did not. A post hoc RT analysis (limited by the variability of a one-observation/participant measure) suggested that it did not: subjects who did not report seeing the square responded to the letter target with an average RT of 1450 ms (st. dev. 600 ms), well in line with their performance on Trial 3. In other words, even though an unseen stimuli could theoretically have affected their RT (e.g. Moore et al., 2003), it failed to do so. In contrast, there was some suggestion in the data that those who did see the square actually slowed down in their response to the letter: the RT for the 18 participants that did see the square was 1661 ms (st. dev. 510 ms). The difference failed to reach significance, mostly because of the inherent variability in the measure. Finally, it should be noted that in the full attention trial, 40 out of 42 observers saw the square, and 38 out of 42 correctly reported its location. So, about 90% of subjects had no trouble seeing the square when not attentively engaged in the central letter task.

Overall, we can conclude that with our task parameters, and using a task-irrelevant stimulus, we can create substantial levels of inattentional blindness (almost 50%, the difference in detection between full attention and inattention trials). What happens when we replace the white square with a much smaller *foil* (i.e. a task-relevant letter)? In this task-relevant group, the critical stimulus was either the capital letter A or E. It should be noted that the two experimental groups were ran simultaneously and that our research assistants were blind to the experimental condition. This is important because it guarantees that all participants received the same instructions and that research assistants could not knowingly or unknowingly affect the outcome of the experiment. So, was this smaller "flanking" letter treated differently by the attention system of our subjects than the square? Absolutely: 34 out of 42 (81%) participants reported seeing the distractor letter, almost twice as many subjects who saw the square. Did the onset of this letter affect their performance in the central letter identification task *before they even knew we would ask questions about this unexpected letter?* Yes, again. Remember, when the critical stimulus was a square, participants in Trial 4 identified the target letter almost perfectly (97% correct). When the critical stimulus was a foil (one of the possible targets), target identification performance dropped to a dismal 69% (chance being 50%). RTs also substantially increased to 2022 ms and were highly variable (st. dev. 1408 ms). Importantly, it was only after participants had completed this response that we asked whether they had seen an unexpected event in the display. By the time we asked, all performance indicators suggested that yes, indeed, the foil had by and large not gone unnoticed. How visible was the foil? In the full attention trial, 40 out of 42 participants reported seeing the foil and 38 out of 42 (90%) correctly reported the foil's location. So, the a priori visibility of the square and letter foil was well matched across conditions, both were at 90%. Given that 81% of participants reported seeing the letter in the inattention trial, our results suggest that for the most part, there was little-to-no inattentional blindness in this condition. In sum, the attention system spontaneously detected and selected the foil (letter distractor), even though participants knew nothing about the possibility that a letter might appear in the periphery, and the act of selecting this letter substantially impacted their performance in the primary task.

Knowing that participants were not blind to the foil, one can do a second pass at the results and ask the question: was there a Flanker-like congruency effect in our experiment? Was the main letter identification task modulated by the congruence between target and foil? Yes again.

The average RT in Trial 4 for participants who had a congruent target–foil relation was 1743 ms (st. error = 175), whereas RT for participants who had an incongruent relation was 600 ms slower: 2359 ms (st. error = 425). When accuracy is analyzed, we found that 100% of participants in the congruent condition reported the correct letter. In contrast, only 31% of responses were correct in the incongruent condition. Finding an accuracy rate below chance suggests that participants were actually reporting the identity of the foil, not that of the target letter. If one recodes performance as a function of foil identity, participants' accuracy on Trial 4 becomes 90.5%, perfectly in line with their performance on Trial 3. This is interesting. The evidence from this experiment suggests not that the foil interfered with the target, but that, in fact, the target may have interfered with the foil (in terms of producing elevated response times). The results further suggest that the Flanker Effect may very well be a failure to select the appropriate of several potential targets (Lachter, Forster, & Ruthruff, 2004; Yigit-Elliott, Palmer, & Moore, 2011).

From this simple experiment we can conclude that, as we had hinted, using one of the possible target letters as a distractor in a selective attention task makes that distractor task-relevant, by our contemporary standards of what "task relevancy" means, but more importantly, by empirical standards as well. In fact, in Section 3 we will propose that when we ask the attention system to identify the presence of As or Es on a display, it will likely do so over the entire display, provided that there is sufficient signal to represent letters in the periphery (i.e. that there is no crowding and that the letters are sufficiently large to be accurately represented). So, the so-called "task-irrelevant" distractors in flankers task are actually always selected by our attention, precisely because they are actually very much task-relevant. Given that the presence and identity of such distractors alters our performance on the primary task (identifying the central target), does that mean that these distractors are *distracting* us in a Flanker Task?

2.6. Distraction or Distractor Interference?

2.6.1. The Current State of Confusion

In 2010, Lavie wrote a brief review on "Attention, Distraction and Cognitive Control under Load" (Lavie, 2010). The title is important because it identifies the concept of distraction as one of the central topics of the article. The conclusions of the article are as follows. If one wants to reduce the likelihood of distraction (at work, for example), it is advisable to work in a perceptually loaded environment. This means we should increase clutter in our desks. Further, we

might be tempted to infer that adding clutter (or otherwise increasing perceptual load) would be advised for situations where attentional focus is required, such as driving, but – as even Lavie (2010) admits – this advice would best be not taken. At gut level, this proposal sounds just wrong and ill-advised. Another important finding to come out of the review is that *increased cognitive load* creates more distractibility. We have not yet defined cognitive load here, but it can intuitively be thought as reflecting the cognitive difficulty of a task or the additional cognitive difficulty brought on by a secondary task. So, one should try to work under conditions of low cognitive load to avoid distractibility. While at first this sounds reasonable, let us translate the technical jargon in this conclusion: Lavie proposes that when we are mentally busy, we are at *our most distractible*, and the more we try to concentrate, the more distractible we will be. This is why she proposes that we should try to work under conditions of low mental load. But, what would be the point of concentrating on a task if not to shut out the world and protect our thoughts from worldly distractions? Lavie's "cognitive load" proposal again is not only counterintuitive, but it is a proposal that flies in the face of the entire inattentional blindness literature (in addition to being at odds with data from her own lab, Macdonald & Lavie, 2008, 2011): if nothing else, the theoretical and applied literature on this phenomenon has taught us that we are *insensitive* to otherwise easily visible events when we are highly focused on a cognitive task, it be judging the lengths of the arms of a cross (Mack & Rock, 1998), getting ready to produce a list of complicated navigation instructions (Simons & Levin, 1998), identifying a briefly presented letter (above), alphabetizing a list of letters (Fougnie & Marois, 2007) or talking on the cell phone (Strayer & Drews, 2007).

So, what gives? We propose that these conclusions (all based on soundly conducted experiments and robust effects) arise from a fundamental misunderstanding of the term "task-irrelevant distractor" and confusion between the concepts of *distraction* and *distractor interference*. Lavie (2010) calls the Flanker Effect: "the most conventional laboratory index of distraction" (p. 144). Lavie's Load Theory was entirely inspired by results from Flankers experiments and most of her work on attention uses variations on "task-irrelevant distractor" paradigms in which the distractors are not really task irrelevant. We have already suggested these are not appropriate paradigms for studying distraction. Distraction is the concept about how events or thoughts that we have *no* a priori *intention* of attending to or thinking about, come nonetheless to be attended or thought of. To reiterate, foils in a flanker task are task-relevant by design, so they fail the basic litmus test to be informative regarding the phenomenon of distraction. Our attention is actively set to select and scrutinize foils.

Thus, (1) we should not be surprised if they end up influencing behavior, and (2) they do not really distract us, as much as they *interfere* with our stated goal of reporting the identity of only one letter in the display (the target). We propose that there is a fundamental difference between these two ways in which our thoughts can be affected by an event (or thought): distraction, which we already defined, and distractor interference. Unlike distraction, *distractor interference* refers to the impact on performance of stimuli that are a priori *relevant* to our behavior but not the target of our task.

Let's return to the example of a visual search task: if the goal is to find a red circle and we bias our visual attention to select red items, the degree to which the number of distractors will *interfere* with our task of finding the red target will be determined by the level of functional overlap (or task relevance) between the distractors in the display and the target. If all distractor shapes are blue, there will be no distractor interference in this task and participants will find the target irrespective of set size. If, on the other hand, there are also red square distractors, these red distractors (because they are candidates) will interfere with our goal of finding the red circle, and RTs will increase with set size. Unless of course, we can bias our attention system to find the target in terms of its shape (rather than its color), in which case, there will be no functional overlap between our a priori *intentions* (of finding circles) and the distractors (squares). In Section 4 of this chapter we will review how we believe selection works in these simple examples. For the time being, it is important to recognize that the set-size effect in visual search is an example of distractor interference, not one of distraction. We are not *distracted* by the distractors, we are *interfered with* by the distractors, and in particular, by those distractors that are relevant to our task of finding the target. In our new terms, we inspect any and all and only candidates.

The same is true with the Flanker Effect and this follows because the flanker task is functionally equivalent to a visual search task without spatial uncertainty (Eriksen & Eriksen, 1974). In other words, it is incorrect to say that we are distracted by foils in a flanker task, we are simply interfered with by them. From this perspective, it is also easy to understand why the actual number of objects in a real-world scene will always be larger than the *functional* set size for any given visual search in that scene (the number of candidates). The functional set size will be determined by the number of objects that are a priori of interest to us in the scene. Looking for a person in a room filled with various furniture items will be accomplished without any effect of the number of pieces of furniture because the furniture cannot exert any distractor interference with our search task since they bear no functional overlap with our task settings

(for finding a person). But, if we are looking for a specific furniture item, then furniture items will be a priori items of interest to attention and will produce distractor interference effects (a significant set-size effect), whereas the number of persons in the room will be irrelevant to performance in that case.

Returning to Lavie's arguments, as explained earlier, we believe that extrapolating from Flanker tasks to the domain of distraction is inappropriate. The problem is that, while the flankers might not be providing information in the sense of information theory, they are highly task-relevant because they are foils—i.e. they are very much like the targets. This greatly contrasts with the truly irrelevant stimuli that are used in most experiments on inattentional blindness, such as the presentation of an unexpected square. Even more, the data from our small experiment confirmed that attention does treat those two situations extremely differently: when we are looking for As and Es, the presentation of an A or E near the target location will not go unnoticed, whereas the presentation of a bigger, brighter white square can be completely ignored. Thus, the foil in a flanker task *will be* processed, which in a way is the point of the Flanker Effect. But, as our results suggest, one cannot generalize between situations with flanking foils (which are actually task-relevant) to those obtained with flanking squares (which are substantially less task-relevant). The two types of distractors are treated in a profoundly different manner by the attention system.

In sum, one cannot, as Lavie does, extend the findings from traditional congruent and incongruent flanking stimuli (the sort that produce Flanker Effects) to task-irrelevant stimuli (the sort that cannot produce distractor interference), like those in most real-life situations pertaining to actual distraction. In fact, we want to argue that finding a Flanker Effect in an experiment is, to some important degree, a measure of the *success* of our attention system (not of its failures) because it demonstrates the degree to which our attention and information-processing system is able to simultaneously perform the task (with accuracies close to 100%) while also picking up several other candidate stimuli from the display. So, whereas Lavie (and many others before her) has characterized the presence of a Flanker Effect as a *limitation* of the selective attention system, we propose that it might index its efficiency. The Flanker Effect is not to be interpreted as evidence that attention works as a late selection filter. Just the opposite: it is clear evidence of attention working as an early filter, a filter whose setting is determined by our goals.

2.6.2. Distractor Interference as a Measure of Attentional Success

Case in point: a beautiful study by Torralbo and Beck (2008) demonstrated that larger Flanker Effects are found when there is less cortical competition

for representation among items in the display. In other words, when the brain can (without the need for biased competition, i.e. constrained spatial attention) represent all letters in the display, there will be larger Flanker Effects, precisely because the brain can successfully represent and select the relevant items in the display, and these representations have, in turn, an opportunity to independently activate response selection processes. The fact that this opportunity exists is what gives rise to distractor interference. Well-represented/identified distractors can activate stimuli–response associations and thereby interfere with our goal of producing a response to the target only. In contrast, when display items are poorly represented, our attention system ends up having to bias the competition to each stimulus separately (via spatial attention, which is effortful and works in serial manner) and thus there are fewer opportunities for distractor interference to arise. In other words, we believe that the *presence* of a Flanker Effect reflects to some degree a level of attentional success: it reflects the success of selecting multiple stimuli from a display. Which stimuli? Those that are relevant to our task, those that we a priori set our system to select from the display. Certainly, the *magnitude* of the Flanker Effect is likely determined by a number of factors (e.g. how small can we make our "spotlight" of spatial attention; the spatial arrangement of targets and distractors; is the target at fixation and distractor in periphery or vice versa; as well as a host of grouping and segregation cues that may further help attention to discard distractors, like color, etc.). The point here, though, is that the fact that there is an *opportunity* for distractor interference should be interpreted as evidence that attention is efficiently selecting from the display the things it cares about.

2.6.3. Connecting Flanker Experiments to Real-World Situations

This is why the recommendations from Lavie are, ultimately, exactly the opposite of what they should be. When she proposes that we should all work with highly cluttered displays (i.e. displays with high perceptual load) so that we can *avoid* Flanker Effects (i.e. in her worldview, *avoid distraction*), her recommendation is exactly opposite to what her data suggest we should do. We should actively seek display arrangements that are conducive to Flanker Effects precisely because the conditions that produce Flanker Effects are those in which we can effortlessly select several *task-relevant* objects from the display in parallel, whenever we need to find them. We will not select task-irrelevant objects, just like we do not select blue circles when looking for a red one, or just like we do not tend to select a square when looking for a letter.

With regards to her second recommendation, Lavie proposes that we should work under conditions of low cognitive load to prevent distractibility by "task-irrelevant" stimuli that is observed under high cognitive load. Again, this is exactly opposite to what one should conclude. She is again confusing distractor interference with distractibility. Distractibility is what has been measured in the inattentional blindness literature: when cognitive engagement is high, we are less sensitive to truly task-irrelevant distractors. And, if cognitive engagement is on the primary task (i.e. on identifying a letter), this will make it more likely that our system will pick up information from the world that is relevant to our task, which is exactly what one should hope it would do. This is also why there was almost no inattentional blindness to letters in our experiment. And what the literature on inattentional blindness has taught us is that a high level of cognitive engagement in the main task comes at the "cost" of less processing of truly task-irrelevant events in the world, so much so, that we can even become completely unaware of otherwise easily detectable events. And this is why our participants failed to see the square. In sum, when we concentrate on a task, (1) we are more likely to select the information that we are looking for in order to complete our task; and (2) we are less likely to select completely irrelevant information—i.e. less likely to be distracted. Countless real-life examples come to mind. When we are engrossed in a TV show, we may fail to hear the teapot whistling (task-irrelevant stimulus). When we are focused on writing a manuscript, we may fail to hear the clock in our office ticking. And when we are focused on a computer project and we want to look for folders on our desktop, we will not be distracted by the background image on our monitors, yet we will likely quickly know *where* all the folders on our desktop are.

Before we move on to the next brief section of this chapter, we want to point out that Lavie is not the only investigator to have misused the term "task-irrelevant" to qualify distractors or to have confused the concepts of distractor interference with that of distraction. Another visible example is Kim, Kim, and Chun (2005), *Proceedings of the National Academy of Sciences* who write "Concurrent working memory load can reduce distraction". Instead of using a Flanker Task to draw this conclusion, they used several variations of a Stroop task. In other words, they were measuring effects of task-relevant stimuli (foils) on target processing. This is not a measure of distraction; it is a measure of distractor interference. To be clear, we are not questioning the replicability of the results in any of the several dozen papers on perceptual load, or any paper on the Stroop effect, or Simon Effect.

These are all tasks that ask the same question: to what degree aspects of a task-relevant stimulus impact target processing. They measure distractor interference. Yet, in our haste to extend our results to the real world, to claim that our experiments can be applied to important real-life challenges, like the understanding of distraction and distractibility, we might be tempted to substitute one term for the other. This is wrong. This is not to say that the results from these experiments do not inform important real-world decisions like, for example, how to design a high-efficiency workplace. They clearly do, as we reviewed in the paragraph above. That said, we must be more careful about what the conclusions that we can draw about our experiments are and avoid overreaching in our explanatory power. To avoid making such mistakes, we believe it is important that we start being more mindful in the use of terms in our discipline, like distraction, distractibility and task (ir)relevance. Or we risk drawing the exact opposite conclusions from our data, as Lavie (2010) has done, or perhaps more subtle, use the results of our experiments to make inferences about phenomena that they cannot inform (as Kim et al., 2005). To this end, we have proposed the use of new terms, candidates and foils, to identifying different types of distractors. These labels are useful because they also help us to determine the explanatory bounds of the effects observed with each type of distractor.

3. A BRIEF CASE STUDY ON DISTRACTION

3.1. Limitations of Unexpected Event Paradigms

We have reviewed one set of experimental paradigms that has been useful in bringing about an initial understanding of distraction: the unexpected event paradigm that produces inattentional blindness. Unfortunately, though powerful, there are two main problems with this particular paradigm if one wants to further explore the issue of distraction. The first concern is simply a question of methodology: the typical unexpected event paradigm produces only one observation per subject. Though one can draw some important theoretical inferences from these experiments (as we have done above), it remains a fundamentally limiting paradigm. The main reason is that, once participants are made aware of the fact that the experimenters will be asking questions about the critical stimulus, they start attending it on every trial. Because of this, the critical stimulus becomes extremely relevant to the subject, so we can no longer argue that future instances of similar stimuli will be treated in the same "unattended" fashion as the first instance was. Good evidence for the "incorporation" of the critical stimulus into the main

attentive task can be seen in Moore, Lleras, Grosjean, and Marrara (2004). In that experiment, participants were asked to report the identity of a centrally presented letter (forward and backward masked). In addition to the target letter, a small dot could appear either to the right or to the left of the target. Participants completed a first block of 64 trials with this setup, followed by a short block of trials where we tested participants' awareness for the small lateralized square, as in an inattentional blindness experiment. Only 16% of participants reported having seen the square in the inattentional blindness trial. Critically, participants were asked to complete a second full set of 64 trials, identical in every way to the first block of trials. We used the accessory Simon effect (e.g. Simon, Acosta, Mewaldt, & Speidel, 1976) as a measure of distractor processing (via distractor interference). Whereas the square did not produce any interference on the letter identification task *before* the inattentional blindness trial, it produced robust levels of distractor interference in the second block of trials. In other words, participants in the first block of trials were unaware of the square (as assessed by the inattentional blindness results) and therefore it failed to impact performance. In contrast, once we asked questions about the square, they not only became aware of it, but incorporated it into their task. From then on, the presence of this square interfered with the responses to the central target on every trial.

A second problem with the unexpected event paradigm is one of generalizability. It is true that the paradigm does reflect a setting that one may find often in the real world (when one is effortfully attending to one event and a completely unexpected event takes place). Yet, there is the perhaps far more common situation of being in a busy environment (and *knowing* some of the characteristics of the environment) while trying to effortfully engage in some mental activity. For instance, one may be trying to read an important email from work while children are playing around. Or one may be talking on the cell phone while driving. In both cases, one is well aware of the characteristics of the environment. And one may decide to only pay attention to one thing (i.e. ignore the children in the first example) or pay mostly attention to one thing (i.e. momentarily decreasing attention to the driving). How will we react to events along these unattended (or significantly underattended) channels of information? This is an important question that has been asked as far back as Treisman (1960). And one that deserves being asked. With respect to the domain of driving, Strayer et al. have shown, for example, that drivers talking on the phone are 10 times more likely to fail to come to a stop at a four-way intersection, and they are significantly less likely to remember traffic-related events that occurred during their driving

(in a simulator) than when the events appeared and they were not talking on the phone (Strayer & Drews, 2007). What Strayer et al. argue is that this is because of inattentional blindness: drivers actually are less aware of the world while on the phone, and this decreased level of awareness does not appear to be mitigated by the traffic relevance of worldly events. That is, once drivers start underattending the driving environment, they seem to lose sensitivity to the driving relevance of events in the world. But one must highlight a certain incompatibility between Strayer's findings and those in the inattentional blindness' laboratory: whereas Strayer and Drews' findings suggest that the relevance of events is not associated with drivers' awareness of those events, the inattentional blindness literature shows that relevance does impact awareness. Clearly, more research needs to be done to understand underattended perception in a paradigm where we can systematically manipulate such things as the degree of cognitive difficulty in a task (cognitive load), the cognitive engagement in a task (motivation), as well as the relevance of events in the unattended/underattended channels of information. Below, we propose one such paradigm.

3.2. A New Paradigm

We here propose a simple paradigm to study actual distraction—i.e. the ability of information in unattended or underattended channels to draw attention onto itself while the observer is actively engaged on a primary task (Lleras & Buetti, 2012a,b). In this task, we ask participants to perform a sustained cognitive task: each trial lasted about 1 min. In the experiments below, we asked participants to perform a 1-back, 2-back, or running arithmetic task. The task-relevant stimuli for this task were presented auditory once every 3 s and each stimulus itself lasted somewhere between 500 and 800 ms. In other words, the participant had between 2200 and 2500 ms of time between task-relevant stimuli. They used this time to perform the required cognitive operation in their minds. The critical manipulation is what happened on the display while participants were focused on performing this ongoing task. We informed participants that while they were performing this cognitive task, completely irrelevant photographs would appear on the computer monitor. They were told that they could ignore those pictures and that they would not be tested on them at any point. There were 360 photos, all taken from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2008). They were selected as being somewhat neutral on valence and not arousing; they were mildly interesting photos. The images appeared 1800–2300 ms after the offset of the previous

task-relevant stimulus and 700–1200 ms before the onset of the next task-relevant stimulus. Images could appear at one of the four possible locations (above, below, to the right or left of the center of the display). Only one image was present at a time. When a new image appeared, it always appeared at one of the three unoccupied locations. An image disappeared exactly at the time the next image in the sequence appeared.

To summarize the experimental task, participants engaged in a complex cognitive task for a period of 1 min at a time. While they did this, images appeared every now and then on the display. Importantly, when a new image appeared, it was the only event in the world, the only "interesting" location to look at on the monitor and its presentation was isolated in time from the presentation of the task-relevant stimuli (that is, it did not compete with other stimuli for sensory attention). A schematic of the trial events is shown in Figure 7.1(a). Figure 7.1(b) and (c) show results from this task in three versions of the task. The experimental stimuli and events were identical in all three experiments. The only thing that differed was the task we asked participants to perform on the task-relevant stimuli. The task-relevant stimuli were a set of audio files with someone's voice reciting a simple mental operation ("plus one", "minus one", "plus two",… up to "minus five"). First group of participants performed a 1-back task on the list of operations, that is, they counted the number of times the same math instruction was presented twice in a row. This was a "low-cognitive load" group. A second group of participants performed a 2-back task on the operations, that is, they counted the number of times in the 1 min trial, the current operation was the same as the operation before the last one. This was a "high-cognitive load" group. Finally, a third group of participants was asked to actually compute all the math operations starting from a three-digit number that was presented to them at the start of the 1 min trial (a running arithmetic task). This was also a "high-cognitive load" group.

As can be seen from Figure 7.1, the data we collect is simply a measure of where people look when a new and very salient photograph suddenly appears in the display. The results speak for themselves. As one would expect (and as would be predicted by all theories of visual attention), when participants are only doing a very easy cognitive task (one that only requires rehearsal of auditory information), participants spontaneously look at our photos: 75% of the time, within 200 ms of a new image appearing, their eyes land on the photo. No surprise there. In fact, if there is a surprise is that the percentage is this *low*: one might have expected something closer to 100%. Be that as it may, it provides us with a baseline level of performance:

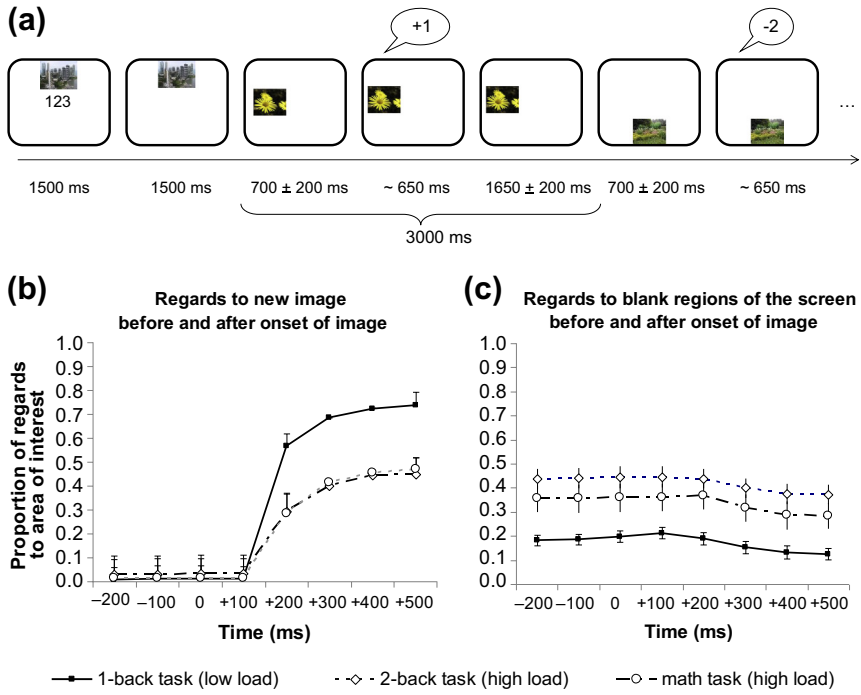


Figure 7.1 Oculomotor capture during mental engagement. (a) Schematic of events in a trial. Every 3 s, participants heard a new operation. The sound file lasted on average 650 ms. A new image always appeared at an unoccupied location, on average 1650 ms after the end of the sound file. (b) Proportion of regards to the new image within 500 ms after image onset (time = 0). Participants reflexively oriented to about 75% of images in the 1-back task, whereas they oriented to fewer than 50% of new image onsets both in the 2-back and math tasks. (c) Proportion of regards directed to empty regions of the display as a function of time across the three cognitive load conditions. All three experimental groups showed a gross insensitivity to the image onset (time 0). If a participant had been looking at an empty region of the display before the image onset, he/she continued to do so after the onset. Furthermore, the degree of regards to "nothing" was about twice as likely in the high-cognitive load conditions as it was in the low-load condition. (For color version of this figure, the reader is referred to the online version of this book.)

these irrelevant (though admittedly expected) images are inspected on a vast majority of instances. However, this is not what happens when we increase the degree of cognitive effort on the main task. The data also show that when people have to actively engage executive resources and apply them to the auditory information, their spontaneous regards of the new images drop to <50%. This is remarkable. Not only did their regards to new images significantly decrease (Figure 7.1(b)), but there is a large percentage

of regards to "nothing" (completely blank regions of the monitor, Figure 7.1(c)), and these regards are completely unaffected by the appearance of the new image: when participants were looking at nothing, they continued to do so after the onset of the image. The conclusion is simple. In identical environments, when participants engage in cognitive tasks that require some degree of executive control, they spontaneously withdraw their attentiveness from irrelevant events.

3.3. Discussion

This new task to study distraction is interesting because it reproduces a real-life situation in which one is trying to complete a mental task, and is focused on doing so, while there are ongoing events in the world. Interestingly, one has a general idea about the structure of the environment: here, that images appear with a certain frequency and that one can ignore them, inasmuch as one wants to ignore them, since they do not inform our decision making on the primary task. The results suggest that we are sensitive to some degree to the cognitive demands of a task and that we tradeoff attention, giving more to our thoughts and less to the world, when cognitive demands are high. Interestingly, this does not seem to be the end of the story. It is difficult to compare performance and mental effort across the two high-cognitive load tasks (perhaps separate measures of effort and mental workload like hear-rate variability and skin conductance would be informative). So, to further investigate the effect of cognitive demands on attention, we performed a follow-up analysis on the math task, because we had a level of control over the actual cognitive difficulty of specific math operations. We measured on a separate, self-timed task, how long it took participants to complete each of the math operations in the experiment. For instance, it took our participants 1.7 s to complete a +1 operation on a three-digit number, whereas it took twice as long to complete a −5 operation (3.2 s). These data suggest a higher degree of cognitive effort is required to complete +5 operations compared to +1 operations. When we analyzed regards as a function of the difficulty of the math operation, we found no measurable effect of the difficulty of the operation on the likelihood that participants will look at the next new image when it appears (Figure 7.2). Thus, even though for our participants it is substantially easier to add 1 to a three-digit number than to add 5, the complexity of this operation does not determine whether they will look or not at the next new image on the screen. This suggests that the "tradeoff" between attentiveness to the outside world (images) and attentiveness to the mental world (cognitive operations)

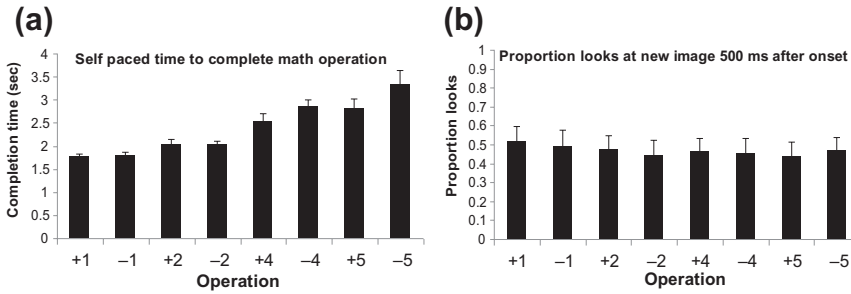


Figure 7.2 Analysis of the math task. (a) Participants completed at the end of the experiment a self-paced version of the main experiment, where they had a chance to advance the operations at their own pace. No images were presented on the screen. Operation times varied greatly as a function of the operation. (b) Proportion regards on the new image, 500 ms after its onset. Comparing panels (a) and (b) demonstrates that in spite of large differences in the degree of difficulty of specific math operations, the type of math operation participants were performing when the image appeared had no impact on their likelihood to look at the new image.

is not driven by a moment-by-moment need for cognitive resources, but by an a priori and general investment of resources. That is, it appears that the actual cognitive resources required to complete a task do not determine the level of distractibility of an individual, rather, it seems that the degree of *anticipated effort* is what determines it. In other words, we believe it is the amount of resources that participants voluntarily set aside to commit to their primary task that determines the attentional tradeoff between thoughts and outside world. The present cognitive difficulty of a task may only play an indirect role in this tradeoff; initially, it may help to calibrate the degree of concentration (effort) to invest in the primary task ("this is going to be tough, I better concentrate"), and it may interact with the degree of motivation of participants ("this is too hard").

Evidence in favor of "effort calibration" in a cognitive task has been documented in the abbreviated vigilance task literature (e.g. Helton & Russell, 2012; Temple et al., 2000) in which participants are asked to engage in brief, but cognitively demanding tasks for periods of about 2 min (each with about 110 events of interest) (Figure 7.3). Performance in the first 2 min trial or "period of watch" (as it is called) is typically superior than performance at all subsequent 2 min trials, as if participants had initially engaged in the task with great motivation, but after the first 2 min trial had decided to settle at a more comfortable, less arduous level of engagement. Similarly, Silvia, Jones, Kelly & Zibaie (2011) produced good evidence that motivation and cognitive difficulty interact to determine performance (see also Gendolla &

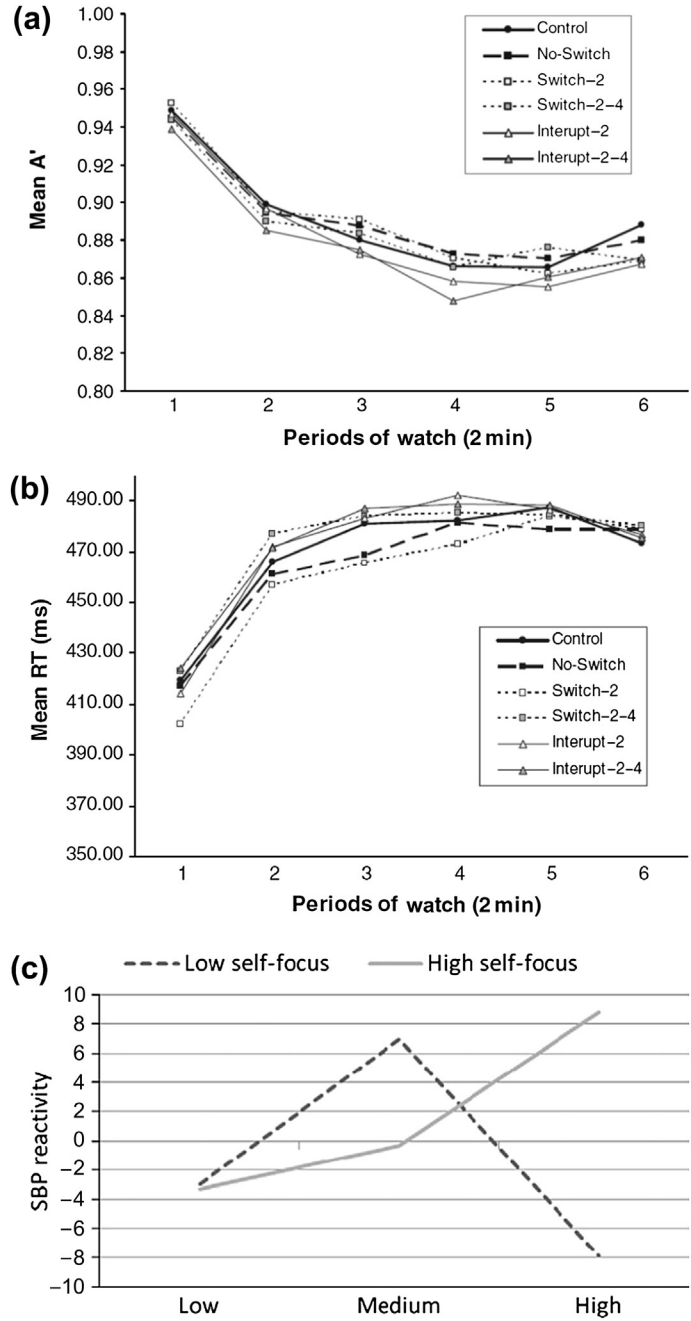


Figure 7.3 (a) Figure 3 from Helton & Russell (2012) showing performance (as indexed by A') in a demanding sustained attention task, as a function of periods of watch.

Richter, 2010; Silvia, McCord, & Gendolla, 2010). In their paper, Silvia et al. assessed motivation in terms of a personality trait ("self-focused attention"), which has been shown to reflects participants' willingness to exert effort in a task, as well as their own level of trait motivation (e.g. Carver & Scheier, 2001; Duval & Silvia, 2001). The authors manipulated the difficulty of a task by changing the response window (long, medium and short = easy, medium and difficult, respectively), using a perceptual discrimination that was difficult to begin with. They measured systolic blood pressure (SBP) reactivity, which is a good physiological marker of effort (e.g. Bongard, 1995; Smith, Ruiz, & Uchino, 2000). Participants high in self-focused attention showed increased SBP reactivity as the difficulty of the task increased. In contrast, for participants low in self-focused attention, SBP reactivity increased from low to medium task difficulty levels, and then it dropped when task difficulty became high. This is evidence that some people can motivate themselves to exert larger effort in difficult tasks than others. Here we propose that this decision of how much effort to put into a task is as critical a determinant of their distractibility as is the cognitive requirements of the attentive task. Note that we believe motivation can also be directly manipulated by incentives or pressure, and we believe those manipulations would result in the same effects on attention. There is a need for research along these lines.

Periods of watch are 2 min intervals during which performance is evaluated. The entire experiment lasts 12 min. (b) Figure 4 from Helton & Russell (2012) showing mean reaction time in the same task. As can be seen from these two figures, participants initially engage strongly in the task (first 2 min), both in terms of their initial effort to be accurate (higher A′) and fast (slower RT). However, performance drops quickly after that initial period and there is little-to-no change in the remaining five periods of watch, either in terms of accuracy or RT. Notice that this is the opposite of a speed-accuracy tradeoff: participants become both less accurate and slower, even though nothing has changed about the task. This is evidence of voluntary changes in effort (or engagement) in the task, as participants calibrate their effort, and settle into the task. (c) Figure 1 from Silvia et al. (2011). Systolic blood pressure reactivity (an index of incurred effort) as a function of task difficulty (low, medium and high) for participants either low or high in self-focus attention, which is a measure of how people evaluate their performance with respect to their own standards and how they motivate themselves to achieve (or not) those standards. People high in self-focused attention will try harder than people low in self-focused attention as success is more significant to them and they will motivate themselves endogenously to achieve better performance. As can be seen, people high in self-focus increase the degree of effort they put in the task as task difficulty increases from low to medium and from medium to high. In contrast, participants low in self-focused attention try harder when difficulty goes from low to medium, but when difficulty further increases to high levels, they stop trying, and SBP reactivity simply drops. *(All figures reprinted with permission)*.

4. A THEORY OF ATTENTION AND DISTRACTIBILITY

4.1. The Need for Inner Focus

Glenberg, Schroeder, and Robertson (1998) presented a wonderful demonstration of the spontaneous tradeoff between attention to the world and attention to our thoughts. In their experiments, participants were asked to do a series of remembering tasks that varied in difficulty. The specific questions were about autobiographical events or general knowledge. In Experiment 3, participants viewed such questions on a computer monitor and the investigators rated the number of times people spontaneously averted their gaze from the monitor (e.g. closed their eyes or looked away). Their results showed that the more people averted their gaze, the better their performance in the memory task. In Experiment 4, they converted this measure to a manipulation: in some conditions, they asked participants to close their eyes while thinking of the answer to the question. They found that people performed better at both the memory task and on a set of simple math problems, as well, when they closed their eyes than when they kept their eyes open. Together, these results are good indicators of two main characteristics of our attention system. First, when we have a need to concentrate on some internal cognitive process, we tend to actively prevent the encoding of potentially distracting information. Second, the prevention of encoding of potentially distracting information is, in fact, functional: we are better thinkers when we shut our eyes. This explains why we tend to avert our gaze from a listener when we are about to start speaking (e.g. Beattie, 1978a,b; Doherty-Sneddon, Bruce, Bonner, Longbotham, & Doyle, 2002; Doherty-Sneddon & Phelps, 2005; Ehrlichman, 1981; Goldman-Eisler, 1967): we are actively trying to shut down a large source of potential distraction (our listener's face) while we are engaged in trying to plan new sentences.

Thus, our instinct to move our eyes away from potential sources of distractions is both a reflection of our intuitive understanding of the hazards of the environment (where distracting information may be found) as well as an effective strategy for controlling the allocation of cognitive resources in times of need. Interestingly, Glenberg et al.'s results also showed that when overly difficult questions were used (such as questions to which people knew they did not know the answer for example), people no longer averted their gaze. They had stopped trying to do the task. They knew that a momentary boost of cognitive focus would not change the fact that they did not know the answer to the question posed. So, they no longer engaged in gaze

aversion. This makes it clear how gaze aversion is not merely a reaction to heightened cognitive demands; it is strategic in nature. We propose that the attentional system has an executive component that is sensitive to *anticipated* cognitive demands (and whether said demands will exceed capacity), as well as *motivational* states to bring about a tradeoff between sensory attention (i.e. attention to world events) and mental attention (i.e. attention to inner cognitive processing). And we propose that it is this tradeoff, when pushed to extremes, that leads to the more blatant failures of awareness observed in inattentional blindness and when driving while talking on the phone. In a momentary need for extreme inner focus, attention to the world suffers.

The need to reorient attention inward may also help explain other psychological phenomena, such as the attentional blink (Raymond, Shapiro, & Arnell, 1992). This phenomenon refers to the finding that people often miss the second of two successive targets, at least when those targets are presented embedded in a rapid visual stream of distractors and the second target appears 200–500 ms after the first one. There are several theories as to why the attentional blink occurs, but most models agree that the blink itself seems to be tightly linked to the need to "save" information from the first target in working memory, so that it can be accessible for response later on. Functionally, this seems like a brief but acute need for inner focus: upon detection of the first target, and given the presence of so many distractors (the other letters in the stream), the participant may need to effortfully pick and store the detected target in working memory. If it is true that an acute need for inner processing triggers a disengagement from potential sources of distraction, then the need to store the first target, once it is detected ("Oh, that's a target! I need to remember it."), in the face of rapid steam of stimuli, may be what causes the blink to occur. This account is not novel per se. What is new is the explicit presentation of a framework from which to understand why consolidation into working memory (or any other demanding and inwardly directed process) would create a blink in the first place (as several theories have argued, Bowman & Wyble, 2007; Chun & Potter, 1995; Craston, Wyble, Chennu, & Bowman, 2009; Di Lollo, Kawahara, Ghorashi, & Enns, 2005; Raymond et al., 1992; Shapiro, Caldwell, & Sorensen, 1997).

4.2. Predicting Inattentional Blindness

We believe our proposed framework for attention allows us to reframe one of the principal challenges in the inattentional blindness literature: how to predict when a given participant will fail to report an unexpected stimulus.

Several studies have shown that more difficult (primary) tasks lead to more blindness (e.g. Cartwright-Finch & Lavie, 2007; Fougnie & Marois, 2007; Simons & Jensen, 2009). But performance in the primary task alone is not a good predictor of inattentional blindness (Cartwright-Finch & Lavie, 2007; but see Simons & Jensen, 2009). In fact, Simons & Jensen (2009) recently demonstrated that someone's skill at a task also fails to predict the degree of blindness. Why is that? The attentional tradeoff we propose is sensitive to the anticipated need for executive control, as well as the motivational state of the participant. The anticipated need for executive control will depend on a number of factors. The self-perceived skill of a participant at a task will play a big role in determining need: cigar-rolling experts will require less-executive oversight of the activity than a beginner, thus, one would predict that the expert will be less likely to suffer from inattentional blindness while engaged in this activity. Once an anticipated need is identified, one must ask the question: how well can the participant fulfill that request? Does he have the ability to precisely match the need with the corresponding resources? It is possible and more research is needed to answer this question, but it is also likely that participants spontaneously assign resources in gross fashion: "I will focus on this task" vs "I will focus on this task a lot," rather than "I will focus on this task with 47% of my resources" vs "I will focus on this task with 66% of my resources." This is not to say that participants are incapable of allocating their attention in fine quanta, but that it is more likely that they spontaneously allocate in relatively large chunks. This may also be a reason why experimental manipulations of the degree of difficulty of the attentive task (aimed at capturing incremental levels of attentiveness in the participants) may not lead to a nicely titrated relationship with performance. It may also explain why the actual difficulty of the mental arithmetic task failed to influence gaze aversion in the task presented above: participants are either attentive to the math task or they are not, as they anticipate a need for focusing on it to succeed. But once that determination has been made, whether a specific operation is easy (+1) or hard (−5) is unlikely to change their overall attentive engagement to the task (Figure 7.2).

Finally, the amount of attentiveness to the main task (and ergo the degree of unattentiveness to the rest of the world) will also be determined by motivational and personality characteristics of each individual. As shown by Silvia et al. (2011), once a task becomes too difficult, some people will continue to increase their effort to match the demands of the task, whereas others will simply stop trying. Factors like the self-focused attention trait (as suggested by Silvia et al.), as well as conscientiousness (e.g. Trautwein,

Lüdtke, Roberts, Schnyder, & Niggli, 2009), will probably be important factors to determine engagement in the attentive task. If one thinks back to experts performing a task that does not require (of them) lots of executive control, one might still see differences in inattentional blindness as determined by their conscientiousness: in spite of their expertise, conscientious experts may deliberately pay more attention to the task, even if they do not need to. And of course, this is likely to change as a function of both intrinsic and extrinsic motivation. If there is a high reward to perform well in the attentive task, one would expect that individuals more responsive to rewards will be more attentive and thereby exhibit larger degrees of inattentional blindness.

In sum, if one wants to predict performance in any situation that would seem to be parallel to inattentional blindness, there are a number of factors that will need to be considered and measured. Failure to take these factors into account will likely produce just noise. Take the Simons and Jensen result, for example, that skill at a motion tracking task failed to predict inattentional blindness. Skill was rated as the speed at which a participant could accurately track to a level of 75% accuracy. Importantly, all subjects in the inattentional blindness task were tested with stimuli moving at a speed that fell more or less at the half-way point of skills in the participant pool. At one end of the spectrum, one is measuring people low in skill at the task, performing at a level beyond their skill. These subjects will either try hard (if conscientious or high in self-focused attention) or simply give up ("it's too hard for me!"). Thus, at the bottom end of the skill distribution, one cannot predict inattentional blindness; it can go either way. The same is true at the top end of the distribution: all these participants can perform the task, but if the difficulty is sufficiently lower than their skill, then (1) they may not see a need for engaging high levels of attention in the task or (2) if they are conscientious, they will likely focus a lot on the task, to make sure that they excel in accordance to their own standards. So, neither at the low nor high end of the skill distribution will skill alone predict inattentional blindness. Admittedly, our proposal remains to be tested, but we hope that it at least provides some guiding parameters for what considerations should be taken when one is trying to predict IB in a large group of people.

4.3. A Look Back at Visual Attention

We will not elaborate much on how visual attention works, except to emphasize some general properties that we believe guide how visual selection takes place. Our starting points are theories like Bundesen's theory

of visual attention as well as Desimone and Duncan's biased competition theory. First, we propose that the visual system uses its "nodes" of specialization to select in parallel and across the visual field stimuli that are processed by those nodes. For instance, if the visual system has a level of specialization that focuses on processing motion, this node can be leveraged to select the presence of particular motions in the world (e.g. feature-based attention, Serences & Boynton, 2007). The same ought to be true for higher or lower nodes, so that one can select on the basis of orientations (V1), shapes (V4), colors (V1, V4a) but also alphanumeric characters (visual word form area, Nestor, Plaut, & Behrmann, 2012), and other objects of expertise (fusiform face area, like faces or cars, if one is a car expert, see McGugin, Gatenby, Gore, & Gauthier, 2012). In other words, one can easily set a filter to detect all faces in a scene. To the extent that those faces are sufficiently well represented at the neural level (i.e. this probably means there is no competition at the level of that nodes' receptive fields from nonpreferred stimuli), they will likely be selected in parallel. Thus, effortlessly, one can know where all the persons in a room are, and likely too, process in parallel properties of the group of selected faces (e.g. Chong & Treisman, 2003, 2005a,b; Haberman & Whitney, 2009).

Note that this "parallel" function of visual attention is both powerful and intuitive. We know that visual information about the entire scene flows in parallel throughout the visual system (i.e. is being analyzed by these different nodes), thus for example, the FFA is analyzing face-like inputs across the visual field all at once, or MT motion information across the visual field. What we are arguing is that attention leverages this massively parallel processing toward biasing a subset of elements in the scene. This allows for efficient selection (or adoption) of candidates across the visual field. Note, too, that there are natural limits to this parallel form of processing. For instance, in addition to competition for representation, cortical magnification will make it so that information farther from fixation will be represented with less fidelity than information at or near fixation, thereby imposing a constraint on the quality of input that is available to this parallel function of attention.

This initial form of parallel selection is a function of relevance (Bundesen, 1990; Fischer & Whitney, 2012). And it is evident in visual search, where observers are extremely efficient at massively rejecting in parallel large sets of items that are not relevant to the search task (Cunningham & Wolfe, 2012). Furthermore, to the extent that distractors are similar to each other and group well, observers can reject these in large groups, as well

(e.g. Anderson, Vogel, & Awh, in press; Duncan & Humphreys, 1989). So, we can both put in place large "adoption" as well as large "rejection" filters to process information in parallel. The output of the "adoption" filter is the set of likely targets for the current task (i.e. *candidates*). Again, assuming that the items in the display can be accurately represented (i.e. no crowding), how fine the "adoption" filter will be in part determined by the ability of a particular "node" to process specific sets of features that discriminate targets from other stimuli in the set. The more diagnostic target features there are (relative to the specific vision node responsible for computing said features), the smaller the set of candidates or likely targets will be (i.e. the more effective the first sweep of selection will be). The less diagnostic the target features are, the larger the set of items is that will be selected in that initial sweep of selection.

What happens next? It depends on the task. In visual search, the task is to select the "one" target among the "selected few" possible candidates at various locations. To perform this operation, a comparison must be made between the candidates and the target template, and this comparison will be made in parallel in working memory and will be determined by the capacity of working memory (Anderson et al., in press). The larger the capacity is, the more of these comparisons can be made simultaneously. The smaller it is, the fewer of these comparisons can be made. So, in fact, the "serial" aspect of visual search observed in most experiments is actually the result of two massively parallel operations (see also Young & Hulleman, 2012). An initial sweep that determines candidates (i.e. a priori targets), that is likely of unlimited capacity (not counting visual acuity limitations or representational crowding). This initial sweep is followed by a second parallel sweep that compares candidates to the target template, also in parallel. However, this second sweep is capacity-limited (Anderson et al., in press; Pashler, 1987). For the current purposes, let us say that the maximum number of items is four. Search then will proceed in serial manner, in groups of four, inspecting the set of candidates, until a match is found (or until acuity and/or crowding require refixations to new locations in the search space). Note that the comparison process will also be limited by the precision of the target representations in memory as well as the confusability between candidates and target templates (e.g. Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007). It is also likely that when the target is fixed, the comparison process will become more and more efficient over time (e.g. Carlisle, Arita, Pardo, & Woodman, 2011; Logan, 1988; Woodman & Arita, 2011).

From this overview, a second important function of attention can be determined: attention must act to decrease the noise in the representation of visual items (as proposed by biased competition theory), either covertly or overtly. That is, when items compete for representation, attention must act on each item individually to improve the representation of that item, at the momentary expense of surrounding items. Indeed, Scalf & Beck (2010) nicely demonstrated that attention to multiple items simultaneously does *not* result on improvements in representation of those items, if those are competing for representation because of their spatial arrangement. Eye movements, guided by attention, then work to solve this problem: by directly gazing at potential items of interest, the representation of those items is greatly improved. Alternatively, attention can be directed serially to each of the competing items to improve, one at a time, the representation of each of those items.

If the task is a Flanker task, the candidates selected by attention in parallel can actually be foils. That is, not only do candidates look like a target, congruent and incongruent flankers have links to potentiated responses (press right/left, for example). Thus, the Flanker Task is actually a task in which response selection processes must select which of the *already selected* foils they want to respond to. The problem comes from the fact that by virtue of having been selected in the first sweep, the foils are already activating potentiated responses, which is why the congruency between foils and target ends up producing a Flanker Effect. Studies in the Flanker Effect literature are, therefore, centered on the processes that allow for more *efficient suppression* of foil interference, or alternatively, better identification of the proper target among the foils (Lachter et al., 2004; Yigit-Elliott et al., 2011). And there are studies showing that the executive control system is actively trying to learn as much as it can from the task ecology to try to be as effective as possible in that regard (Max & Tsal, 2011, 2012). Yet, the critical message that should be remembered is that the presence itself of a Flanker Effect represents a specific sign of attentional success: the foil was successfully selected by our "adoption" filter as a candidate. The relative reduction (or elimination of a Flanker Effect when one was expected) reflects success of a different kind: success in the efficient suppression of foil interference. But note that the simple Flanker Effect can be modulated by both factors. In displays with "high perceptual load," a Flanker Effect may not be observed for two categorically different reasons: because the flanker was not selected as a candidate (perhaps because of crowding or competition for representation, which requires slow allocation of spatial attention to disambiguate

visual representations) or because foil interference was entirely suppressed (perhaps because the executive system found ways to more efficiently suppress response-related foil processing).

5. CONCLUSIONS

A Flanker Effect experiment is an experiment that studies the response interference that is created by selected foils on target processing—what people may call a form of "distractor interference". Another example of distractor interference is seen in the color-word Stroop task, where foils are words that semantically activate the same response tokens as the colors of the ink in which the word is written (which is what determines the response). So, there are various forms of distractor interference effects, but, in general, it is important to recognize that the effects arise because the distractors are in fact "foils" of some sort (foils because of their visual similarity to the target as in the traditional flanker or their semantic overlap with the target information as in the classic Stroop). The attention system efficiently selects foils from a display (as it does with candidates). As we have argued before then, paradigms like the Flanker and Stroop cannot measure distractibility and cannot be used to study distractibility because the stimulus causing the interference was selected in the first place. Inspecting candidates and foils is a critical part of the task. It is what is intended from the start. Given that foils are present in the display, inspecting them is in an important way part of what the participant must do to successfully complete the task.

We have reviewed a number of issues regarding various selective attention paradigms and more specifically, we have identified the problematic use of certain labels within this literature (Figure 7.4). We have proposed a new labeling system for qualifying "distractors" across various tasks, so as to minimize issues with the interpretation of results from these paradigms. Importantly, the labels themselves make it clear what the explanatory power and the explanatory domain are of a given stimulus and task set. One cannot and should not try to generalize results from one type of distractors to another set of distractors without proper evidence that such generalization is warranted. Conducting a Flanker Effect and concluding something about distraction is therefore problematic. Having defined this new set of labels is important because it allows investigators who are interested in using our paradigms to understand other aspects of cognition (such as cognitive problems in clinical populations) to know which tasks and stimuli are most appropriate for studying their phenomenon of

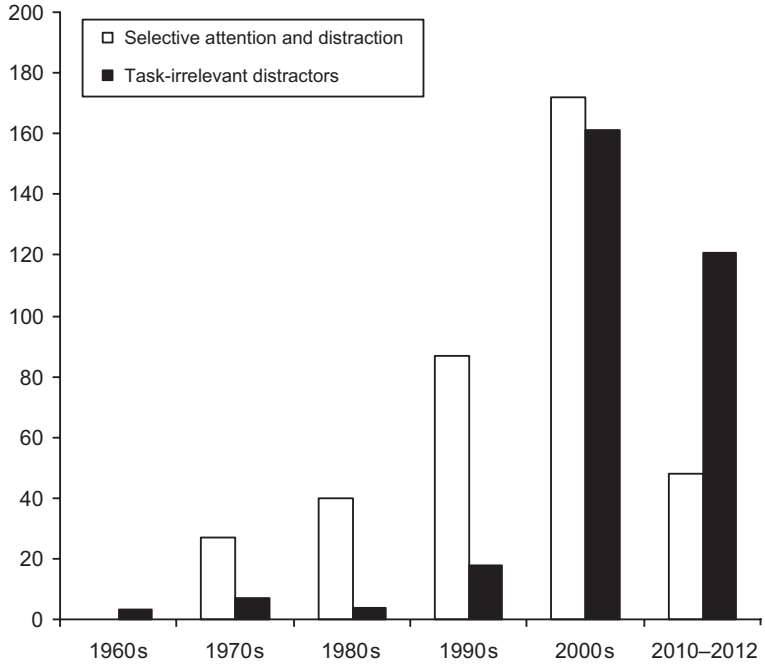


Figure 7.4 Citation counts each decade for two different searches in PsychInfo. In white are the citation counts for the keyword "distraction" in all fields within the specific subject "selective attention". In black are the citation counts for the search "task-irrelevant distractors" or "task-irrelevant distractor". This citation count simply illustrates the sudden explosion of articles using the terms "task-irrelevant distractors" in the last dozen years (2000–2012). The citation count for the conjunction selective attention and distraction is presented here as a baseline. This is instructive as it documents the research interest in understanding distraction, within the selective attention literature. That said, most selective attention paradigms are not appropriate for drawing conclusions about distraction, if they use candidates or foils as stimuli. With respect to the use of the term "task-irrelevant distractor," its rise in prominence in the literature is clearly worrisome as it is likely people are misusing and/or misunderstanding the meaning of the term: 75 of those papers (almost 40%) also include the word "distraction" as a descriptor of their paper.

interest. As an example, if one is interested in understanding distractibility in aging or in Alzheimer's disease, it would be simply inappropriate to use a Flanker or a Stroop task.

With respect to the critical difference between distractor interference and distractibility, one final comment is warranted. It is important to understand that distractor interference effects are all "within-task" effects. The subject never feels like he or she has "changed" what he/she is doing while

performing a Flanker or a Stroop task. The instructions for the task are set and the participant feels in compliance throughout the trial. In terms of executive control, the goal structure to complete the task does not change, even when one is suffering from severe levels of distractor interference, as in the traditional Stroop task. A person trying to name the ink of a color word throughout the trial is struggling with completing one specific goal: naming the ink (not saying the word) of the colored word. That much is clear. A corollary to this is that no matter the type of task-relevant distractor, a task-relevant distractor can never alter the boundaries of the task space. In contrast, the study of distraction is itself a study of changes to the task space. Distraction is the taking possession of the mind by a stimulus or thought that the subject *never* intended to process in the first place. It is a thought or event that interferes with the "goal stack," therefore changes the task space. Participants are suddenly doing (or thinking about) something other than their stated intention. If prompted, participants can answer the question "are you currently doing the task you are supposed to be doing". If distracted, the answer to that question will be no (Giambra, 1989; McVay & Kane, 2009, 2012; Smallwood et al., 2003), because they can recognize that their goal stack has changed.

What is particularly pernicious about distraction is that we are not always aware that a change in the goal stack has occurred. That is troublesome. That is why understanding distraction is a matter of crucial importance to us as we try to understand problems like "distracted driving" or rumination in anxiety and depression. If we are not doing what we are meant to be doing and are not aware of it, that is a problem. The example that jumps to mind is distracted reading (Smilek, Carriere, & Cheyne, 2010). Often we realize we have been "reading" a book but have no idea of what it is that we just read, because our goal of reading was pushed down our goal stack to a lower priority than, for example, our preoccupation with Mom's dinner plans for thanksgiving. In the visual world, understanding sensory distractibility is also of crucial importance to understanding issues like performance at work or human response to alert and alarm systems. A number of interesting questions remain completely open and thus far are largely unanswered: What events in our intended task are capable of bringing us back to our main task? The death of the main character in our book? In the case of distracted driving, will the change of a traffic light to red bring us back from our cell phone conversation and into the driving task again? This issue of "competition for goal status" is fascinating and will likely be the source of much fruitful research in the years to come.

REFERENCES

Alexander, R. G., & Zelinsky, G. J. (2011). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8).

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111.

Anersdon, D. E., Vogel, E. K., & Awh, E. A common discrete resource for visual working memory and visual search. *Psychological Science*, in press.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items, regardless of complexity. *Psychological Science*, *18*(7), 622–628.

Beattie, G. W. (1978a). Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, *17*(1), 1–9.

Beattie, G. W. (1978b). Sequential temporal patterns of speech and gaze in dialogue. *Semiotica*, *23*, 1–24.

Bichot, N., Rossi, A., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*(5721), 529–534.

Bongard, S. (1995). Mental effort during active and passive coping: a dual-task analysis. *Psychophysiology*, *32*(3), 242–248.

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, *114*(1), 38–70.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547.

Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience*, *31*(25), 9315–9322.

Cartwright-Finch, U., & Lavie, N. (2007). The role of perceptual load in inattentional blindness. *Cognition*, *102*(3), 1–20.

Carver, C. S., & Scheier, M. F. (2001). *On the self-regulation of behavior*. New York: Cambridge University Press.

Chong, S., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 1–12.

Chong, S., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1–13.

Chong, S., & Treisman, A. (2005b). Statistical processing: computing the average size in perceptual groups. *Vision Research*, *45*(7), 1–10.

Choo, H., & Franconeri, S. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception & Psychophysics*, *72*(1), 86–99.

Chun, M., & Potter, M. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 109–127.

Cooper, J. M., Vladisavljevic, I., Medeiros-Ward, N., Martin, P. T., & Strayer, D. L. (2009). An investigation of driver distraction near the tipping point of traffic flow stability. *Human Factors*, *51*(2), 261–268.

Craston, P., Wyble, B., Chennu, S., & Bowman, H. (2009). The attentional blink reveals serial working memory encoding: evidence from virtual and human event-related potentials. *Journal of Cognitive Neuroscience*, *21*(3), 550–566.

Cunningham, C. A., & Wolfe, J. M. (2012). Extending "Hybrid" visual X memory search to very large memory sets and to category search. *Paper presented at the 53th Annual Meeting of the Psychonomic Society*.

Darlington, T., & Talbot, E. B. (1898). A study of certain methods of distracting the attention. III. Distraction by musical sounds; the effect of pitch upon attention. *American Journal of Psychology*, *9*(3), 332–345.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 1–30.

Deutsch, J., & Deutsch, D. (1963). Some theoretical considerations. *Psychological Review*, *70*, 80–90.

Di Lollo, V., Kawahara, J., Zuvic, S., & Visser, T. A. W. (2001). The preattentive emperor has no clothes: a dynamic redressing. *Journal of Experimental Psychology: General*, *130*(3), 479–492.

Di Lollo, V., Kawahara, J., Ghorashi, S., & Enns, J. (2005). The attentional blink: resource depletion or temporary loss of control? *Psychological Research*, *69*(3), 191–200.

Doherty-Sneddon, G., & Phelps, F. (2005). Gaze aversion: a response to cognitive or social difficulty? *Memory & Cognition*, *33*(4), 727–733.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology*, *38*(3), 438–445.

Driver, J., & Tipper, S. P. (1989). On the nonselectivity of 'selective' seeing: contrasts between interference and priming in selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 304–314.

Dulsky, S. G. (1932). Discussion: what is a distractor? *Psychological Review*, *39*(6), 590–592.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Duval, T., & Silvia, P. (2001). *Self-awareness & causal attribution: A dual systems theory*. Boston: Kluwer Academic.

Ehrlichman, H. (1981). From gaze aversion to eye-movement suppression: an investigation of the cognitive interference explanation of gaze patterns during conversation. *British Journal of Social Psychology*, *20*(4), 1–9.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non-search task. *Perception & Psychophysics*, *16*, 143–149.

Estes, W. K., & Taylor, H. A. (1966). Visual detection in relation to display size and redundancy of critical elements. *Perception & Psychophysics*, *1*, 9–16.

Fecteau, J. (2007). Priming of pop-out depends upon the current goals of observers. *Journal of Vision*, *7*(6), 1.

Fischer, J. T., & Whitney, D. (2012). Distractor suppression: attention gates visual coding in the human pulvinar. *Nature Communications*.

Folk, C. L., Remington, R. W., & Wright, J. H. (1994). The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 317–329.

Folk, C. L., Leber, A. B., & Egeth, H. E. (2008). Top-down control settings and the attentional blink: evidence for nonspatial contingent capture. *Visual Cognition*, *16*(5), 616–642.

Fougnie, D., & Marois, R. (2007). Executive working memory load induces inattentional blindness. *Psychonomic Bulletin and Review*, *14*(1), 142–147.

Franconeri, S., Alvarez, G., & Enns, J. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1003–1012.

Gendolla, G. H., & Richter, M. (2010). Effort mobilization when the self is involved: some lessons from the cardiovascular system. *Review of General Psychology*, *14*(3), 212–226.

Giambra, L. M. (1989). Task-unrelated thought frequency as a function of age: a laboratory study. *Psychology and Aging*, *4*(2), 136–143.

Glenberg, A., Schroeder, J., & Robertson, D. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658.

Goldman-Eisler, F. (1967). Sequential temporal patterns and cognitive processes in speech. *Language and Speech*, *10*(2), 122–132.

Grice, G. R., & Gwynne, J. W. (1987). Dependence of target redundancy effects on noise conditions and number of targets. *Perception & Psychophysics*, *42*, 29–36.

Grice, G. R., Canham, L., & Boroughs, J. M. (1984). Combination rule for redundant information in reaction time tasks with divided attention. *Perception & Psychophysics*, *35*, 451–463.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception & Performance*, *35*(3), 718–734.
Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin and Review*, *18*(5), 855–859.
Helton, W., & Russell, P. N. (2012). Brief mental breaks and content-free cues may not keep you focused. *Experimental Brain Research*, *219*(1), 37–46.
Hillyard, S. A., & Münte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Perception and Psychophysics*, *36*, 185–198.
Hollingworth, A., Matsukura, M., & Luck, S. J. Visual working memory modulates rapid eye movements to simple onset targets. *Psychological Science*, in press.
Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.
Jaensch, E. R. (1929). *Grundformen menschlichen Seins*. Berlin: Otto Elsner.
Kastner, S., O'Connor, D. H., Fukui, M. M., Fehd, H. M., Herwig, U., & Pinsk, M. A. (2004). Functional imaging of the human lateral geniculate nucleus and pulvinar. *Journal of Neurophysiology*, *91*(1), 438–448.
Kim, S., Kim, M., & Chun, M. (2005). Concurrent working memory load can reduce distraction. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16524–16529.
Kirchner, H., & Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776.
Lachter, J., Forster, K. I., & Ruthruff, E. (2004). Forty years after Broadbent: still no identification without attention. *Psychological Review*, *111*, 880–913.
Lang, P., Bradley, M., & Cuthbert, B. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8, Gainesville, FL: University of Florida.
Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*(3), 339–354.
Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, *19*(3), 143–148.
Lin, J., Franconeri, S., & Enns, J. (2008). Objects on a collision path with the observer demand attention. *Psychological Science*, *19*(7), 686–692.
Lleras, A., & Buetti, S. (2012a). Ocular capture by abrupt onsets is severely reduced during mental arithmetic: evidence against cognitive load theory. *Paper Presented at the 53rd Annual Meeting of the Psychonomic Society*.
Lleras, A., & Buetti, S. (2012b). Where do the eyes go when you think? Away from visually salient information. *Paper presented at the 12th Annual Meeting of the Vision Sciences Society*, Naples, Florida. 12(9), 1261, http://dx.doi.org/10.1167/12.9.1261.
Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.
Macdonald, J., & Lavie, N. (2008). Load induced blindness. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1078–1091.
Macdonald, J., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception & Psychophysics*, *73*(6), 1780–1789.
Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.
Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, *22*(6), 657–672.
Maljkovic, V., & Nakayama, K. (1996). Priming of pop-out: II. The role of position. *Perception & Psychophysics*, *58*(7), 977–991.
Max, R., & Tsal, Y. (2011). Efficient selection is modulated by target-distractor discriminability and not by perceptual load. *Poster presented at the 52nd Annual Meeting of the Psychonomic Society*, Seattle, WA, November 4, 2011.
Max, R., & Tsal, Y. (2012). Exact temporal locus of visual distraction. *Poster presented at the 53rd Object Perception Attention and Memory Workshop*, Minneapolis, MN, November 15, 2012.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the FFA related to perceptual performance. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), 17063–17068.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(1), 196–204.

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "D'oh!": working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 525–549.

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247–279.

Miller, J. (1987). Priming is not necessary for selective-attention failures: semantic effects of unattended, unprimed letters. *Perception & Psychophysics*, *41*, 419–434.

Moore, C., & Egeth, H. (1997). Perception without attention: evidence of grouping under conditions of inattention. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 339–352.

Moore, C., Grosjean, M., & Lleras, A. (2003). Using inattentional blindness as an operational definition of unattended: the case of surface completion. *Visual Cognition*, *10*, 299–318.

Moore, C., Lleras, A., Grosjean, M., & Marrara, M. (2004). Using inattentional blindness as an operational definition of unattended: a response-end effect. *Visual Cognition*, *11*, 705–719.

Mordkoff, J. T., & Halterman, R. (2008). Feature integration without visual attention: evidence from the correlated flankers task. *Psychonomic Bulletin and Review*, *15*, 385–389.

Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 520–538.

Mordkoff, J. T. (1996). Selective attention and internal constraints: there is more to the flanker effect than biased contingencies. In A. F. Kramer, M. G. H. Coles & G. Logan (Eds.), *Converging operations in the study of visual selective attention* (pp. 483–502). Washington, DC: APA.

Most, S., Simons, J., Scholl, B., Jimenez, R., Clifford, E., & Chabris, C. (2001). How not to be seen: the contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science*, *12*(1), 9–17.

Most, S., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What you see is what you set: sustained inattentional blindness and the capture of awareness. *Psychological Review*, *112*(1), 217–242.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*(4), 605–617.

Neider, M., & Zelinsky, G. (2011). Cutting through the clutter: searching for targets in evolving complex scenes. *Journal of Vision*, *11*(14).

Neisser, U., & Becklen, R. (1975). Selective looking: attending to visually specified events. *Cognitive Psychology*, *7*(4), 480–494.

Neisser, U. (1979). The control of information pickup in selective looking. In A. D. Pick (Ed.), *Perception and its development: A tribute to Eleanor J. Gibson* (pp. 201–219). Hillsdale, NJ: Lawrence Erlbaum.

Nestor, A., Plaut, D., & Behrmann, M. (2012). Orthographic form processing—a multivariate investigation of its neural basis. *Cerebral Cortex*, Cereb. Cortex firxt published online June 12, 2012, http://doi.10.1093/cercor/bhs158.

Norman, D. A. (1968). Towards a theory of memory and attention. *Psychological Review*, *75*, 522–536.

O'Connor, D., Fukui, M., Pinsk, M., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*(11), 1203–1209.

Pashler, H. (1987). Detecting conjunctions of color and form: reassessing the serial search hypothesis. *Perception & Psychophysics*, *41*(3), 191–201.

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 1–20.

Purcell, B., Heitz, R., Cohen, J., Schall, J., Logan, G., & Palmeri, T. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*(4), 1113–1143.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574–590.

Raymond, J., Shapiro, K., & Arnell, K. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860.

Rensink, R. (2000). Seeing, sensing, and scrutinizing. *Vision Research*, *40*(10–12), 1–19.

Scalf, P. E., & Beck, D. M. (2010). Competition in visual cortex impedes attention to multiple items. *Journal of Neuroscience*, *30*(1), 161–169.

Schall, J., Purcell, B., Heitz, R., Logan, G., & Palmeri, T. (2011). Neural mechanisms of saccade target selection: gated accumulator model of the visual-motor cascade. *European Journal of Neuroscience*, *33*(11), 1991–2002.

Serences, J., & Boynton, G. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 1–12.

Shapiro, K., Caldwell, J., & Sorensen, R. (1997). Personal names and the attentional blink: a visual "cocktail party" effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 504–514.

Shiffrin, R. M., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology*, *93*(1), 72–82.

Silvia, P., McCord, D., & Gendolla, G. (2010). Self-focused attention, performance expectancies, and the intensity of effort: do people try harder for harder goals? *Motivation and Emotion*, *34*, 363–370.

Silvia, P., Jones, H., Kelly, C., & Zibaie, A. (2011). Trait self-focused attention, task difficulty, and effort-related cardiovascular reactivity. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *79*(3), 1–6.

Simon, J., & Rudell, A. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304.

Simon, J., Acosta, E., Mewaldt, S., & Speidel, C. (1976). The effect of an irrelevant directional cue on choice reaction time: duration of the phenomenon and its relation to stages of processing. *Perception & Psychophysics*, *19*, 1–7.

Simons, D., & Chabris, C. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Simons, D., & Jensen, M. (2009). The effects of individual differences and task difficulty on inattentional blindness. *Psychonomic Bulletin and Review*, *16*(2), 398–403.

Simons, D., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review*, *5*, 644–649.

Smallwood, J., Obonsawin, M., & Heim, D. (2003). Task unrelated thought: the role of distributed processing. *Consciousness and Cognition*, *12*(2), 169–189.

Smilek, D., Carriere, J. S.A., & Cheyne, J. A. (2010). Out of mind, out of sight: eye blinking as indicator and embodiment of mind wandering. *Psychological Science*, *21*(6), 786–789.

Smith, T., Ruiz, J., & Uchino, B. (2000). Vigilance, active coping, and cardiovascular reactivity during social interaction in young men. *Health psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *19*(4), 382–392.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*.

Strayer, D. L., & Drews, F. (2007). Cell-phone induced inattention blindness. *Current Directions in Psychological Science*, *16*, 128–131.

Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive distraction while multitasking in the automobile. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 54, pp. 29–58). Burlington: Academic Press.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Temple, J., Warm, J., Dember, W., Jones, K., LaGrange, C., & Matthews, G. (2000). The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. *Human Factors*, *42*(2), 183–194.

Tinker, M. A. (1925). Intelligence in an intelligence test with an auditory distractor. *American Journal of Psychology*, *36*(3), 467–468.

Torralbo, A., & Beck, D. (2008). Perceptual-load-induced selection as a result of local competitive interactions in visual cortex. *Psychological Science*, *19*(10), 1045–1050.

Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical and Statistical Psychology*, *25*, 168–199.

Trautwein, U., Lüdtke, O., Roberts, B., Schnyder, I., & Niggli, A. (2009). Different forces, same consequence: conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of Personality and Social Psychology*, *97*(6), 1115–1128.

Treisman, A., & Geffen, G. (1967). Selective attention: perception or response? *Quarterly Journal of Experimental Psychology*, *19*(1), 1–17.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*, 242–248.

Tseng, Y. C., Glaser, J. I., Caddigan, E., & Lleras, A. *Attentional decision making in pop-out visual search*, Submitted for publication.

von Mühlenen, A., & Lleras, A. (2007). No-onset looming motion guides spatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1297–1310.

Wan, X., & Lleras, A. (2010). The effect of feature discriminability on the inter-trial inhibition of focused attention. *Visual Cognition*, *18*, 920–944.

Watson, D. G., & Humphreys, G. W. (1997). Visual marking: prioritizing selection for new objects by top-down attention inhibition of old objects. *Psychological Review*, *104*, 90–122.

Wenger, M. J., & Townsend, J. T. (2000). Basic tools for attention and general processing capacity in perception and cognition. *Journal of General Psychology: Visual Attention*, *127*(1), 67–99.

Wolfe, J., Alvarez, G., Rosenholtz, R., Kuzmova, Y., & Sherman, A. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception & Psychophysics*, *73*(6), 1650–1671.

Wolfe, J., Vo, M., Evans, K., & Greene, M. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84.

Wolfe, J. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review*, *1*(2), 202–238.

Woodman, G. F., & Arita, J. T. (2011). Direct electrophysiological measurement of attentional templates in visual working memory. *Psychological Science*, *22*(2), 212–215.

Yantis, S., & Egeth, H. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 661–676.

Yigit-Elliott, S., Palmer, J., & Moore, C. M. (2011). Distinguishing blocking from attenuation in visual selective attention. *Psychological Science*, *22*(6), 771–780.

Young, A. H., & Hulleman, J. (2012). Eye movements reveal how task difficulty moulds visual search. *Journal of Experimental Psychology: Human Perception and Performance*.

This page intentionally left blank

INDEX

Note: Page numbers followed by "f" and "t" indicate figures and tables respectively.

D

E

I

M

Q

R

S

T

U

V

W

This page intentionally left blank

CONTENTS OF PREVIOUS VOLUMES

VOLUME 40

VOLUME 41

VOLUME 42

VOLUME 43

VOLUME 44

VOLUME 45

VOLUME 46

VOLUME 47

VOLUME 48

VOLUME 49

VOLUME 50

VOLUME 51

VOLUME 52

VOLUME 53

VOLUME 54

VOLUME 55

VOLUME 56

VOLUME 57

VOLUME 58

This page intentionally left blank