CHOICE BLINDNESS

PETTER JOHANSSON

Choice Blindness

Choice Blindness

The incongruence of intention, action and introspection

Petter Johansson

Lund University Cognitive Studies 130

For further information, see authors' webpage:
www.lucs.lu.se/people/petter.johansson

Cover and book design: David de Léon
Cover photo: Mark Hanlon

Printed in Sweden, Media-Tryck
ISBN 91-628-6954-X
ISSN 1101-8453
ISRN LUHFDA/HFKO-1017-SE
Lund University Cognitive Studies 130

To my family, and my family of friends

"Era uma vez, em um reino distante, cientistas mostraram a voluntários alguns pares de fotografias de rostos de mulheres. 'Qual lhes parece mais atraente?' os cientistas perguntavam. Quando o voluntário revelava sua escolha, os cientistas então pediam que ele descrevesse verbalmente as razões para explicar sua escolha.

Mas o que os voluntários não sabiam é que os cientistas eram traquinas, e algumas vezes usavam um passe de mágica para trocar as fotos depois que a escolha havia sido feita. Assim, pediam ao voluntário para explicar por que havia escolhido o rosto que, em verdade, não havia escolhido."

"It was a time, in a distant kingdom, scientists had shown to the volunteers some pairs of photographs of faces of women. 'Which them seems more attractive', the scientists asked. When the volunteer disclosed its choice, the scientists then asked for that it described the reasons verbally to explain its choice.

But what the volunteers did not know it is that the scientists were traquinas [rascals], and some times used a magician pass to change the photos later that the choice had been made. Thus, they asked for to the volunteer to explain why it had chosen the face that, in truth, it had not chosen."

An article about choice blindness in Portuguese, automatically translated to English through Babelfish

CONTENTS

Publication histories *xi*
Acknowledgements *xiii*

Introduction 1

Paper one
From Change Blindness to Choice Blindness 43

Paper two
Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task 63

Paper two: Appendix
Supporting Online Material 73

Paper three
How Something Can be Said About Telling More Than We Can Know 93

Paper four
Magic at the Markeplace 135

Publication Histories

The publication histories for the papers included in the thesis are as follows:

Paper one

Paper one is based on the following presentations:

Johansson, P., Hall, L., & Olsson, A. (2004). *From change blindness to choice blindness*. Towards a Science of Consciousness 2004, Tucson, Arizona, April 7–11, 2004.

Hall, L., Johansson, P., Olsson, A., & Sikström, S. (2004). *Choice blindness and verbal report*. The Association for the Scientific Study of Consciousness, 8th Annual Meeting, University of Antwerp, Belgium, June 25–28, 2004.

Johansson, P., Hall, L., Olsson, A., & Sikström, S. (2004). *Facing changes: choice blindness and facial attractiveness*. The 28th International Congress of Psychology, Beijing, China, August 8–13, 2004.

Paper two

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*, 116-119.

Paper three

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (in press). How something can be said about telling more than we can know. *Consciousness and Cognition*.

Paper four

Hall, L., Johansson, P., Tärning, B. Deutgen, T., & Sikström, S. (2006). Magic at the marketplace. *Lund University Cognitive Studies, 129*.

Acknowledgements

I have during the years used many different images to describe what it is like to write a thesis. My present favourite is from Werner Herzog's movie *Aguirre: The Wrath of God*, starring Klaus Kinski. Against all advice, he sets out on a small raft to find El Dorado – the legendary land of gold. The only map used is his hunches and hopes. Carried by the currents of the Amazon River, he travels further and further into the wilderness. But the quest is a disaster. There are no signs of gold; there are no signs of anything. Each day the same, day after day. The river runs fast, but only in one direction. And there is no turning back. In the end, he stands alone on the raft. Stark mad, raving about future riches and rewards.

I now see that my case differs in at least two respects. I may not have found gold, but at least I came ashore. The journey ends here. Secondly, I have not been alone on the raft. Rather than dying off, the crew has been constantly growing. So the following are the people I would like to thank for cheering me on or keeping me company on this trip.

First of all, Peter Gärdenfors, professor of the department of Cognitive Science at Lund University and my first supervisor. I would actually like to thank him most for what he did last. Our long discussions regarding the rhetorical composition of the introduction and the thesis were both fun and very valuable. On a more general level, I would like to thank him for creating the special atmosphere of enthusiasm and intellectual curiosity we have at the department. This is something he is often praised for, but it is a credit he deserves.

Next in line is my current supervisor Sverker Sikström. I would like to thank him for his empirical know-how and the scientific rigor he added to our team. You might not have noticed, but I have learnt a lot of things from you.

I would also like to thank all the past and present members of the Cognitive Science department, for all the seminars and the discussions we have had over the years: Martin Bergling, Petra Björne, Nils Dahlbäck, Philip Diderichsen, Pierre Gander, Agneta Gulz, Kenneth Holmqvist, Jana Holsánová, Nils Hulth, Birger Johansson, Magnus Johansson, Paulina Lindström, Petter Kallioinen, Peter Kitzing, Lars Kopp, Maria Larsson, Jan Morén, Mathias Osvath, Tomas Persson, Annika Wallin and Jordan Zlatev.

Christian Balkenius deserves special credit for helping us with the diagrams for Paper 2 in the thesis, a template I have shamelessly used for all work done since. He has also been my interface to the world of Mac, and has given priceless help and advice when I have been in a tight spot the hours before a deadline without a clue what to do.

I would also like to thank our brilliant secretary (well, you are the best!) Eva Sjöstrand, for always taking care of everything that needs to be taken care of.

There are also a number of people I have collaborated with when writing the papers in this thesis. Andreas Olsson is co-author on Paper 2, and I thank him for his input on how to best go from our results and ideas to a proper article. Betty Tärning, co-author on paper 3 and 4, has been invaluable as she has performed experiments, catalogued data, as well as transcribed verbal reports. Apart from this, she has contributed greatly in our often long and sometimes seemingly aimless group discussion on what to do with choice blindness. Andreas Lind was essential in the making of Paper 3, both with his linguistic expertise as well as his tireless devotion to getting things done (I agree, it is more fun to work hard). Finally, Thérèse Deutgen played a large part when making magic at the marketplace for Paper 4.

I would also like to thank a number of people at the department of linguistics at Lund University for all their help and advice concerning the linguistic analyses performed in Paper 3: Mats

Andrén, Victoria Johansson, Joost van de Weijer and Jordan Zlatev. While some of you might not agree with our interpretation of the data, I hope you still thought our project was of some interest.

In addition to those that had to, several people read and commented on my introduction. I would like to thank you all for your time, I really appreciated your input: Lars Brink, Ingegärd Johansson, Markus Karlsson, Peter Kitzing and Björn Peterson.

The people that have meant most to me in my daily life writing this thesis are David de Léon, Jens Månsson and Lars Hall. Through the long lunches and late nights, you guys have made being at work fun. I especially thank David for all the artefactual intelligence you put into the making of this book. I admire your eye for beauty, and share your appreciation for the good in life. And Lars, I am not going to even try to list the things I thank you for. You are a brilliant man, and my dearest friend.

*

There are also several people and institutions I would like to mention outside the immediate environment of Lund University.

I would like to thank *Bank of Sweden Tercentenary Foundation* for funding large parts of my Ph.D. studies. This was a separate project from the thesis, which resulted in the edited volume Gärdenfors and Johansson (2005). *Cognition, Education and Communication Technology*. Lawrence Erlbaum Associates.

I would similarly like to thank *The New Society of Letters at Lund* for financing a year at University of East London.

I thank UEL for hosting me that year, and especially my personal host Tom Dickins for making my stay very rewarding on a both personal and professional level. He taught me that “Evolution is the Way and the Truth” and that “Clarity is All” – I hope he is not too disappointed by the near absence of both these things in the thesis. At EUL I also realised that the pub is the best place for academic discourse, and in addition to Tom I would also like to especially thank Eike Adams, Chris Pawson and Qazi Rahman for teaching me this.

Professor Nick Humphrey at London School of Economics invited me to attend his weekly work-in-progress seminar on evolutionary psychology. I thank him for this as well as for letting me present my own ideas at this forum.

I would also like to thank Julie McNiff, for giving me warmth and shelter, and for showing me how London should be lived.

*

Peter Rosengren guided us through the fine arts of card magic; I thank him for this (and for not laughing at our clumsy efforts). Having a real magician in the group has been the source of envy for colleagues all over the world. I would also like to thank Karl Berseus and Axel Adlercreutz for our long discussions at the “Swedish Magicians’ Circle” conference, regarding how to make objects disappear.

Anders Hall deserves special credit for all his help with my computers over the years. The final act was his efforts the last day before handing in my thesis (earlier this morning!), when Windows suddenly decided to implode. He saved me today, as he has done many times before.

I would like to thank my old friend Magnus Dittmer for all the summer weeks I spent with you and Marie, and for giving me some perspective on life – there are many ways to live it happily.

The last few years, I have shared apartment with Markus Karlsson, Hans Appelqvist, Hilderun Gorpe and Malin Skoglund. I would like to thank them for being part of a home that has made me want to return from work as well.

*

Finally, I would like to thank my closest family.

First of all, my girlfriend Marie. To be we is the best part of my life. Finishing this thesis was always going to be hard, but knowing that I had already won made it so much easier. You know I will always point at your picture.

My aunt Eva and my grandmother Elsa, for all the light and laughter, and for caring so much for me.

My sister Lovisa and her family: Lars, Rasmus and Klara. For letting me share your space, the warmest and safest place in the world.

My father Leif, who taught me the beauty of obsession. It is not what you do but how you do it that matters.

My mother Ingegärd and Bo, who taught me to love thoughts, but also convinced me to try to put some data in. You are right: no one will listen if I just talk. Thank you mum, for always being there.

And to not exclude anyone, I would also like thank myself, for pulling this off without falling apart.

Introduction

Look at the two faces on the book cover. Try to decide which one of them you find more attractive. After you have made up your mind, focus on the face you preferred, and explain to yourself why you liked that one better. Now imagine I told you that you actually preferred the other face. After your decision – but before you started to talk – I switched the position of the pictures, so you are now looking at the face you did *not* choose. When you gave your reasons you were in fact looking at the opposite of your choice.

Would you take my word for it, or would you find it hard to believe?

If you think you would have noticed the manipulation you are not alone; this is what most people think. But however unlikely it may seem, it is not at all certain that you *would* have seen the switch. And had you not seen when the pictures changed places, you are also quite likely to give a long and elaborate description of why you chose this face and not the other.

Despite its brevity, this scenario contains all the major components of the thesis. The work presented is an empirical and theoretical exploration of the finding that people are prone to miss even dramatic mismatches between what they want and what they get. The fact that it is possible to manipulate the relation between people's intentions and the outcome of their actions *without them noticing* is what my collaborators and I have dubbed *choice blind-*

ness. This effect is demonstrated in a series of experiments, using both different stimuli and different experimental methods.

But not only were the participants in our experiments blind to the manipulation of their choices, they also offered introspectively derived reasons for preferring the alternative they were given instead. The second major component of this thesis is thus the participants' verbal reports explaining choices they did not intend to make. These reports are analysed both in isolation and in relation to reports from non-manipulated choices. By comparing the content of the verbal reports with the properties of the chosen items it is possible to establish that the reports are sometimes "confabulatory" – i.e. when the participants refer to unique features of the initially non-preferred face (e.g. a pair of earrings) as being the reason for choosing this alternative rather than the other. As an additional finding, the reports stemming from manipulated choices seem to be just as rich and elaborate as the ones given in non-manipulated trials.

Finally, I consider the experimental methodology to be a finding of its own. We have created a number of different experimental procedures in which we generate a mismatch between what the participants intend to choose and the outcome they experience as being their choice. By using a binary choice task, we can always be certain that our participants actually wanted the opposite of what they were given. All the empirical work presented shares these general characteristics.

Thus, the three things I see as novel in the thesis are the choice blindness effect, the verbal reports based on manipulated choices, and the experimental approach as such. Throughout the book and the rest of the introduction, this is what it is all about.

In this introduction, each of the four papers is presented with a very compressed descriptive recapitulation of the experiments, the results and the conclusions drawn. The papers are then discussed in terms of related topics and theory, organised around the three major themes identified above.

I consider Paper 1 and 2 as well as the *Supporting Online Material* accompanying Paper 2 as belonging to the same project, and they will be summarised and discussed together in relation to the choice blindness effect. The theoretical backdrop for this discussion is the nature of *Folk Psychology*, and the use of belief-desire explanations in cognitive science modelling. I will argue that our results represent a substantial problem for philosophers and cognitive scientists that connect their models too closely to a Folk Psychological model of the mind. As such, the choice blindness effect challenges the commonsense assumption that beliefs, desires and intentions, are entities in the brain. Instead, our results are better understood within the framework of the *Intentional Stance* (Dennett, 1987), in which beliefs and desires are seen as predictive tools we use in our attempts to make sense of ourselves and others.

The discussion of Paper 3 will be focused on the analyses of the verbal reports. The natural context of this discussion is the perennial battle in psychology and philosophy regarding the validity of introspective self-reports. Extra attention is given to the debate following the publication of Nisbett and Wilson (1977), an article which strongly questions the accuracy of introspection. I will argue that while our results can be given a similar interpretation as was given Nisbett and Wilson's, our experimental method is a significant step forward. Still, one conclusion must be that that our results indicate that we know a lot less about ourselves than we think we do.

In relation to Paper 4, I expand on the idea of using our experimental approach as a more general research tool, and give a glimpse of future studies planned.

Choice Blindness

I consider Paper 2 to be the centrepiece of the thesis, and the paper best served to introduce the approach as a whole. I will therefore start with the summary of Paper 2.

Summary Paper 2: ***Failure to detect mismatches between intention and outcome in a simple decision task.*** The participants in the study of Paper 2 were shown two pictures of female faces, and were instructed to point at the face they found most attractive. After pointing, the chosen picture was given to the participants, and they were asked to explain why they preferred the picture they now held in their hand. Unknown to the participants, using a double-card ploy, the pictures were sometimes covertly exchanged mid-trial. Thus, on these trials, the outcome of the choice became the opposite of that intended by the participants (see Figure 1 in Paper 2).

Each of the 120 participants performed 15 choice trials, of which three were manipulated. The time given to make a choice, and the similarity of the face-pairs were varied. For time, three choice conditions were included: one with two seconds of deliberation time, one with five, and a final condition where participants could take as much time as they liked. For similarity, a high and a low similarity set of target faces was used.

A trial was classified as detected if participants showed any signs of detection in immediate relation to the switch (such as explicitly reporting that the faces had been switched, or indicating that something went wrong with their choice), or if the participants voiced any suspicion in the debriefing session after the experiment.

Counting all forms of detection across all experimental conditions, no more than 26% of the manipulated trials were detected. There were no significant differences in detection rate between the two groups of stimuli used. For viewing time, the 2-second and 5-second conditions did not differ in detection rate, but there were significantly more detections in the free viewing time condition.

The verbal reports were also recorded, transcribed and analysed. Of primary interest is the relation between the reports given in manipulated and non-manipulated trials. In the non-manipulated trials the participants just answered why they had preferred the chosen picture, but when doing the same thing in the non-detected manipulated trials the participants described and gave reasons for a choice they did not intend to make. The two classes of reports were analysed on a number of different dimensions, such as the level of emotionality, specificity and certainty expressed, but no substantial differences between manipulated and non-manipulated reports were found.

The experiment also established the extent to which a report could be matched to the picture originally chosen or to the manipulated outcome received – i.e. if the participants talked about the face they thought more attractive first or the one they ended up with after the switch was performed. The conclusion drawn in Paper 2 is that the relationship between intention and outcome may sometimes be far looser than current theorising has suggested. As such, choice blindness warns of the dangers of aligning the technical concept of intention too closely with commonsense. The analyses of the verbal reports shows that in some trials we can be certain that the participants confabulate or construct their answers in line with the manipulations made, as they refer to unique properties of the initially non-preferred face. The lack of differentiation between the manipulated and non-manipulated reports casts doubt on the origin of the non-manipulated reports as well; confabulation could be seen to be the norm and truthful reporting something that needs to be argued for.

The Supporting Online Material functions as an appendix for Paper 2. Several aspects are expanded and detailed, such as the experimental procedure, statistical measures used, detection criteria, the analyses of the verbal reports, and the relation to previous studies.

Summary Paper 1: *From change blindness to choice blindness.* Paper 1 is a precursor to Paper 2, in terms of both theory and empirical method. The participants either had to choose which of two abstract patterns they found most aesthetically appealing or which of two pictures of female faces they found most attractive. Fifteen trials were used, of which three were manipulated. The choice task was presented on a computer screen, and the participants had to indicate their choice by moving the cursor to the chosen picture. When all the choice trials were completed, an unannounced memory test was introduced. The participants had to look at all the pairs again, without time-constraint, and try to remember which face or pattern they previously preferred. The result was similar to that in Paper 2, as the participants showed considerable levels of choice blindness. The memory test revealed that the participants had been influenced by the manipulations made, and tended to remember the manipulated outcome as the alternative they originally preferred.

Surprise, surprise

So why do I think our experimental results are an interesting finding? From a commonsense perspective, choice blindness seems a baffling phenomenon. How can someone choose *x*, and then not notice when given *y* instead? Do we not know what we want when we make a choice? Given the lack of similarity between the faces (see Picture 1 in SOM), how is it possible *not* to notice if they are swapped? This does not seem to fit well with our ordinary intuitions of how we function.

But it is not just the description of the experiment and the results that people find surprising. In the debriefing session after the experiment in Paper 2, all participants were asked a series of increasingly specific questions to investigate whether they suspected in any way that something had gone wrong ("What did you think about the experiment?", "Did you find anything odd about the experiment?" and "Did you notice anything strange about the stimuli presented in the experiment?"). Participants who revealed no signs of detection were then presented with a hypothetical scenario describing an experiment in which the faces they choose between are secretly switched (i.e. the very experiment they had just participated in), and asked whether they thought they would have noticed such a change. The result shows that, of the participants who failed to notice any of the manipulations, 84% believed that they *would* have been able to do so. Accordingly, many participants also showed considerable surprise, even disbelief at times, when we debriefed them about the true nature of the design. We call this effect "choice blindness blindness"; i.e. the overconfidence in our own ability to detect choice-manipulations (For a similar meta-cognitive error in relation to change blindness, see Scholl, Simons, & Levin, 2004). In my opinion, this is also the strongest evidence there is that we have discovered something genuinely contra-intuitive.[1]

Our commonsense intuitions are also a good starting point for a more theoretically grounded discussion of choice blindness. In philosophy and cognitive science, the totality of our everyday psychological explanations is referred to as Folk Psychology (Bogdan, 1991; Christensen & Turner, 1993; Greenwood, 1991). When we try to make sense of other people, or when we answer

[1] This is also a strong argument with regards to the question how we can know that the participants really did not detect the manipulations. Maybe the participants saw all manipulations but just did not tell us? But to first confidently claim that they think they would have noticed a switch, and then "feign" surprise and deliberately lie when asked if they saw the manipulations, is something that seems a very odd thing to do. In addition, counting all experiments mentioned or described in this thesis, we have tested around 470 participants and classified around 790 trials as non-detected choice manipulations. It does not seem likely that we have misclassified all of them. The issue of forms and levels of detection is further discussed in Supporting Online Material and in an interchange with a commentary on Paper 3 (Hall, Johansson, Sikström, Tärning & Lind, in press; More & Haggard, in press).

questions such as *why* we preferred one picture over another, we phrase these answers in mental state descriptions such as beliefs and desires. For example:

"She must think that no one can see her through the window."
"Probably, he just really really wanted that apple."
"I thought you thought that I believed you to be innocent."

We are all experts on Folk Psychology, it is a language we are fluent in from a very early age. When examined more closely, Folk Psychological descriptions have certain law-like regularities, such as:

If X wants Y and believes that it is necessary to do Z to get Y, X will do Z

If Petter wants ice-cream and believes it is in the freezer, he will open the freezer and take one out. But they work as explanations as well as predictions – if Petter is seen opening the freezer and taking an ice-cream, he most likely wants ice-cream as well. We use the framework of Folk Psychology all the time, to understand and make sense of both ourselves and others. We believe, desire, intend, want, hope, think, fear, etc. But despite being a seemingly indispensable tool for understanding and interacting with each other, our Folk Psychological constructs are problematic entities. What exactly *are* beliefs, desires and intentions?[2]

[2] But we sometimes feel the limits of Folk Psychology. In the 110th minute of the 2006 world cup final, Zinedine Zidane suddenly head-butts Marco Materazzi and is sent off. The most celebrated player of the modern era; the captain of the French team; a true hero of the people. He declared that he would retire after the tournament, and then defied age and expectations and played some of the best games of his career. And in the last act, he puts his entire legacy at risk. More than a billion spectators sit stupefied in front of the television set. *Why did he do it?* In the replay it is clear that the Italian defender says something. Zidane hesitates for almost a second, as if contemplating the alternatives, and then charges. The only possible explanation is the words said, but how can they have had the force to make him do what he did? From a Folk Psychological perspective, it is interesting to note that everyone agreed that it *must* have been something extremely offensive or vile, or some deeply personal matter. The magnitude of the insult does not only need to match the future consequences disregarded, it is also the personality we have pinned on Zidane after getting to know him for 15 years watching him play. Still, in this case, it does not feel like we will ever understand the action.

The Great Divide

The status and nature of Folk Psychology is an old battleground in philosophy of mind and cognitive science, crisscrossed with trenches and fronts opened in all directions. What everyone seems to agree on is that Folk Psychology is a very powerful tool in explaining and predicting people's behaviour, but apart from that, they disagree on just about everything else. Philosophers argue about how well it actually functions as a coherent scientific theory (Churchland, 1981), while developmental psychologists disagree on both how and when we acquire the mental concepts we use in later life (Astington, 1993; Gopnik, 1993). Primatologists discuss to what extent our near neighbours share our belief-desire type of "theory of mind" (Premack & Woodruff, 1978), while others argue whether a "theory of mind" is a prerequisite for the development of a Folk Psychology in the first place (Baron-Cohen, 1994, 1995).

But there are two main threads in the debate. What do we *do* when we understand each other using the conceptual framework of Folk Psychology, and to what extent does the theory of Folk Psychology correspond to what is actually going on in the mind (or the brain)? Regarding the first question, there are two major positions: You are either a theory-theorist and argue that we apply the *theory* of Folk Psychology as any other theory when we explain what people do (Gopnik, 1993), or you are a simulation-theorist and argue that we primarily understand each other through a kind of mental role-playing in which we put ourselves in other people's position and thereby "experience" what mental states they are likely to have (see e.g. Goldman, 1993). I will return to this question briefly when discussing introspection in the next chapter summary, but in relation to choice blindness the second question is more important. So, in what sense do the entities of Folk Psychology such as beliefs, desires and intentions exist?

The philosophical position that assumes Folk Psychology to describe real things residing in the head is called Intentional Realism, and the foremost champion of this doctrine is Jerry Fodor (1983, 2000). According to him, it is Folk Psychology all the way in. The reason Folk Psychology works so well is because it happens to be true. In a distant future, when we have mapped out the workings of the brain, we will find the equivalents of beliefs and desires.

They will be discovered to be the fundamental building blocks in the internal cognitive machinery that governs our behaviour. He is an adamant defender of this position, and he does not take his job lightly: "if commonsense psychology were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species" (1987, p. xii). Despite this, I have my allegiances elsewhere.

Daniel Dennett explains both what Folk Psychology is and how we use it within the same theoretical framework: The Intentional Stance. Some additional background is necessary to appreciate his position. Dennett (1987) presents a taxonomy of stances or viewpoints from which to predict or understand any system. First we have the Physical Stance, from which systems are predicted by exploiting information about their physical constitution. Since, in the end, humans are nothing more than extremely complex physical systems we are in principle predictable with this method. Next we have the Design Stance, from which one understands the behaviour of a system by assuming it is composed of elements with functions, i.e. that it has a certain design, and that it will behave as it is designed to do under various circumstances. Finally, there is the Intentional Stance, from which one predicts a system by treating it as an approximation of a rational agent. We attribute the beliefs, desires and goals it ought to have, taking into consideration previous actions, verbal statements and available options[3].

[3] To complicate things further: in philosophy there is also an underlying debate on the nature of *intentionality*, which is a technical concept referring to the ability of one thing to be *about* something else. The word "turnip" refers to a specific vegetable; a turnip as such can not refer. Apart from words and symbols, mental entities can also be about other things: I believe there is gold at the end of the rainbow, I think about what to eat for lunch. Brentano (1874/1973) famously stated that as physical objects cannot be about other things, but mental states can, mental states cannot be reduced to physical states or entities – *the irreducibility of the mental*. This can either be interpreted as supporting some form of dualism (Chisholm, 1966); or that in an absolute sense, mental states do not exist and therefore we cannot have a proper science about them (Quine, 1960). Both Fodor and Dennett opt for other alternatives: Fodor claims that mental states *are* physical states and get their meaning or content through causal links to the objects they refer to, while Dennett agrees with Quine that beliefs and desires do not exist as objects, but claims them to exist as relations seen from the intentional stance, whose content ultimately can be derived from the rationality presupposed by an evolutionary perspective. Much abbreviated – depending on your perspective this debate is either extremely important or largely irrelevant for the present thesis, but I will nevertheless relate it no further here.

The Intentional Stance is in Dennett's view the backbone of our Folk Psychology, and it is the rationality assumption that is the guiding principle when we create a psychological explanation. In his own words:

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (Dennett, 1987; p. 17)

Within this framework, every system that can be profitably treated as an intentional system by the ascription of beliefs, desires, etc., also *is* an intentional system in the fullest sense (see Dennett, 1987; 1991a). Not just human beings but countries, banks, butterflies – even the lowly thermostat – have beliefs and desires if we gain any predictive leverage from ascribing such states to them. Dennett is thus very inclusive regarding what can be considered to *have* beliefs and desires, as well as what should be considered to *be* beliefs and desires. They exist as patterns in the world, to be seen from the Intentional Stance (Dennett, 1991b). With this perspective, it is not surprising that he does not think that belief-desire prediction reveals the exact internal machinery responsible for the behaviour.

> We would be unwise to model our scientific psychology too closely on these putative *illata* (concrete entities) of folk theory. We postulate all these apparent activities and mental processes in order to make sense of the behavior we observe – in order, in fact, to make as much sense possible of the behavior, especially when the behavior we observe is our own [...] each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing. (Dennett, 1987; p. 91, emphasis in original)

When we explain our own behaviour in terms of Folk Psychology, we do this by applying the Intentional Stance towards ourselves as well. I observe myself and interpret my actions rather than getting to know my beliefs and desires from the inside. And after explaining a certain act and having clad my behaviour in words, the description of the mental entities deemed responsible for my actions now has a concrete existence not previously enjoyed: "The intentions are as much an effect of the process as a cause – they emerge as a product, and once they emerge, they are available as standards against which to measure *further* implementation of the intentions" (Dennett, 1991a, p. 241, emphasis in original).

If we take a look at our experiment, the behaviour of the participants seems to make sense given Intentional Stance theory. What the participants seems to be doing is to make interpretations. They see themselves act, and assume that the picture they reached for and were given also was the picture they intended to choose. All the external evidence points in this direction, it is a reasonable conclusion to draw given the circumstances. They took the card, so they must have wanted it. But the things they say do not need to be actual descriptions of what went on in their heads prior to the decision. Some kind of decision-making process made them choose one face over the other, but the "reasons" responsible for this do not need to correspond to the things they say. And there need not be any higher-order intention *in the brain* to choose one face over the other, the outcome of the internal evaluation might only result in the motor act of pointing to the face preferred. The reasons the participants give is their own interpretation of *why* they must have wanted this picture rather than the other. In a sense, they inform themselves as much as everybody else about what they wanted when they perform and then explain their actions.

It is of course hard to draw any strong ontological conclusions from our experiments; it would be silly to say that we have shown Intentional Realism to be false. But it is also quite evident that our results better fit Dennett's perspective than Fodor's. If Folk Psychology is an instrument of interpretation, it should be possible to make "mistakes" about ourselves – e.g. to make a belief-desire interpretation that does not fit with the lower-level implementation of the action. As it now stands, one possible explanation why the participants in our experiments did not detect the mismatch between their intentions and the outcome of their actions could simply be that the prior intentions (as conceptualised by Intentional Realism) do not exist. Intentions are not well specified concrete entities; they are abstractions we use to make sense of behaviour. There are processes in the brain that are responsible for the evaluation that led to the action, but there is no well-specified internal description of what the participants intended to do in addition to that. Something must precede the action, but that process does not need to exist in a format that is comparable to the Folk Psychological description of what went on.

But in relation to our experiments, the problem for Intentional Realism becomes more vivid when we leave the high grounds of philosophical controversy and instead look at more specific cognitive science models of human behaviour. Even if not explicitly endorsed, Intentional Realism about Folk Psychological constructs is a ubiquitous feature in cognitive science. In line with the reasoning of Fodor, many researchers have taken the apparent success of Folk Psychology as evidence that there must be corresponding processes in the brain that closely resemble the goals and intentions postulated by the theory.

Letting the intentions out of the box

In cognitive psychology and cognitive science, a frequently used tool for describing cognitive and behavioural relations is the flowchart model. When it comes to goal-directed behaviour, one thing the models often have in common is that in the uppermost region of the chart, a big box sits perched governing the flow of action. It is the box containing the Prior Intentions (Brown & Pluck, 2000; Jeannerod, 2003). In these models, intentions and goals are discrete entities with very specific identifiable properties.

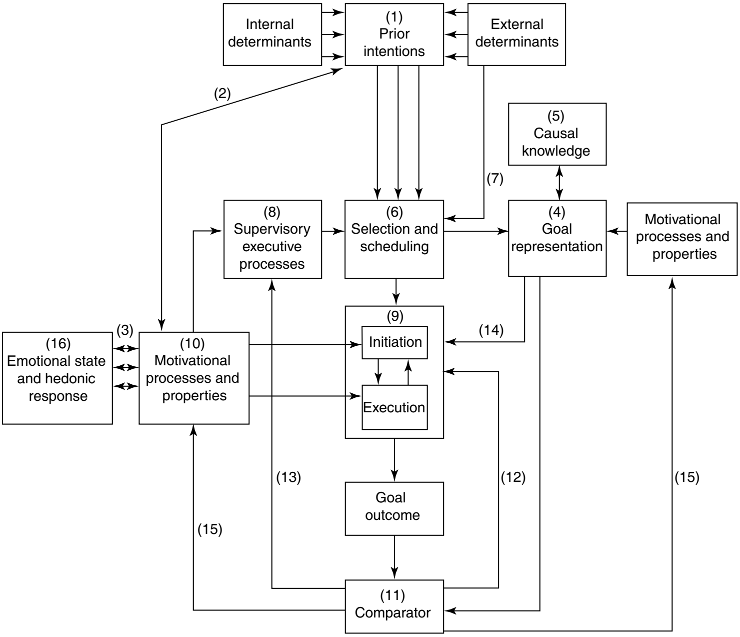

Internal determinants
(1) Prior intentions
External determinants
(2)
(5) Causal knowledge
(7)
(8) Supervisory executive processes
(6) Selection and scheduling
(4) Goal representation
Motivational processes and properties
(9)
(3)
(16) Emotional state and hedonic response
(10) Motivational processes and properties
Initiation
(14)
Execution
(13)
(12)
(15)
Goal outcome
(15)
(11) Comparator

Figure 1. Model of goal-directed behaviour, from Brown and Pluck (2000).

The model above is labelled a neuro-cognitive and a neuro-philosophical formulation of goal-directed behaviour (Brown & Pluck, 2000; Jeannerod, 2003). The model is in itself a synthesis of other models from several different areas in cognitive science, such as cognitive and functional anatomy of will and volition (Ingvar, 1999; Spence & Frith, 1999), neurobiology of reward (Schultz, 1999), and philosophical descriptions of purposeful behaviour (Searle, 1983).

According to this flow-chart, a goal-directed action is driven by the Prior Intention. For an action to be goal-directed the system needs to have an internal representation of the goal, as well as knowledge of particular actions that will lead to achieving the goal. The action is controlled through feedback from the Comparator, which compares and evaluates the goal outcome against the goal representation. The output from the Comparator is used to maintain or stop the ongoing action, and will further influence the motivational processes involved in the task.

It is in relation to a model like this that choice blindness as a phenomenon becomes very hard to account for. Arguably, choosing and taking the more attractive of two pictures of faces must be considered a goal-directed behaviour. The action performed in our experiment has all the components of the model, but still the mismatch between the intention and the outcome is not detected. The Comparator should have stopped the process when the participants received the opposite of their choice, but it didn't. How can this be?

One explanation could be that in models like this, the internal representation of the goal state only concerns low-level features; maybe they are only meant to describe actions on a motoric level, such as reaching for the remote or tying one's shoes. But Brown and Pluck (2000) do not put any restrictions on the kinds of actions or level of goal specificity that this model is supposed to handle:

> Within neuroscience, the construct of GDB [goal-directed behaviour] is increasingly being used to operationalize a broad spectrum of purposeful actions and their determinants, from the simplest single-joint movement, to the most complex patterns of behaviour. GDB is construed as a set of related processes by which an internal state is translated, through action, into the attainment of a goal. The 'goal' object can be immediate and physical, such as relieving thirst, or long-term and abstract, such as being successful in one's job or the pursuit of happiness. (p. 416)

Apparently, both intentions and goals can be both abstract and complex, and for the Comparator to fill any function it must be able to detect when the higher-order goals are obtained or not.

Another possible objection to protect the model is that the Comparator just did not do its job this time. Maybe checking the relation between intentions, goals and results is optional rather than essential? But the ability to compare the prior goals with the outcome obtained is an ever-present feature in action modelling, and is seen to be fundamental for a great number of things:

> Our ability to judge the consequences of our actions is central to rational decision making [...] A key component to survival in a constantly changing environment is the ability to evaluate the consequences of one's actions and to adapt one's behavior accordingly. (Walton, Devlin, & Rushworth, 2004; p. 1259)

> Flexible behavior requires a system for relating responses to the current context and one's goals. (Badre & Wagner, 2004; p. 473)

> Adaptive goal-directed behavior involves monitoring of ongoing actions and performance outcomes, and subsequent adjustments of behavior and learning. (Ridderinkhof, Ullsberger, Crone, & Nieuwenhuis, 2004; p. 443)

> [T]he anterior cingulate cortex (ACC) has a fundamental role in relating actions to their consequences, both positive reinforcement outcomes and errors, and in guiding decisions about which actions are worth making. (Rushworth, Walton, Kennerley, & Bannerman, 2004; p. 410)

> Flexible adjustments of behavior and reward-based association learning require the continuous assessment of ongoing actions and the outcomes of these actions. The ability to monitor and compare actual performance with internal goals and standards is critical for optimizing behavior. (Ridderinkhof, van den Wildenberg, Segalowitz & Carter, 2004; p. 135)

> Voluntary action implies a subjective experience of the decision and the intention to act [...] For willed action to be a functional behavior, the brain must have a mechanism for matching the consequences of the motor act against the prior intention. (Sirigu et al. 2004; p. 80)

So it does seem as if the Comparator plays a substantial role in many theories. Just looking at the quotations above, to be able to compare the goals with the outcome of one's actions is deemed of vital importance for as diverse things as rational decision making, learning and voluntary action. The Comparator should have been on full alert when the choice was executed.

To connect with the discussion of Fodor's version of Intentional Realism, a third alternative is of course that there is nothing in the box. Or, at least, whatever process fills the role of initiating the action or representing the desired goal-state, it does not correspond to what could be expected from a Folk Psychological perspective. The model still works if the only thing that is supposed to be represented is the motor action: I point, reach, and pick up the picture to the right. That is what I did, so I did the right thing. But the model is clearly meant to be more than this. If I reach for a beer but end up with a glass of milk in my hand, I should notice, because that was not what I wanted! In relation to standard models of goal-directed behaviour, I think choice blindness is a genuine problem that needs to be addressed.

A small caveat is called for here. I do not claim that it is impossible to consciously deliberate the reasons back and forth for a particular choice, and we certainly can remember (some) of the things we tell ourselves when doing so. And we can set up "goals" like quitting smoking and then notice when we fail to achieve them. But the things we *say* to ourselves when trying to quit smoking should not be the starting point when we try to build models for how our cognitive machinery represents the mechanisms for our actions. They are Folk Psychological constructions, given their exactness through the language we use, not by a reality they describe.

Introspection and verbal reports

Summary Paper 3: ***How something can be said about telling more than we can know.*** The experimental method in Paper 3 is identical to the one used in Paper 2. The participants were shown pairs of female faces and were asked to choose which one they found more attractive. After the choice had been performed, the participants were sometimes asked to explain their choice. Eighty participants completed 15 trials each, of which three were manipulated. The deliberation time for performing the choice was fixed to four seconds for all conditions. The set of faces was different from that used in Papers 1 and 2.

The important difference in relation to the study described in Paper 2 is the collection of introspective verbal reports. This study was divided into two different conditions. In the first condition the participants were simply asked why they preferred the chosen picture. The same question was asked in the second condition, but now the experimenter encouraged the participants to elaborate their answers up to one full minute of talking time. This was done both by the use of positive verbal and non-verbal signals and by interjecting simple follow-up questions.

Two major methods were used in the comparative analyses of the verbal reports: relative word frequency and latent semantic analyses. Based on relevant research, such as automatic lie-detection and

> language development, a large number of variables were compared for manipulated and non-manipulated reports. Examples are: filled and unfilled pauses, words marking uncertainty, specific and non-specific nouns, positive and negative adjectives, lexical density and diversity. Of the total 30 variables measured for long as well as short reports, only two variables were statistically different in manipulated and non-manipulated reports. In latent semantic analyses, by analysing the contextual usage of words in a large corpus (i.e. a collection of text), a "semantic space" is constructed representing the relative distance between the words in the corpus. This space can in turn be used to calculate the difference between two other corpora. In our analyses, we found no difference between manipulated and non-manipulated reports. In contrast, large discrepancies were found between our male and female participants, both with latent semantic analyses and with several of the linguistic frequency variables. The detection of sex differences shows that it is possible to detect differences in our corpus with the methods we have used, which thereby gives strength to the overall conclusion that there are very few differences between manipulated and non-manipulated reports.

No difference that makes a difference

To better appreciate the discussion of the theoretical context of this study, a few words on the underlying reason for examining the verbal reports. First of all, it is interesting that the participants *do* talk in the manipulated trials, that they say anything at all. As they are asked to explain a choice they did not make, saying "I don't know" or "I wanted the other one!" would seem the more natural thing to do.

Secondly, it is interesting to analyse what the participants actually say, to find out to what extent they give reasons referring to the original choice or the manipulated outcome. Due to the nature of the stimuli it is often quite hard to determine which of the faces has the "pretty nose" or the "nice haircut" the participants might claim to have been influential in their decision. But sometimes the features referred to are unique for the manipulated picture, such as the earrings, the dark hair or a hint of a smile. In these cases we can be certain that the reports are constructed after the fact, and thus in some sense are confabulatory.

Thirdly, it is interesting to compare the manipulated and the non-manipulated reports. The amount of difference detected says something about the "normality" of the manipulated reports. If the reports have the same amount of detail, the same number of pauses and markers of uncertainty, the same amount of emotional content, and so on, then there is nothing "wrong" with the reports generated in the manipulated trials. This also serves as an implicit marker whether the participants on some unconscious level have detected or registered the manipulation, as detection might have asserted itself, for example, in an increase of markers of uncertainty.

And finally, the lack of differentiation between manipulated and non-manipulated reports also says something about the "authenticity" of the non-manipulated reports. If there are no or few differences between manipulated and non-manipulated reports, and we know that the manipulated reports at least to some extent are confabulatory, then this might indicate that the same mechanism is responsible for both types of reports. In this roundabout way, it could be argued that the problems of finding differences between manipulated and non-manipulated reports are due to the fact that they are *both* confabulatory. No difference that makes a difference.

Know Thyself

Hardly any concept in the history of psychology and philosophy of mind has generated more controversy than introspection (Lyons 1986). Since Descartes' dualist vision of a mind fully transparent to the self, the pendulum regarding just how much we think we know about ourselves *from the inside* has swung back and forth several times. Early experimental approaches such as the German Gestalt psychology (e.g. Wertheimer, 1912) relied heavily on the ability to report accurately on one's perceptual experiences. This was in turn followed by Methodological Behaviourism (Watson, 1913; Skinner, 1938), in which behaviour is supposed to be explicable without reference to intermediate mental states, leaving little interest for what people claimed to know about the workings of their own minds. Despite not being necessarily committed to introspection, the cognitive movement that came to replace Behaviourism as foundational for psychological research at least put the mental back on the map. Still, prominent researchers such as Ericsson and Simon (1980; 1998) believe that by using techniques such as "think aloud" during problem solving, we get an accurate picture of what is actually going on when we make decisions and solve problems.

A parallel and not entirely coincidental development can be seen in philosophy. In the phenomenological tradition, Husserl developed the notion of *epoché*, which translates to an isolation of the inner experience from theories or preconceptions of how the world works. The subjective perspective is essential for understanding the mind, and the goal to strive for is the "purest" form of introspection (Husserl, 1900/1970). Wittgenstein questioned this very idea in his famous discussion of the private object (Wittgenstein, 1953). It is of course not entirely clear what Wittgenstein would recommend as psychological practice, but he is at least often interpreted as arguing against the possibility of isolating an experience and then saying something meaningful about it. In *Concept of Mind*, Ryle (1949) was a bit more straightforward in his attack on introspective knowledge:

> The sort of things that I can find out about myself are the same as the sort of things that I can find out about other people, and the methods of finding them out are pretty much the same. A residual difference in the supplies of the requisite data makes some difference in degree between what I can know about myself and what I can know about you, but these differences are not all in favor of self-knowledge. (p. 155)

For Ryle, mental talk was to be understood as dispositions to act, not as descriptions of causally active entities. Despite being out of favour nowadays, Ryle's Logical Behaviourism inspired many later thinkers, such as Sellars (1963) and Dennett (1987; 1991a).

In modern days, the debate over the use and utility of introspection has been seamlessly intertwined with the discussion of the "easy" and the "hard" problems of consciousness (Chalmers, 1996) that is, what can be known about consciousness from the first person perspective (introspection) compared to the third person perspective (the standard scientific method).

The partisanship is as fierce as ever concerning the philosophical problems of consciousness and introspection, with cemented positions and slight chances of resolution or reconciliation (Block, 1995; Chalmers, 1996; Dennett, 1993; Rorty, 1993).

In cognitive science, something like a consensus has emerged around a picture of the mind as primarily being made up out of unconscious machinery (e.g. see Gazzaniga, 2004). It is clear that large parts of what is going on in the brain do not ever reveal themselves to introspection (Dehaene & Naccache, 2001; Wilson, 2002; LeDoux, 1996). But there is also a steadily growing appreciation for the central role introspective reports can play in, for example, cognitive neuroscience research, triangulating the reports with behaviour and brain activity (Jack & Shallice, 2001; Jack & Roepstorff, 2002).

There are many forms and aspects of introspection, as there are many different things we can know about ourselves, our experiences and our mental states (Schwitzgebel, 2002). To lump all threads together in one quick historical sweep does not do justice to the intricacies of all positions held and argued for. For example, in relation to phenomenal states or qualia (things like seeing red or the softness of a kiss), I cannot claim that our experiments have much to say. Regarding self-knowledge and introspection as such, I am primarily concerned with higher-order mental states such as beliefs and desires. And in relation to this, what introspection can tell us about what we believe and what we desire, our experimental results clearly support an anti-introspectionist view. If we are supposed to know our own minds from the inside, we should know why we do what we do. And when asked to describe why we chose a face we in reality did not prefer, we are not supposed to just fabricate reasons (at least not without knowing that this is what we are doing). In our experiments, it is evident that the participants do not have perfect access to their underlying cognitive machinery. But despite being a striking demonstration that we don't always know why we do things, the results of our experiments do not have as great an impact on philosophy of mind as they might have had some decades back. Few philosophers today believe us to be infallible concerning our own mental processes. However, in relation to the previously mentioned debate about how we use Folk Psychology, introspective knowledge *is* essential for philosophers such as Goldman (1993), as we are supposed to understand the behaviour of other people through an internal simulation of what we would have believed and desired had we been in their shoes. If we use our own mind as a model to understand others, it is a bit curious that we have such a lack of understanding of how we function ourselves.

Regardless of what we actually *do* know about our own mental life, one interesting aspect of self-knowledge is that for most people it does *feel* as if we know ourselves from the inside.

As in our case, when I tell you why I made a particular choice, I just assume that I am right. Where this sense of knowing comes from is of course contested (i.e. does it feel right because in general we *are* right? Goldman, 1993; Gopnik, 1993), but most people debating introspection agree that this is a prevalent part of the psychological sphere. One reason why it feels as if we have this special authority about ourselves is that we are very seldom proven wrong. However strongly I suspect that "being sorry" does not accurately describe your present condition, when you tell me that this is how you feel, there is no external evidence for me to use against your claim. But this is true in relation to ourselves as well, i.e. we rarely realise that we are wrong in our self-explanations. As Nisbett and Wilson (1977; p. 256) say: "disconfirmation of hypotheses about the workings of our own minds is hard to come by." This is also a genuine problem when doing experimental work on self-knowledge. Without any means to question the validity of people's verbal reports, it is also difficult to say how much of it is true. Most often, the correctness of people's introspective reports is just taken for granted.

We have solved this problem in our experiment. We do not need to take on the burden of explaining the mechanism behind the original choice – why they preferred one face over the other in the first place. Given the structure of the manipulation, we just *know* that the participants did not want what they got. By setting up this mismatch between what they wanted and what they received, we now have a way of demonstrating when experimental participants are manifestly wrong about themselves. And as such it is a novel tool in research on self-knowledge. And in addition, it is also a way to show both to ourselves and to others that we do not know as much about ourselves as we think we do.

As was the case in the previous discussion of choice blindness and Folk Psychology, the implications of our approach are perhaps better seen when we connect it to a more specific research tradition in cognitive science.

How something can be said

Paper 3 takes as its starting point the classic article “Telling More Than We Can Know: Verbal Reports on Mental Processes” by Nisbett and Wilson (1977). It is one of the most cited articles of all times in psychology as well as philosophy, and it surfaces in the most diverse circumstances. But what did they actually say to stir such a controversy?

At the outset, Nisbett and Wilson make clear that they are interested in mundane verbal interactions, such as giving and taking reasons, asking questions, making judgements, stating preferences, etc. In our daily lives, we are confronted with countless questions that rely upon our higher-order cognitive processes: “Why do you like him?” “How did you solve this problem?” “Why did you take that job?” (1977, p. 232). We answer such questions with apparent ease, and we ask them ourselves believing that others can tell why they do what they do. Nisbett and Wilson thought this confidence ill-founded. They had collected a lot of relevant research from neighbouring fields, as well as performing a large number of experiments themselves. Their own (rather harsh) verdict:

> [T]here may be little or no direct introspective access to higher order cognitive processes. […] when people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. (Nisbett & Wilson, 1977; p. 232)

They had reviewed large parts of the then burgeoning experimental social psychology literature, with topics such as cognitive dissonance, insufficient justification, and attribution theory, and found a lot of support for their conclusion.

An example of the kind of studies they leaned on is Zimbardo’s famous grasshopper experiment (Zimbardo et al, 1969). This study is also a nice illustration of the insufficient justification effect, as well as a telling example what was allowed before the reign of ethics committees.

The group of participants consisted of students recruited to an outdoor survival training course. Naturally, to survive outdoors, an essential skill is learning to eat what nature has to offer. On this topic, how to best capture, prepare and eat grasshoppers was explained to the participants. Half of them were instructed by a nice and warm person, sensitive to their discomforts, interacting in a friendly manner with his assistants, etc. The other half were given an angry and hostile instructor, yelling at his co-workers, laughing at the participants, and so on. After the "eating" was done, the participants had to indicate what they actually thought of the experience. In line with insufficient justification theory, the group with a non-pleasant instructor liked the taste better than the other group (a few even took extra grasshoppers home to share with their friends and families). The logic of insufficient justification theory is sometimes a bit hard to follow, but to explain using the terms of the theory: In the first group, the "dissonance" between disliking grasshoppers and still eating them could be reduced by "thinking" that they did it because the instructor was such a nice man, and as the dissonance was accounted for by referring to the instructor, the participants did not need to change their negative attitudes towards eating grasshoppers. But in the second case, the participants could not find a sufficient justification for why they ate those disgusting grasshoppers, so they changed their attitude towards liking them instead. It is the same argument as in experiments in which you like a boring task more if you get paid less; as it can not have been the money that made you do it, you must just have liked it!

But what is important here is that the participants themselves are not aware that their attitude has been influenced by the behaviour of the experimenter. If asked why they would not have known that the perceived likeability of the instructor was the reason they now (believed themselves) to like eating grasshoppers.

Among Nisbett and Wilson's own experiments, the most pertinent to our experiments is the stocking and nightgown study. Under the pretence of a consumer survey, people walking by in a shopping centre were invited to evaluate articles of clothing. The participants were either asked to indicate which one of four different nightgowns they preferred, or to evaluate four identical pairs of nylon stockings. When they had made their choice, they were asked why they had chosen the article in question. As reported by Nisbett and Wilson: “there was a pronounced left-to-right position effect, such that the right-most object in the array was heavily overchosen. For the stockings, the effect was quite large, with the right-most stocking being preferred over the leftmost by a factor of almost four to one” (1977, p. 243). In contrast to this, none of the participants mentioned position as having a possible influence on their choice; not surprisingly, they commented on the quality or texture of the fabric instead. Nisbett and Wilson themselves were not able to provide a systematic explanation of why position should be such an important factor. Their suggestion was that people might examine the items from left to right and hold of judgement until the last one in the array had been explored. But what is important here is not really *how* the ordering influenced the evaluation, the interesting part is that we know that it had an effect but still did not show up in the participants’ own explanations.

The stocking and nightgown study nicely captures the spirit of the Nisbett and Wilson approach, showing that we sometimes are unaware of which stimulus influences our behaviour. It is also relevant because it bears a structural resemblance to our studies: several items are evaluated, one of them is publicly chosen as the one preferred, and the choice is later explained to the experimenter. But there are also some important differences. Naturally, I consider our choice blindness experiments to represent a methodological step forward. By listing some of the arguments directed against the studies of Nisbett and Wilson, we can see to what extent that is true.

Ecological validity. Nisbett and Wilson have been accused of using unimportant and contrived tasks in their experiments: It is somewhat strange to choose the one preferred of *identical* stockings at a clothing retailer (Kraut & Lewis, 1982; Kellogg, 1982). It is not unreasonable to believe that our introspective capacities may be diminished under such circumstances (Smith & Miller 1978). In contrast, choosing which face one finds more attractive is a very straightforward task, reflecting a simple type of judgement that people often make in their daily lives. While not being the most important task imaginable, many people have very strong opinions about facial attractiveness. Compared to the studies of Nisbett and Wilson (and to psychological experiments in general), evaluating faces is as interesting as it gets.

Verbal reports. Despite the title of their article, very little was done with the verbal reports in Nisbett and Wilson (1977). Apart from registering whether the influential stimuli were mentioned or not, no thorough or comparative analyses were performed. In most of the experiments the introspective reports were also generated several minutes (or even hours) after the critical behaviour occurred. Several critics therefore argued that the impoverished and "incorrect" verbal reports were due to a memory effect (Ericsson & Simon, 1980). The participants had simply forgotten why they did what they did. Ericsson and Simon (1980; 1993) put this in contrast to their own protocol analyses and "think-aloud" technique, in which the participants "reveal" their actual trains of thought by verbally stating what they think while performing a task. If done properly, with the correct timing, this is supposed to yield a "correct" description of our cognitive processes:

> [T]he validity of verbally reported thought sequences depends on the time interval between the occurrence of a thought and its verbal report, where the highest validity is observed for concurrent, think aloud verbalizations. For tasks with relatively short response latencies (less than 5–10 seconds), subjects are able to recall their sequences of thoughts accurately immediately after the completion of the task and the validity of this type of retrospective reports remains very high. (Ericsson, 2002; p. 3)

In our experiments, the reports were solicited only a few seconds after the choice was made, immediately after the participants had received the chosen picture. According to the quotation above, this is well within the time margin Ericsson has set up for delivering accurate descriptions of our cognitive processes. What the participants say in our experiment should be a true reflection of why they chose one picture over the other. In a way Nisbett and Wilson's studies did not, our results seem to challenge this position.

It should also be noted that in our experiments the participants had been informed at the beginning of the sessions that we would ask them about their reasoning, thus cueing them to reason deliberately, and to attend to their reflective processes.

Individual vs. group effects. In most of the experiments presented in Nisbett and Wilson (1977), the discrepancies between action and introspection can only be discerned in group-level response patterns, not for each individual (Quatrone, 1985; Quatrone & Jones, 1980, Smith & Miller, 1978; White, 1988). In the stocking and nightgown experiment above, it is impossible to say which of the participants were influenced by the positioning of the items, we only know that some of them must have been influenced as we know that from a statistical perspective there is an ordering effect. In our experiments, we *know* that the participants did not want the photograph received in the manipulated trials. Whatever the participants say, it will be in contrast to what they originally intended to choose. This design also gives us the two classes of verbal reports to compare and contrast. And at the very minimum, in the manipulated reports describing unique features of the non-chosen picture, we have unequivocally shown that normal participants may produce confabulatory reports when asked to describe the reasons behind their choices. This too goes beyond what was established by Nisbett and Wilson.

I think we are allowed to say that our experiment is a methodological improvement on what was employed by Nisbett and Wilson. We solve several of the problems they were criticised for, as well as providing a methodological platform for new experiments. Our experimental design is the first to give cognitive scientists the opportunity to systematically study how confabulatory reports are created and how they relate to standard or "truthful" reports about choice behaviour. In the end, this will hopefully enable us to also say something about the *general properties* of introspective reports.

Methods and Methodology

Summary Paper 4: ***Magic at the marketplace.*** The experiment took place inside a local supermarket, and the participants were recruited after being asked if they wanted to participate in a consumer preference test. The test consisted of tasting or smelling two sorts of jam and two sorts of tea. When the participants had made their choice of which jam or tea they preferred they got to sample the chosen item again, and were asked to explain why they liked this one better. For each participant, either the tea or the jam condition was manipulated. By using prepared jars with two separate compartments containing both varieties of jam or tea, the experimenter could switch the position of the two jams or teas by simply turning both jars upside down (see Figure 1 in Paper 4). When the participants sampled the third time they were given of the non-chosen product, and at the same time they were asked why they liked this taste or smell better,

In total, 180 participants took part in this experiment. The similarity within the pairs was established in a pilot study. The six pairs used in the experiment ranged from relatively similar to distinctively dissimilar. A trial was categorised as detected if the participants voiced any concerns immediately after tasting or smelling the switched jam or tea or if the participants at the end of the experiment in any way claimed to have noticed the manipulation. A manipulated trial was also considered detected if the participant thought that the taste or smell had changed the second time it was sampled.

Half the participants also received either a package of tea or a jar of jam as a gift. The jam or tea chosen by the participants in the manipulated trials was also the product used as gift. In addition, several other factors were measured in the experiment. When sampling the first time, the participants rated both sorts of jam and both sorts of tea with regard to how good they tasted or smelled. After the choice, the participants rated how easy it was to discriminate between the two choice options, and also indicated how confident they were in their choice.

Counting all conditions and all forms of detection, 32.2% of the manipulated tea trials and 33.3% of the manipulated jam trials were detected. There was an increased rate of detection for the least similar compared to the most similar pair for both tea and jam. The gift was associated with lower detection rate for tea but not for jam. A larger discrepancy in attractiveness rating was associated with higher degree of detection for jam but not for tea. Comparing manipulated and non-manipulated trials, the perceived ease of distinguishing between the items in the pairs was higher for non-manipulated trials for tea but no difference was found for jam. There were no differences in rated confidence between the manipulated and the non-manipulated trials for either tea or jam.

> The major conclusion drawn is that choice blindness is further established as a robust effect in decision making, extending the findings from previous research using visual stimuli to the modalities of taste and olfaction.

The Wedge

At the beginning of the introduction, I identified three things as novel in this thesis: Choice blindness, the verbal reports and the experimental methodology as such. The first two entries on this list have been discussed in relation to Papers 1, 2 and 3. Accordingly, Paper 4 will be primarily used as a platform for a discussion of the experimental methodology. I will give some background for why and how we came up with the idea of doing the kind of studies described in the thesis, and also present some planned future work on choice blindness.

From a methodological perspective, it is important to point out that the experimental approach was deduced from our theoretical background rather than the other way around, i.e. we did not invent the experiments first and then try to find a suitable context for them. Being very much influenced by Daniel Dennett, my colleague Lars Hall and I had for a long time thought that there must be some experimentally testable consequences of his Intentional Stance theory. We had previously made a distinction between (the classical concept of) introspection and a more Dennettian mode of self-knowledge based on self-observation, which we called *extrospection* (Hall 2003, Hall & Johansson 2003a). To emphasize the potential of extrospection as a tool for self-understanding, we had applied this concept in the domains of educational psychology and self-control (Hall & Johansson 2003b, Hall, de León & Johansson, 2002), but thus far we had not made a direct empirical test of the theory.

Given this perspective it ought to be possible to influence people's interpretations of themselves by controlling what evidence they have available for their extrospective reasoning. As Dennett claims in the long quotation I used previously, every one of us is an: "inveterate auto-psychologist, *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing" (Dennett, 1987; p. 91, emphasis in original). As we see it, choice blindness can be used as a wedge to pry apart the otherwise "inseparable mix" of the things we do and the things we say about ourselves.

An interesting further application of this methodology is to examine what happens *after* the choice (what Dennett 1991a calls The Hard Question: *And then what happens?*). In Paper 1, a memory test used after the completion of the choice experiment revealed that the participants tended to remember the manipulated outcome as being what they originally preferred. But the more interesting question is what becomes of the participants preferences and attitudes; what would for instance happen if they had to do the same choice again, would they pick the alternative they initially thought was better or the mismatched option they ended up with?

We have recently begun to explore this question. In the experiment that formed the basis for the introspective reports that were analysed in Paper 3, the participants had to choose between two faces, pick the one they preferred, and give either a short or a long verbal report explaining their choice. But in addition to this, their later preferences were also probed in several different ways. All participants were presented with the pairs a second time and had to choose the picture preferred once again. In one condition, the participants also had to rate on a numerical scale how attractive they thought both pictures were directly after having given their verbal reports. The results showed that the participants were clearly influenced by the manipulations made, as they were much more likely to pick the originally non-preferred face the second time they had to evaluate a pair. But perhaps even more interestingly, this tendency was correlated with the participants "involvement" in the choice, i.e. if they had given short or long reports, and if they had numerically rated the pictures after the first choice (see Hall, Johansson, Tärning & Sikström, in preparation). [4]

We think this is a very interesting avenue of exploration. What will happen with these "induced" preferences over time? Will they

[4] This paper was meant to be included in the thesis, but life could no longer wait.

transfer to more general attributes (like preferring brunettes)? Will they be modulated by other choices? In a sense, choice blindness can be used as an instrument to measure how much we influence *ourselves* by the choices we make.

A brief note on magic

The experimental procedure in Paper 2 was developed in cooperation with the eminent Swedish close-up magician Peter Rosengren. The technique used is called "black art" (Dondrake, 2003), which is a method of concealing something black against a black background (e.g. the ropes carrying the attractive assistant when she appears to float in mid-air on stage). In the manipulated trials, the experimenter held two cards in each hand, with the card shown fitted with a black back side of the same material as the black desk cover that served as the surface of the experiment. When the "chosen" picture was slid to the participant, the front card stayed on the table. Generally, black art can be used effectively even at a very close range, but since we needed to conduct our experiment in a brightly lit office environment we also used some sleight of hand, through which the extra card is hidden by the experimenter's sleeve until it is raked back and falls down in a hidden compartment at the end of the table (see Picture 1 in Paper 2).[5]

The technique used in Paper 4 has its origin in a long discussion we had with two professional magicians at the yearly "Swedish Magicians' Circle" conference, Karl Berseus and Axel Adlercreutz. But it was Lars Hall who came up with the brilliant idea of gluing two jars together and thereby creating a single jar with two separate compartments. In this experiment, we also used two experimenters working together to conceal the manipulation, as the first experimenter waits to execute the switch until the participant moves his or her attention to the other experimenter to answer a question about how well they liked the sampled item.

Interestingly, while the techniques of the experiment are imported from the domain of magic, the purpose of the experiments is more or less the opposite of what magicians usually want to

[5] In the experiment in Paper 2, only two participants were removed for having seen the procedure – as they would say in the classic poker movie *Rounders*: I only got caught with a hanger twice!

achieve. In card magic, the performer must take great pains to ensure that the participants and the members of the audience are able to remember which card was initially chosen. Otherwise, when the act reaches its finale, they would simply be unable to notice that anything magical had taken place. But in our experiments the whole point is the participants not noticing the change; in this case, we have to wait for the applause until we are published! Despite this, it is safe to say that it has been a lot more fun to invent and perform the experiments than to analyse the data obtained.

The future of choice blindness

As we see it, there are a great number of possible variations and extensions that can be made in relation to the experiments we have produced so far. In both Paper 1 and Paper 4, we briefly discuss the possibilities of using the methodology of choice blindness as a more general tool in psychological research. Here, I would just like to give a short overview of some of the things we have started on or plan to do in the near future.

We do not yet know the limits of choice blindness. For instance, while it seems as if it would be impossible to swap two pictures of Marilyn Monroe and Marilyn Manson without the participants noticing, it is still an empirical question how dissimilar or how "unequal" two pictures can be. We also need to investigate more rigorously the importance of parameters related to the memory of the choice, such as the encoding time (i.e. the time participants are allowed to deliberate upon their choice), the occlusion interval (i.e. the time the chosen stimuli is invisible when the manipulation is performed), and the retention interval (i.e. the time until the mismatch detection is tested).

But we can also change the stimuli as well as the task to be performed by the participants. Both abstract patterns and male and female faces have been tested, but perhaps change blindness would disappear if other stimuli were used (as someone remarked in an Internet chat-forum after the *Science* publication: "Who cares about pictures of young women – had it been pictures of new cars there is no way I would have missed the switch!"). We could also use more "culturally" charged stimuli, such as brands or logotypes (fake or real), and ask question in line with standard marketing research, for instance, which symbol is more energetic, youthful, dynamic etc. If we keep faces as stimuli but instead change the task, we could, for example, vary the importance of the choice, such as letting the participants choose which of two persons they were going to have a cup of coffee with, or which one they would prefer to employ at their company.

Large parts of the research done on face processing have been on aspects relevant from an evolutionary perspective, and much of this research is easily adapted to our approach (Penton-Voak, & Perrett, 2000; Perrett et al., 1999). For example, we could systematically vary the symmetry of the faces, or change the task to things like which person would you rather have a long-term relationship with as compared to a one-night stand. It could be suspected that changes made on more evolutionarily important choices should also be more easily detected, but again, this is an empirical question.

To expand on the issue of verbal reports and confabulation, instead of a complete identity switch, we could just add potentially salient features, such as earrings or a smile, and see if any of these features were mentioned in the participants' explanations of their choice. If they were, this would add even more strength to the suspicion that reasons stated for choices are often constructed "after the fact". But there are no grounds for *not* including verbal reports in all or most of the experiments, and thereby building a large "database" of various forms of manipulated and non-manipulated reports.

One large class of data that we have yet to work with is implicit measures, such as galvanic skin response, eye-tracking, ERP and fMRI. This type of measures is interesting for several different reasons. First of all, they might reveal specific response patterns that differentiate between manipulated and non-manipulated trials, indicating that, despite the participants' own conscious denial of having detect a manipulation, some parts of the cognitive system actually "noticed" that something went wrong with the choice. There is a large literature on change blindness and change detection in general that is connected to this issue (see Simons & Silverman, 2004). Secondly, there might be patterns in, for example, the saccadic movement of the eyes that are indicative of whether a change *is going to be* detected. Perhaps the detected manipulations are encoded differently? Thirdly, there might be ways to connect the verbal reports to, for example, patterns revealed by ERP. Are there any differences in activity between giving confabulatory and "ordinary" verbal reports?

By keeping the methodology and just varying the stimuli and the task, a large number of interesting experiments could be made. But we could also expand on the method, using new "magic" tricks, such as the prepared jars in Paper 4. With methods like this we could try changing real objects rather than just pictures, as well as further exploring choices in other sensory modalities than vision.

As suggested by the inclusion of implicit measures, we can also focus on other aspects of the participants' responses. One interesting (and underdeveloped) feature in Paper 4 is the certainty measure – i.e. the participants' own rating of how certain or confident they felt in their choice. We found no differences between manipulated and non-manipulated trials, which means that the participants were just as confident in a choice they did not intend to make as in one they did make without alterations. The use of self-rating scales of certainty is a prevalent component in psychological research on decision making (Baranski & Petrusic, 1998; Petrusic & Baranski, 2003; Pallier et al., 2002). The fact that it is possible to switch the outcome of people's choices without this making a mark on how confident they are in those choices ought to say something about the precision of this type of self-rating measures.

Similarly, the study in Paper 4 can be used as a starting point in a more thorough investigation of decision making and consumer behaviour. In what circumstances are we blind to changes in our consumer choices? How does a non-detected manipulation affect, for instance, how much we are willing to pay for a certain item, or how satisfied we are with a certain product after we have bought and used it? There are of course many other ways of working with choice blindness to illuminate previous research on choices and decision-making, as well as the use of "introspective" verbal reports in psychological research.

Another approach would be to further enquire into the participants' self-understanding in our experiments. What do they themselves think they do when they answer the question *why* they performed a choice; do they think they have access to their own psychological processes, or do they think they just report the most likely causes when looking at the picture the second time? How certain are they that what they say actually captures the reasoning process responsible for their decision? What would happen if we instead asked *how* they came to that conclusion – would they attempt a more causal account compared to a *why*-question, or would they say that they just don't know? The terms introspection and confabulation have a very special meaning in philosophical jargon, but what does it correspond to when laypersons try to describe themselves and the actions they perform?

Despite being a both brief and shallow run-through of some of the things on our to-do list in the next few years, I hope it has served the purpose of showing that choice blindness as a concept extends further than the four studies presented in the thesis.

The End of the Beginning

There are of course many more things I would like to say in relation to my thesis. But it is time to stop here and let the papers talk for themselves.

As a final note, I would like to point out that even if this is my thesis, the work behind it is very much a collaborative effort. Lars Hall and I have worked on this project for a very long time, and during the last few years our duo has turned into a full group. Therefore, I would like to share the credit with all those listed as co-authors on the papers, but take the blame myself for all faults to be found herein.

References

Astington, J. W. (1993). *The child's discovery of the mind.* Cambridge, MA: Harvard University Press.

Badre, D., & Wagner, A. (2004). Selection, integration, and conflict monitoring: Assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron, 41*(3), 473–487.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology Human Perception and Performance*, *24*(3), 929–45.

Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Current Psychology of Cognition, 13*(5), 513–552.

Baron-Cohen, S. (1995). *Mindblindness: an Essay on Autism and Theory of Mind*. Cambridge, MA: MIT-Press.

Block, N. (1995). On a Confusion of a Function of Consciousness. *Behavioral and Brain Sciences, 18*(2), 227–287.

Bogdan, R. J. (1991). (Ed.). *Mind and Commonsense Psychology*, Cambridge: Cambridge University Press.

Brentano, F. (1874/1973). *Psychology from an Empirical Standpoint*. New York: Humanities Press.

Brown, R. G., & Pluck, G. (2000). Negative symptoms: the 'pathology' of motivation and goal-directed behaviour. *Trends in Neurosciences, 23*(9), 412–417.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chisholm, R. (1966). *The Theory of Knowledge*. Englewood Cliffs, NJ: Prentice Hall.

Christensen, S., & Turner, D. (1993). (Eds.). *Folk Psychology and the Philosophy of Mind.* New York: Lawrence Erlbaum Associates.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*, 67–90.

Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*, 1–37.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991a). *Consciousness explained*. Boston: Little, Brown & Company.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, *89*, 27–51.

Dennett, D. C. (1993). Back from the Drawing Board. In B. Dahlbom (Ed.). *Dennett and his Critics: Demystifying Mind*. Oxford: Blackwell.

Drake, D. (2003). Dondrake's Black Art Breakthroughs. Dondrake.

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review, 87*(3), 215–251.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

Ericsson, K. A. (2002). Protocol analysis and verbal reports on thinking. Retrived 15 September 2006, from: .http://www.psy.fsu.edu/faculty/ericsson/ericsson.proto.thnk

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Gazzaniga, M. S. (2004). (Ed.). *The New Cognitive Neurosciences III: Third Edition*. Bradford Books.

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, *16*, 15–28.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*, 1–14.

Greenwood, J. D. (1991) (Ed.). *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.

Hall, L. (2003). *Self-Knowledge/Self-Regulation/Self-Control: A Ubiquitous Computing Perspective*. Ph.D. Thesis. Lund University Cognitive Studies, 113.

Hall, L., & Johansson, P. (2003a). Introspection and Extrospection: Some Notes on the Contextual Nature of Self-Knowledge. *Lund University Cognitive Studies, 107*.

Hall, L., & Johansson, P. (2003b). Self-Regulation in Education: A Ubiquitous Computing Perspective. *Lund University Cognitive Studies*, *111*.

Hall, L., de Léon, D., & Johansson, P. (2002). The Future of Self-Control: Distributed Motivation and Computer-Mediated Extrospection. *Lund University Cognitive Studies, 95*.

Hall, L., Johansson, P., Tärning, B., & Lind, A. (in press). How Something Can Be Said About Telling More Than We Can Know: Reply to Moore and Haggard. *Consciousness and Cognition.*

Hall, L., Johansson, P., Tärning, B., & Sikström, S. (in preparation). Choice Blindness and Preference Change. *Lund University Cognitive Science.*

Husserl, E. G. A. (1900/1970). *Logical Investigations*. Vol 1-2.

Ingvar, D. (1999). On volition: a neurophysiologically oriented essay. In B. Libet, A. Freeman, and K. Sutherland (Eds.). *The Volitional Brain*. Exeter: Imprint Academic.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus-response to script-report. *Trends in Cognitive Sciences*, 6, 333–339.

Jack, A. I., & Shallice, T. (2001). Introspective physicalism as an approach to the science of consciousness. *Cognition*, *79*, 161–196.

Jeannerod, M. (2003). Consciousness of action and self-consciousness. A cognitive neuroscience approach. In J. Roessler, and N. Eilan (Eds.). *Agency and self awareness: Issues in philosophy and psychology*. Oxford: Oxford University Press.

Kellogg, R. T. (1982). When Can We Introspect Accurately About Mental Processes. *Memory & Cognition, 10*(2), 141–144.

Kraut, R. E. and Lewis, S. H. (1982). Person Perception and Self-Awareness - Knowledge of Influences on Ones Own Judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

LeDoux, J. E. (1996). *The emotional brain*. New York: Simon and Schuster.

Lyons, W. (1986). *The disappearance of introspection*. Cambridge, MA: MIT Press.

Moore, J. W., & Haggard, P. (in press). Commentary on "How Something Can Be Said About Telling More Than We Can Know: On Choice Blindness and Introspection". *Consciousness and Cognition.*

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Pallier, G, et al. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, *129*(3), 257–99.

Penton-Voak, I. S. and Perrett, D. I. (2000.) Female preference for faces change cyclically: Further evidence. *Evolution and Human Behaviour, 21*, 39-48.

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., and Edwards, R. (1999.) Symmetry and Human Facial Attractiveness. *Evolution and Human Behaviour, 20*, 295-307.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychon Bull Rev, 10*(1), 177–83.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Quattrone, G. A. & Jones, E. E. (1980). Perception of Variability within in-Groups and out-Groups - Implications for the Law of Small Numbers. *Journal of Personality and Social Psychology, 38*(1), 141–152.

Quattrone, G. A. (1985). On the Congruity between Internal States and Action. *Psychological Bulletin, 98*(1), 3–40.

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Ridderinkhof, K. R., Ullsberger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The Role of the Medial Frontal Cortex in Cognitive Control. *Science*, *306*, 443–447.

Ridderinkhof, K. R., van den Wildenberg, W. P. M., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition*, *56,* 129–140.

Rorty, R. (1993). Holism, Intrinsicality, and the Ambition of Transcendence. In B. Dahlbom (Ed.), *Dennett and his Critics: Demystifying Mind*. Oxford: Blackwell.

Rushworth, M. F., Walton, M. E., Kennerley S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Science*, *8*(9), 410–417.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Scholl, B. J., Simons, D. J., & Levin, D. T. (2004). 'Change blindness blindness': An implicit measure of a metacognitive error. In D. T. Levin (Ed.), *Visual metacognition in adults and children: Thinking about seeing*. Westport, CT: Greenwood/Praeger.

Schultz, W. (2000). Multiple reward systems in the brain. *Nature Reviews: Neuroscience, 1*, 199–207.

Schwitzgebel, E. (2002). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies, 9*(5–6), 35–53.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Sellars, W. (1963). *Science, Perception and Reality*. London: Routledge and Kegan Paul.

Simons, D. J., & Silverman, M. (2004). Neural and behavioral measures of change detection. In L. M. Chalupa and J. S. Werner (Eds.). *The Visual Neurosciences*. Cambridge, MA: MIT Press. 1524–1537.

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., & Haggard, P. (2004). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience*, *7*, 80–84.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Smith, E. R., & Miller, F. D. (1978). Limits on Perception of Cognitive-Processes - Reply to Nisbett and Wilson. *Psychological Review, 85*(4), 355–362.

Spence, S. A., & Frith, C. D. (1999). Towards a functional anatomy of volition. *Journal of Conciousness Studies*. *6*, 11–29.

Walton, M. E., Devlin, J. T., & Rushworth. M. F. S. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, *7*(11), 1259–1265.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*, 158–177.

Wertheimer, M. (1912). Experimental studies of the perception of movement. *Zeitschrift für Psychologie*, *61,* 161–265.

White, P. A. (1988). Knowing More About What We Can Tell - Introspective Access and Causal Report Accuracy 10 Years Later. *British Journal of Psychology, 79*, 13–45.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.

Zimbardo, P., Weisenberg, M., Firestone, I., & Levy, B. (1969). Changing Appetites For Eating Fried Grasshoppers with Cognitive Dissonance. In P. Zimbardo (Ed). *The Cognitive Control of Motivation: The Consequences of Choice and Dissonance*. Glenview, IL: Scott Foresman, 44–54.

PAPER ONE

From Change Blindness to Choice Blindness

Petter Johansson, Lars Hall & Sverker Sikström

Abstract: The phenomenon of change blindness has received a lot of attention during the last decade, but very few experiments have examined the effects of the subjective importance of the visual stimuli under study. We have addressed this question in a series of experiments by introducing choice as a critical variable in change detection. Participants were asked to choose which of two pictures they found more attractive. For stimuli we used both pairs of abstract patterns and female faces. Sometimes the pictures were switched during to choice procedure, leading to a reversal of the initial choice of the participants. Surprisingly, the subjects seldom noticed the switch, and in a post-test memory task, they also often remembered the manipulated choice as being their own. These findings indicate that we are prone to miss changes in the world even if they have later consequences for our own actions. In analogy with change blindness, we call this phenomenon choice blindness.

Introduction

Even if naïve participants often express bewilderment and disbelief during change blindness experiments, the *results* of these experiments no longer surprise cognitive scientists working in the field. In the last decade, a mass of empirical studies of change blindness have been published in the journals of cognitive science and vision research (Angelone, Levin & Simons, 2003; Simons, 1996; Simons & Chabris, 1999). The general phenomenon has been divided into various sub-fields, with respect to both the theoretical approaches (Mitroff, Simons & Franconeri, 2002; Rensink, 2002; Simons & Levin, 1997), and the techniques used (Grimes, 1996; O'Regan, Rensink & Clark, 1999; Smilek, Eastwood &

Merikle, 2000). Change blindness is now a standard example in cognitive science and vision research, on a par with the Stroop effect and the Kaniza triangle.

The common denominator in experiments on change blindness is that the participants fail to detect changes in a scene when the change is accompanied by some other visual disturbance. If the same changes had occurred in plain sight, with no interruptions in the visual stream, they would have been detected instantaneously. While the exact mechanisms have not yet been agreed upon (Simons, 2000), experiments involving change blindness have deepened our understanding of the visual system, particularly in mapping out the fine-grained properties of attention (Rensink, 2000; Tse, Sheinberg & Logothetis, 2003).

More controversially, change blindness has also served as a focal point in the debate about the nature of visual consciousness (the so called *Grand Illusion Debate*, see Noë, 2002), where the radical proposal have been made that change blindness shows that we all have a drastically false conception of our own visual experiences (e.g. Blackmore, 2002). A less dramatic conclusion drawn from these experiments is that we represent the world in much less detail than what was previously thought. Instead, when we need to be informed, we just direct our attention toward those features of the visual environment that is of current importance. As O'Regan and Noë (2002) says, we "allow the world to be its own best model". Thus, in this process, we rely on the stability of the world, and we implicitly assume that it does not change in undetectable ways.

However, surprisingly little research have been aimed at investigating our ability to detect changes when the stability of the world is of great importance to us – i.e. when changes in the visual environment have effects in relation to our *intentions* and *actions*. As Rensink (2002) writes:

> [T]he study of change detection has evolved over many years, proceeding through phases that have emphasized different types of stimuli and different types of tasks. *All studies, however, rely on the same basic design*. An observer is initially shown a stimulus… a change of some kind is made to this stimulus… and the response of the observer is then measured (p. 251, our emphasis)

We believe that the full potential of change blindness as a tool for studying the human mind is far from realized. Why should change blindness be used only to study *visual* aspects of cognition? In this paper, we are interested in the possibility of modifying the standard design of change blindness experiments. Our approach involves embedding change manipulations in a simple decision task where the participants are to choose which one of two items they find more attractive. The question is, will the participants fail to notice changes even for stimuli they have intentionally chosen?

Change blindness experiments can be divided into two main categories: explicit and implicit change detection tasks. In the first category, the participants are explicitly instructed to look for changes. One technique is the so called "flicker" paradigm, in which an original and an altered picture are shown in rapid succession with a blank screen inserted between them. The task for the participants is to say when they first detect what the change is (e.g. Rensink, O'Regan & Clark, 2000). Another example is "one-shot", or forced choice, detection experiments, where the participants see only two pictures in succession and then have to decide *if* something was changed or not (e.g. Pashler, 1988). Two important findings from this line of research is that we more easily detect changes to objects that are in the centre of interest in a scene, and that we more readily detect changes to objects or features with higher subjective importance, such as changes made for pictures of human faces (Davies & Hoffman, 2002). However, despite being an important tool for examining the fine details of our visual awareness, as we seldom go about actively trying to detect possible changes in our environment, it could be argued that the findings from these types of studies does not generalize well to "normal" use of vision.

In contrast to this, in implicit change detection tasks, the participants are uninformed about the actual purpose of the experiment. For example, in an experiment involving unexpected cuts in a movie sequence, only 2/3 of the participants took notice when one of the actors was replaced by another actor (Angelone, Levin & Simons, 2003). Similar experiments have also been conducted in real world settings, and with equally dramatic results. In an often quoted study by Simons and Levin (1998), an experimenter approached students and staff on a university campus and started to ask for directions to a nearby building. After a short while the conversation was interrupted by two men carrying a large door between the two persons talking. During this brief intermission the experimenter switched places with one of the men carrying the door, who then continued the interaction about campus directions, Quite remarkably, no more than 7 of 15 participants noticed the switch in this experiment.

But what is often missing in implicit change detection tasks is an active element by which the participants engage in the situation. In Simons and Levin's study, the participants interacted with the person standing in front of them, but they were not required to evaluate or scrutinize the features or characteristics of that person. The setting was an open and friendly situation at a university campus, and the identity of this stranger was of no importance to the participants of the study. On the other hand, if someone had asked for directions in a dim-lit alley at the outskirts of St: Petersburg, or if the interaction had concerned a job interview for a position in the participants own firm, it would have been an entirely different affair. For the participants in Simon and Levin's study, the interaction had no future consequences; it was the kind of encounter in which you know you are not likely to ever meet the other person again.

By including choice as a critical variable in our experimental design, we mean to explore the effects of changing the participants' role in the task. We believe that by letting the participant choose the stimuli that is being changed, the situation becomes drastically altered compared to standard change blindness experiments.

Experiment 1

Method

Participants. Twenty undergraduate students (12 female) at Lund University participated in the study. They received a cinema ticket for their participation. The experiment was described as a test of rapid, intuitive judgment of aesthetic beauty. All participants were naïve about the actual purpose of the experiment.

Material. As stimulus material we used abstract patterns collected from various websites containing "artistic" computer wallpaper for non-commercial use. The presentation size on the screen was 5x5 cm. The pictures were organized in pairs, roughly matched for similarity and attractiveness, covering a range from "similar" to "not so similar". The matching was performed by the authors.

Procedure. Experiment 1 consisted of a simple binary choice task, where participants had to choose which one of two abstract patterns presented on a computer screen they found most aesthetically appealing (see Figure 1). Each trial began when the participants clicked on a left-aligned start-icon that made two patterns appear on the right side of the screen. Participants were given 1500ms to consider their choice, then a beep was played, and they had to move the cursor to the preferred pattern. In addition, the cursor trajectory had to pass through one of two small, color-coded, intermediate squares corresponding to either the upper or the lower pattern on the right. These two squares only became visible after the sound was played, and to prevent learning-effects the vertical position of the squares was randomized within their half of the screen. The upper square was always red and the lower square was always blue, and when the participants passed through one of these squares, the entire screen flashed in matching color for 50ms. The intermediate square and the screen flash were explained to the participants as a way to help them keep the "pace" of the experiment.

After the participants completed their choice, the indicated pattern was framed in the same color as the prior intermediate box, and the non-chosen picture was removed from the screen. The chosen picture remained on the screen for an additional 1500ms after the choice was completed. If the participants had not yet managed to complete a choice 1500ms after the sound alert, the trial ended, and was categorized as a mistrial. The full experiment consisted of 15 trials.[1]

For each participant, on 3 of these trials a change manipulation was introduced (see Figure 1c). On a manipulation trial, the attention-grabbing properties of the midway square and the 50ms screen flash were used to conceal the fact that the two choice alternatives switched places while the participants were moving the cursor across the screen. The manipulation always occurred on trial 7, 10 and 14, but the presentation order of the pairs was randomized.

After all 15 trials had been completed, the participants were given an unannounced memory test. The same pairs of patterns were once again presented, and the participants were asked to indicate which one of the two patterns they had previously found most appealing. In this phase, no time constraints were imposed.

Before the experiment started, the participants were given 10 practice trials. After the experiment all participants were debriefed, and asked whether they consented to have the data from their trials included in the analysis.

[1] In total, 149 of the 900 trials (16.5%) in the three experiments were classified as mistrials and were removed from further analyses. There were no differences between manipulated and non-manipulated trials in the number of mistrials.

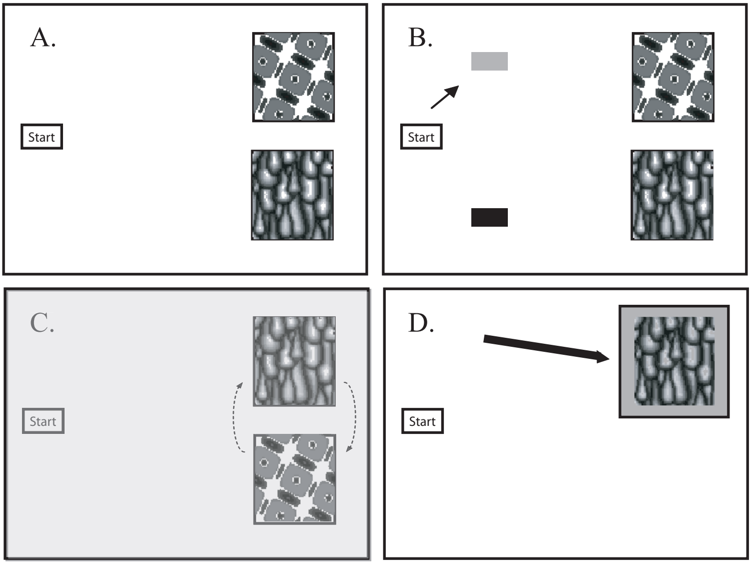


Figure 1. Step-by-step progression of a manipulated trial. **A.** The participants press the start icon and the two pictures emerges on the right hand side. **B.** After 1500ms a beep is played, and the participants moves the cursor to the midway square corresponding to the chosen picture. **C.** When the cursor hits the square the screen is occluded for 50ms. **D.** The participants continue the movement to the chosen (but now altered) picture, and when it is reached the non-chosen alternative is removed from the screen. The chosen picture is then framed and remains visible for 1500ms. Note, for purposes of illustration the pictures are here somewhat magnified compared to their size in the experiment.

A trial was classified as detected if participants showed signs of detection concurrent with the switch (such as explicitly reporting that the patterns had been switched, or that something went wrong with their choice, or by showing signs of confusion and surprise), or if they later in post-experiment interviews claimed to have detected a switch or sensed that something went wrong. For participants that did not show any concurrent signs of detection, a series of increasingly specific questions were asked to make sure their responses were not misclassified as non-detected: "What did you think about the experiment?", "Did you find anything odd with the experiment?" and "Did you notice anything strange with the stimuli presented in the experiment?". At this point, if the participants still revealed no sign of having noticed anything odd with the experiment, they were told that we planned a follow-up study in which the patterns presented sometimes would switch place in mid-trial (i.e. an actual description of the current experiment), and asked if they believed that *they* would have noticed such a switch. Finally, participants were asked if they had noticed anything in the current experiment resembling the hypothetical switches that we had just described. If they answered no to this question, we concluded that they did not consciously notice any of the manipulations made during the experiment.

The post-test memory task was included to measure if a (non-detected) manipulation would influence what the participants remembered as their "own" choice, i.e. if the original choice or the manipulated outcome would be remembered as the picture preferred. The memory task also serves as an independent measure that the pictures were processed after the manipulation was performed. If there would be no differences between the manipulated and the non-manipulated trials on the memory task, it could mean that the pictures were not fully processed after the switch.

Results

In Experiment 1, only 19% of the manipulated trials were categorized as detected (see Figure 3). Of the non-manipulated choices, 86% were remembered correctly in the post-test memory task. For the manipulated trials, the original choice was remembered in 61% of the trials (see Figure 4). The distribution differs significantly between non-manipulated and manipulated trials ($\chi^2_1 = 6.95$, $p = 0.0084<0.5$), showing that the manipulation influence what the participants remembered as being their own choice.

Experiment 2

In the post-test interviews in Experiment 1, most participants described the choice task as being both "real" and meaningful. Nevertheless, it could still be argued that there is something slightly artificial about evaluating abstract patterns, as it is something most people have very little experience of.[2] To provide a more critical test of our approach, we therefore chose to use human faces as the stimuli in Experiment 2. In contrast to abstract patterns, most people have had lots of practice in evaluating faces, and they often have strong opinions about attractiveness. Given this, it seems likely that we would be better at detecting manipulations of faces than most other stimuli.

Method

Participants. Twenty undergraduate students (11 female), at Lund University participated in the study. They were given a cinema ticket for their participation. The experiment was described as a test of rapid, intuitive judgment of attractiveness. All participants were naïve about the actual purpose of the experiment.

Material. Experiment 2 used gray-scale pictures of female faces (taken from the University of Stirling database (PICS), see Figure 2). The pictures were organized in pairs, roughly matched for similarity and attractiveness. The matching was performed by the authors. The presentation size on the screen was 5x5 cm.

Procedure. As in Experiment 1, participants were given the task to choose the picture they preferred the most. However, the exact wording of the instructions was changed from “choose the pattern you find most aesthetically appealing” to “choose the face you find most attractive”. The procedure employed was the same as that in Experiment 1, using 15 trials, three of which were manipulated.

[2] But this is not true for all participants. For instance, an architect student had very strong views on the use of symmetry and what colours could be mixed without unbalancing the picture etc. When the actual procedure was revealed she simply refused to believe that something like that could have taken place.

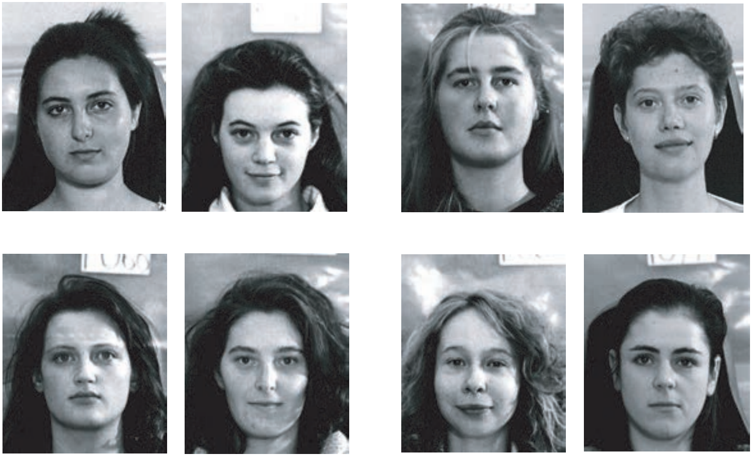


Figure 2. Examples of pairs of pictures used.

Results

In Experiment 2, the detection rate for the manipulated trials was 12% (see Figure 3). This detection rate does not differ statistically from Experiment 1. The participants remembered their choices in 87% of the trials in the post-test memory task. For the manipulated trials, the participants indicated their "original" choice as being what they chose for 76% of the trials (see Figure 4). This number does not differ significantly from the results of the non-manipulated trials.

Experiment 3

At the outset, it seemed likely to us that the change of stimulus material would lead to a difference in detection rate between Experiment 1 and Experiment 2. However, this was not the case. One possible explanation is that there are other factors than the nature of the stimuli that are more important in determining the detection rate. For instance, it may be the case that the relative "distance" between the items paired are not equivalent in the two experiments, e.g. that the face pairs differed less in similarity or attractiveness compared to the pairs of patterns used in Experiment 1[3]. Another possible explanation is that the participants did not fully process the faces after the switch in Experiment 2. In Experiment 1 there was a difference between the manipulated and non-manipulated trials in the post-test memory task. The only candidate explanation for this result is the manipulation itself – i.e. it can be assumed that the participants did in fact look at the pictures after the switch, but did not realize they had been switched, and then later remembered the altered alternative as the one they preferred. But in Experiment 2 there were no differences in the memory test, and therefore we have no independent measure that the participants actually attended to the faces after the switch. And if this would be the case, then it is not that surprising that very few manipulation trials were reported as detected. Thus, to make sure that the manipulated item was properly processed after the manipulation, in Experiment 3, we included a rating task of the chosen and non-chosen faces directly after the choice was completed. Now the pictures stayed visible on the screen until they were rated for attractiveness.

Method

Participants. Twenty undergraduate students (10 female) at Lund University participated in the study. They received a cinema ticket for their participation. The experiment was described as a test of rapid, intuitive judgment of attractiveness. All participants were naïve of the actual purpose of the experiment.

Material. Experiment 3 used the same set of female faces as in Experiment 2.

Procedure. The procedure was the same as Experiment 2 with the following exceptions. After the choice had been indicated on the screen, the chosen picture stayed visible and the participants were asked to rate the face on scale for attractiveness from 1 to 9. The

[3] The use of two sorts of stimuli should be seen more as a further test of the general phenomenon of "choice blindness" rather than a thorough comparison between the likelihood of detection for patterns and faces.

picture remained on the screen until the participants had typed their numerical rating in a box next to the picture. After the chosen picture was rated it was removed, and the non-chosen picture emerged, and the participants were asked to rate this alternative as well. After the participants had done so, the next trial began. As in the previous experiments, the full set consisted of 15 trials, three of which were manipulated.

Results

The detection rate in Experiment 3 was 39%. This is a significantly higher level of detection compared to Experiment 2 (χ^2_1 = 8.75, p = 0.0031<0.05).

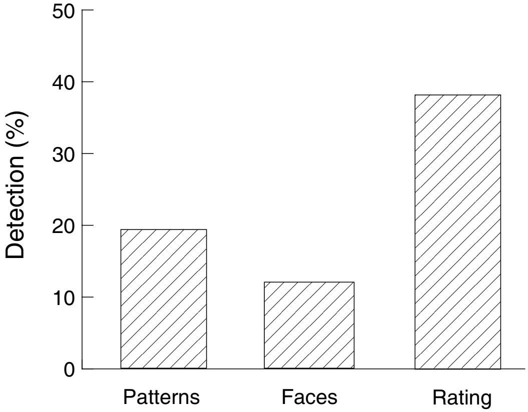


Figure 3. Detection frequency in the three experiments.

The result on the memory test differed markedly when comparing non-detected manipulated and non-manipulated trials. In the non-manipulated trials, the participants remembered their original choice in 92% of the trials. But for the manipulated trials, only 33% of the originally chosen pictures were later remembered as being what the participants preferred initially. The difference between manipulated and non-manipulated trials is significant, (χ^2_1 = 69.62, p = 0.00001<0.05).

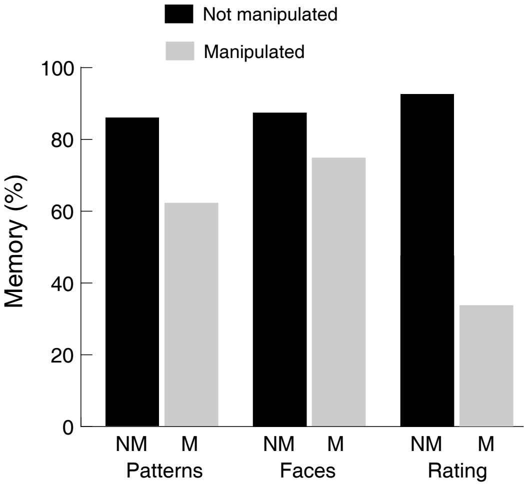


Figure 4. Memory of initial or original choice.

We also analyzed the attractiveness rating. In the rating phase, when the faces were presented again, the participants "knew" that the first face to be rated was the face they originally chose. This means that the participants ought to rate the first face higher, as that was the alternative they thought was the more attractive just a few seconds ago. This is also what we found. In 89% of the non-manipulated trials, the ratings of the participants were consistent with their initial choice. The same was true for the manipulated trials. In 67% of the manipulated trials the participants rated the first picture higher, even though this picture was not the one originally chosen.

Discussion

We have described three experiments involving a simple choice task in combination with a covert manipulation of the outcome of the choices made. The participants in our experiments often failed to notice that the outcome of their choice became the opposite of what they intended, an effect we have termed *choice blindness*. In the experiments presented we have varied both the stimuli used and the choice procedure. The first experiment used abstract patterns, and in the second and third experiment we used pictures of female faces. In all three experiments, the majority of the manipulations remained undetected, indicating that choice blindness is a robust phenomenon.

But given the counter-intuitive nature of the result, we need to carefully consider some objections and alternative interpretations.

To be certain that the pictures were attended to after the switch, we changed the procedure somewhat in the final study. In Experiment 3, we left the faces on the screen until an attractiveness rating was made by the participants. As we see it, it is very difficult to imagine that the participants did not look at the pictures when performing this task.

A similar question is how can we know that 1500ms is enough to form an opinion about aesthetic preference. According to recent research we are remarkably fast at forming opinions about the appearance of faces (Todorov, Mandisodza, Goren, & Hall, 2005). For example, it has been shown that an attractiveness evaluation of a face made after as short exposure as 100ms correlates highly with judgments made after free viewing time (Willis & Todorov, 2006). This indicates that 1500ms is sufficient time to decide at least which of two faces are the more attractive.

In the final experiment, the detection rate also rose to 39%, which differs significantly from the first two experiments. This difference can perhaps be interpreted as an indicator that the pictures were not fully processed after the switch in Experiment 1 and 2. However, a more likely reason for the increased level of detection in Experiment 3 is that the participants were allowed to look at *both* pictures, and that they could make a more explicit comparison when they first rated the manipulated and then the initially preferred picture.

But the most obvious and not yet fully addressed objection to the results is that perhaps the participants actually did notice all the manipulations, but for some reason they just did not tell us. It is possible, but we find it quite unlikely. As was described in the procedure section of the first experiment, the debriefing after the experiment involved asking the participants a series of questions, the last one being if they thought they would have notice if a switch had been made during a “similar” experiment. Of the participants that did not notice any manipulations during the experiment, 85% believed that they would have detected such a switch if it had been performed. When the actual purpose of the experiment was finally revealed, the participants showed considerable surprise, and sometimes even questioned our claim that we *had* switched the pictures. This often strong reaction is hard to account for if the participants really knew about the manipulations. By answering yes to the meta-question about whether they think they *would* have noticed a manipulation, the participants in a sense also set the norm for what should be expected of them. To answer yes to the first question and then deliberately "lie" when asked whether they detected any manipulations seems a very strange thing to do.

Given that the experiments presented here reveal a genuine effect, and that choice blindness is a robust and replicable phenomenon, what are the possible routes to go from here?

Compared to most change blindness studies our small series of experiments employed a quite radical change – i.e. a full identity switch. But nothing dictates that choice blindness experiments must be confined to switches between binary alternatives. For practical purposes it would seem that probing more detailed features of a choice is a more promising option than completely reversing it. An interesting extension of Experiment 3 in the present article would be to replace the rating task with a question, and just ask the participants *why* they preferred the chosen face. If we added salient stimuli such as a pair of earrings, or if we changed the mouth into a smile, it would be very interesting to see if these features would surface in the participants' verbal motivations for their choices.

In relation to the explicit and implicit attraction of different commercial products, choice blindness would seem to be a promising tool to systematically gauge consumer sensitivities (e.g. give them a choice between Coke and Pepsi, artfully make the swap, sit back, and see what will happen). In this way, choice blindness experiments may be able to expose quite drastic discrepancies between the subjective "feel" of a choice (as measured by verbal report) and what properties of the chosen object that are actually relevant for the decision.

Another potential avenue of exploration lies in the wider study of preference formation and change. While it would be a considerable stretch of the imagination to say that the rating procedure of our Experiment 3 induced a new preference for the non-detected manipulated choices, we believe it clearly contains a seed for preference change. When the rating task was included the participants were much more likely to single out the manipulated faces as their original choice, as was revealed by the subsequent memory test. But what happens with this "response tendency" after the experiment is done? Is it immediately forgotten, or does it linger on in the system, perhaps ready to assert itself with a different set of stimuli, or even in a wholly different context? A natural extension of our experiments would be to give participants a longer series of choices and to try to measure to what extent features of the manipulated choice feed back and influence further choices in the series. Evidence indicates that consistent exposure to a particular type of face also changes facial preferences in the direction of that prototype (Kramer & Parkinson, 2005). This is one mechanism by which feedback from manipulated choices might influence future preferences. But more explicit inferential mechanisms might also be at work. According to the prominent tradition of self-perception theory in social-psychology (Bem, 1967) the simple fact that a particular choice has been made is the best evidence participants have of what they actually prefer. With choice blindness manipulations we believe this hypothesis could be given a more rigorous testing than with traditional dissonance reduction paradigms (e.g. see Festinger, 1957; Harmon-Jones & Mills, 1999).

Finally, we believe choice blindness may reveal something about the concept of intention as it is currently understood in psychological research. An obvious question about our results concerns whether choice blindness in any way differs from the general phenomenon of change blindness. Certainly, the surface description of the experiment differs from that of standard change blindness experiments, but this does not necessarily mean that the mechanisms differ. For change blindness in general it is natural to consider "erasing" or "overwriting" of contents in visual short term memory as a promising candidate mechanisms behind the effect (e.g. see Rensink, 2002), but these explanations are not nearly as compelling in the case of choice blindness. The intention to choose something is not supposed to be instantly forgotten. Intentions are supposed to be the guiding structures behind our actions (and phenomenologically speaking, this is what many people claim them to be). But if this is the case, how can the participants in our study intend to choose X, and then 1500ms later fail to notice when they end up with Y? It might be possible to treat our study as simply a variation on the change blindness theme. But in that case, the added ingredient of intention and choice must be explained.

References

Angelone, B. L., Levin, D. T., & Simons, D. J. (2003). The relationship between change detection and recognition of centrally attended objects in motion pictures. *Perception, 32*, 947-962.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*, 183-200.

Blackmore, S. (2002). There is no stream of consciousness. *Journal of Consciousness Studies, 9*, 17-28.

Davies, T. N., & Hoffman, D. D. (2002). Attention to faces: A change-blindness study. *Perception, 31*(9), 1123-1146.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.). *Vancouver studies in cognitive science: Perception* (Vol. 2, pp. 89-110). New York: Oxford University Press.

Harmon-Jones, E., & Mills, J. (1999). *Cognitive Dissonance: Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association.

Kramer, R. S., & Parkinson, B. (2005). Generalization of mere exposure to faces viewed from different horizontal angles. *Social Cognition, 23*, 125-136.

Mitroff, S. R., Simons, D. J., & Franconeri, S. L. (2002). The siren song of implicit change detection. *Journal of Experimental Psychology: Human Perception and Performance, 28(4),* 798-815.

Noë, A. (Ed.). (2002). *Is the visual world a grand illusion?* Thorverton: Imprint Academic.

O'Regan, J. K., & Noe, A. (2002). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939-1011.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of "mudsplashes". *Nature, 398*, 34.

Pashler, H. E. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369-378.

Rensink, R. A. (2000). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition, 7*, 345-376.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245-277.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition, 7*, 127-145.

Simons, D. J. (1996). In sight, out of mind: when object representation fail. *Psychological Science*, *7*(5), 301-305.

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition*, *7*(1-3), 1-15.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059-1074.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people in a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644-649.

Simons, D.J., & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Science, 1*, 261-267.

Smilek, D., Eastwood, J. D., & Merikle, P. M. (2000). Does unattended information facilitate change detection? *Journal of Experimental Psychology: Human Perception and Performance, 26,* 480-487.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623-1626.

Tse, P. U., Sheinberg, D. L., & Logothetis, N. K. (2003). Attentional enhancement opposite a peripheral flash revealed by change blindness. *Psychological Science, 14,* 91-99.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), *592-598.*

PAPER TWO

Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task

Petter Johansson, Lars Hall, Sverker Sikström
& Andreas Olsson

Abstract: A fundamental assumption of theories of decision making is that we detect mismatches between intention and outcome, adjust our behavior in the face of error, and adapt to changing circumstances. But is this always the case? We investigated the relation between intention, choice, and introspection. Participants made choices between presented face-pairs on the basis of attractiveness, while we covertly manipulated the relationship between choice and outcome that they experienced. Participants failed to notice conspicuous mismatches between their intended choice and the outcome they were presented with, while nevertheless offering introspectively derived reasons for why they chose the way they did. We call this effect choice blindness.

Introduction

A fundamental assumption of theories of decision making is that intentions and outcomes form a tight loop (*1*). The ability to monitor and to compare the outcome of our choices with prior intentions and goals is seen to be critical for adaptive behavior (*2-4*). This type of cognitive control has been studied extensively, and it has been proposed that intentions work by way of forward models (*5*) that enable us to simulate the feedback from our choices and actions even before executing them (*6*, *7*).

However, in studies of cognitive control, the intentions are often tightly specified by the task at hand (*8-10*). While important in itself, this type of research may not tell us much about natural environments where intentions are plentiful and obscure, and where the actual need for monitoring is unknown. Despite all its shortcomings, the world is in many ways a forgiving place to implement our decisions in. Mismatches between intention and outcome are surely possible, but when we reach for a bottle of beer, we very seldomly end up with a glass of milk in our hands. But what if the world was less forgiving; what if it instead conspired to create discrepancies between the choices we make, and the feedback we get? Would we always be able to tell if an error was made? And if not, what would we think, and what would we say?

To examine these questions we created a novel choice experiment, which permitted us to surreptitiously manipulate the relationship between choice and outcome that our participants experienced.

We showed picture-pairs of female faces to 120 participants (70 female), and asked them to choose which face in each pair they found most attractive. In addition, on some trials, immediately after their choice, they were asked to verbally describe the reasons for choosing the way they did. Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other (Fig. 1). Thus, on these trials, the outcome of the choice became the opposite of what they intended.

Each subject completed a sequence of 15 face-pairs, three of which were manipulated. The manipulated face-pairs always appeared at the same position in the sequence, and for all of these pairs participants were asked to state the reasons behind their choice. Verbal reports were also solicited for three trials of non-manipulated pairs (*11*).

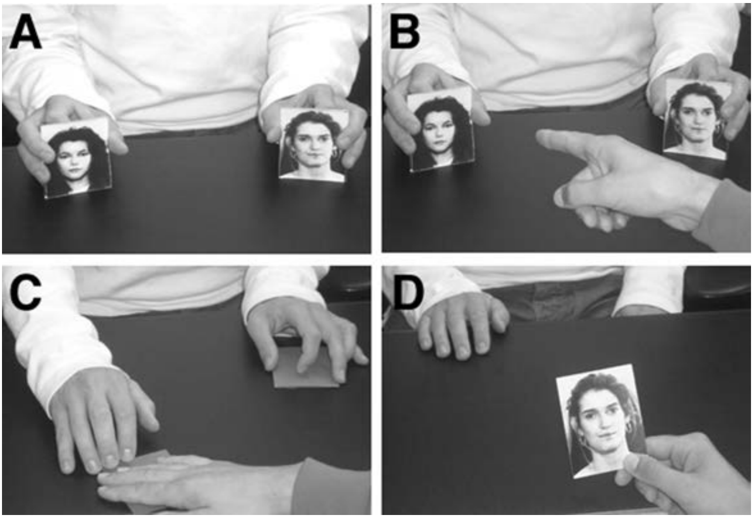


Figure 1. A snapshot sequence of the choice-procedure during a manipulation trial. **A.** Participants are shown two pictures of female faces and asked to choose which one they find most attractive. Unknown to the participants, a second card depicting the opposite face is concealed behind the visible alternatives. **B.** Participants indicate their choice by pointing at the face they prefer the most. **C.** The experimenter flips down the pictures and slides the hidden picture over to the participants, covering the previously shown picture with the sleeve of his moving arm. **D.** Participants pick up the picture, and are immediately asked to explain why they chose the way they did.

The experiment employed a three by two between-group factorial design, with deliberation time and similarity of the face-pairs as factors. For time, three choice conditions were included: one with two seconds of deliberation time, one with five, and one where participants could take as much time as they liked. Participants generally feel that they are able to form an opinion given two seconds of deliberation time (supporting online text). Nevertheless, the opportunity for participants to enjoy free deliberation time was included to provide an individual criterion of choice.

For similarity, we created two sets of target faces, a high similarity (HS) and a low similarity set (LS) (Fig. S1). Using an interval scale from 1–10 where 1 represents "very dissimilar" and 10 "very similar", the HS set had a mean similarity of 5.7 (SD=2.1), and the LS a mean similarity of 3.4 (SD=2.0).

Detection rates for the manipulated pictures were measured both concurrently, during the experimental task, and retrospectively through a post-experimental interview (*11*, supporting online text).

There was a very low level of concurrent detection. With a total of 354 manipulated trials performed, only 46 (13%) were detected concurrently. Not even when participants were given free deliberation time and a set of low similarity faces to judge, were more than 27% of all trials detected this way. There were no significant differences in detection rate between the 2s and 5s viewing time conditions, but there was a higher detection rate in the free compared to the fixed viewing time conditions ($t(118) = 2.17$, $p = 0.03 < 0.05$). Across all conditions there were no differences in detection rate between the HS and the LS set (Fig 2A). In addition, there were no significant sex or age differences in detection rate. Tallying all forms of detection across all groups revealed that no more than 26% of all manipulated trials were exposed.

But these figures are inflated even so. The moment a detection is made, the outlook of the participants change: they become suspicious, and more resources are diverted to monitoring and control. To avoid such cascading detection effects it is necessary to discard all trials after the first detection is made. Fig. 2B shows detection rates with this correction in place. The overall detection rate is significantly lower ($t(118) = 3.21$, $p = 0.0017 < 0.05$), but no prior conclusions are affected by the use of this data-set (for the percentage of participants that detected the manipulation, see Fig. S2).

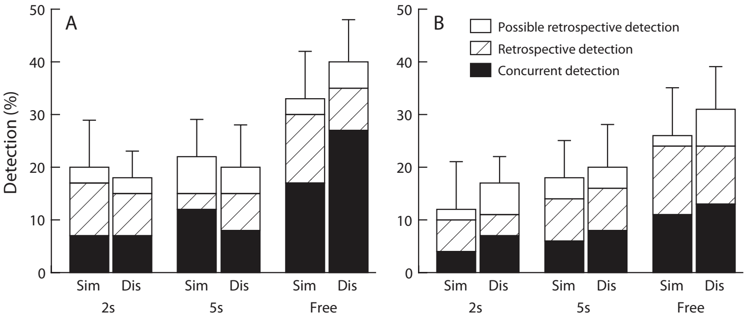


Figure 2. Percent detection, divided into deliberation time and similarity. **A**. All trials included. **B**. Corrected for prior detections.

Our experiment indicates that the relationship between intentions and outcomes may sometimes be far looser than what current theorising has suggested (9, 6). The detection rate was not influenced by the similarity of the face-pairs, indicating the robustness of the finding. The face-pairs of the LS set bore very little resemblance to each other, and it is hard to imagine how a choice between them could be confused (S1, supporting online text). The overall detection rate was higher when participants were given free deliberation time. This shows the importance of allowing individual criteria to govern choice, but it is not likely to indicate a simple subjective threshold. The great majority of the participants in the 2s groups believed themselves to have had enough time to make a choice (as determined by post-test interviews), and there was no difference in the actual distribution of choices among the pairs from fixed to free deliberation time.

Next, we examined the relationship between choice blindness and introspective report. One might suspect that the reports given for non-manipulated (NM) and manipulated (M) trials would differ in many ways. After all, the former reports stem from a situation common to everyday life (revealing the reasons behind a choice) while the latter reports stem from a truly anomalous one (revealing the reasons behind a choice one manifestly did not intend to make).

We classified the verbal reports into a number of different categories that potentially could differentiate between NM- and

M-reports. For all classifications we used three independent blind raters, and interrater reliability was consistently high (Supporting Online Text, Table S1). We found no differences in the number of empty reports (when participants were unable to present any reasons at all), or in the degree to which reports were phrased in present or past tense (which might indicate whether the report is made in response to the present face, or the prior context of choice). Neither did the length of the statements, as measured by number of characters, differ between the two sets (NM = 33, SD = 45.4, M = 38, SD = 44.4), nor the amount of laughter present in the reports (with laughter being a potential marker of nervousness or distress). We found significantly more dynamic self-commentary in the M-reports ($t(118) = 3.31$, $p = 0.001<0.05$). This is an interesting type of report in which participants come to reflect upon their own choice (typically by questioning their own prior motives), although even in the M-trials such reports occurred infrequently (5%).

We rated the reports along three dimensions: emotionality, specificity and certainty (using a numeric scale from 1–5). Emotionality was defined as the level of emotional engagement in the report, specificity as the level of detail in the description, and certainty as the level of confidence in their choice the participants expressed. There were no differences between the verbal reports elicited from NM and M trials with respect to these three categories (Fig. S3). This is a striking result. Seemingly, the M-reports were delivered with the same confidence as the NM-ones, and with the same level of detail and emotionality. One possible explanation is that overall engagement in the task was low, and this created a floor effect for both NM and M-reports. However, this cannot be the case. All three measures were rated around the midline on our scale (Emotionality = 3.5, SD = 0.9, Specificity = 3.1, SD = 1.2, Certainty = 3.3, SD = 1.1). Another possibility is that the lack of differentiation between NM and M-reports is an indication that delivering an M-report came naturally to most of the participants in our task. On a radical reading of this view, a suspicion would be cast even on the NM-reports. Confabulation could be seen to be the norm and truthful reporting something that needs to be argued for.

To scrutinize these possibilities more closely we conducted a final analysis of the M-reports, adding a contextual dimension to the classification previously used. Fig. 3 shows the percentage of M-reports falling into eight different categories. The *specific confabulation* category contains reports that refer to features unique to the face participants ended up with in a manipulated trial. As these reports cannot possibly be about the original choice (i.e. “I chose her [the blond woman] because she had dark hair”), this would indeed be an indisputable case of “telling more than we can know” (*12*). Equally interesting is the *original choice* category. These are reports that must be about the original choice, because they are inconsistent with the face participants ended up with (i.e. “I chose her because she smiled [said about the solemn one]”). Here, despite the imposing context of the manipulated choice, vestiges of the original intention are revealed in the M-reports. Analogous to the earlier example of confabulation, this would be an unquestionable case of truthful report.

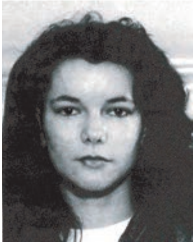
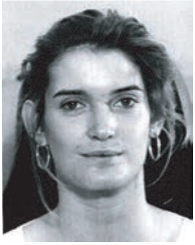
Type	%		
Specific Conf.	13.3		She's radiant. I would rather have approached her at a bar than the other one. I like earrings! [M]
Detailed Conf.	17.3	She look like an aunt of mine I think, and she seems nicer than the other one. [F]	
Emotional Conf.	9.3		Yes, well, [laughter] she looks very hot in this picture. [M]
Simple Conf.	10.8		Just a nice shape of the face, and the chin. [M]
Relational Conf.	21.3		I thought she had more personality, in a way. She was the most appealing to me. [F]
Uncertainty	11.6	Eh.. I don't know. [F]	
Dynamic report	5.2		Oh, [short laughter] Why did I choose her? She looks very masculine! [M]
Original choice	11.2	Because she was smiling.[F]	

Figure. 3. Frequency distribution of the contents of the M-reports aligned along a rough continuum from confabulatory to truthful report. Sample sentences (translated from Swedish) are drawn from the set of reports for the displayed face-pair. Letters in brackets indicate whether the report was given by a male [M] or a female [F] participant. The specific confabulation category contains reports that refer to features unique to the face participants ended up with in an M-trial. The detailed and emotional confabulation categories contain reports that rank exceptionally high on detail and emotionality (>4.0 on a scale from 1–5). The simple and relational confabulation categories concerns reports where the generality of the face descriptions precluded us from conclusively associating them with either of the two faces (i.e. everybody has a nose, or a personality). The category of uncertainty contains reports dominated by uncertainty (<2 on a scale from 1-5). The next category contains the dynamic reports. The final category contains reports that refer to the original context of choice.

In summary, when evaluating facial attractiveness participants may fail to notice a radical change to the outcome of their choice. As a notable extension of the well known phenomenon of change blindness (*13*), we call this effect *choice blindness* (supporting online text). This finding can be used as an instrument to estimate the representational detail of the decisions that humans make (*14*). We do not doubt that humans can form very specific and detailed prior intentions, but as the phenomenon of choice blindness demonstrates, this is not something that should be taken for granted in everyday decision tasks. While the current experiment warrants no conclusions about the mechanisms behind this effect, we hope it will lead to an increased scrutiny of the concept of intention itself. As a strongly counter-intuitive finding, choice blindness warns of the dangers of aligning the technical concept of intention too closely with common sense (*15*, *16*).

In addition, we have presented a novel method for studying the relationship between choice and introspection. Classic studies of social psychology have shown that telling discrepancies between choice and introspection can sometimes be discerned in group-level response patterns (*12*), but never for each of the individuals at hand. In the current experiment, using choice blindness as a wedge, we were able to 'get between' the decisions of the participants and the outcomes they were presented with. This allowed us to show, unequivocally, that normal participants may produce confabulatory reports when asked to describe the reasons behind their choices. But more importantly, the current experiment contains a seed of systematicity for the study of choice and subjective report. The possibility of detailing the properties of confabulation that choice blindness affords, could give researchers an increased toehold in the quest to understand the processes behind truthful report.

References and Notes

1. K. R. Ridderinkhof, W. P. M. van den Wildenberg, S. J. Segalowitz, C. S. Carter, *Brain Cogn.* **56**, 129 (2004).
2. K. R. Ridderinkhof, M. Ullsberger, E. A. Crone, S. Nieuwenhuis, *Science* **306**, 443 (2004).
3. M. Ullsperger, D. Y. Cramon, *Cortex* **40**, 593 (2004).
4. M. E. Walton, J. T. Devlin, M. F. S. Rushworth, *Nat. Neurosci.* **7**, 1259 (2004).
5. P. Haggard, S. Clark, *Conscious. Cogn.* **12**, 695 (2003).
6. R. Grush, *Behav. Brain Sci.* **27**, 377 (2004).
7. D. M. Wolpert, K. Doya, & M. Kawato, *Phil. Trans. R. Soc. Lond. B* **358**, 593 (2003).
8. J. G. Kerns, J. D. Cohen, A. W. MacDonald III, R. Y. Cho, V. A. Stenger, C. S. Carter, *Science* **303**, 1023 (2004).
9. R. Hester, C. Fassbender, H. Garavan, *Cereb. Cortex* **14**, 986 (2004).
10. H. C. Lau, R. D. Rogers, P. Haggard R. E. Passingham, *Science* **308**, 1208 (2004).
11. Materials and methods are available as supporting material on *Science* Online.
12. R. E. Nisbett, T. D. Wilson, *Psychol. Rev.* **84**, 231 (1977).
13. R. A. Rensink, *Annu. Rev. Psychol.* **53**, 245 (2002).
14. D. J. Simons, R. A. Rensink, *Trends Cogn. Sci.* **9**, 16 (2005).
15. D.C. Dennett, *The Intentional Stance* (MIT-Press, Cambridge, MA, 1987).
16. D. M. Wegner, *The Illusion of Conscious Will* (MIT-Press, Cambridge, MA, 2003).
17. We thank D. de Léon, K. Holmqvist, P. Björne, P. Gärdenfors, T. Dickins, N. Humphrey, A. Marcel, Q. Rahman, E. Adams, N. Bolger, B. Cohen, and E. Phelps for useful comments and discussions. We also thank C. Balkenius for providing the illustrations for the article, and P. Rosengren for invaluable advice concerning the use of card magic techniques. This work was supported by The New Society of Letters at Lund (PJ), The Knowledge Foundation, The Erik Philip-Sörensen Foundation (LH), and the Swedish Research Council (SS).

PAPER TWO: APPENDIX

SUPPORTING ONLINE MATERIAL

Material and Methods

Participants. One hundred and twenty participants (70 female) participated in the study (mean age ± SD, 26 ± 8.3). Participants were drawn from a mixed student and non-student population. As a cover story for the experiment participants were told that the experimenters were interested in choice and facial attractiveness. After the experiment participants were debriefed about the true nature of the design, and given the opportunity to voice any concerns. All participants then gave informed consent. Two participants were removed from the subsequent analysis because they were immediately able to discern how the card trick was performed (due to flawed presentations by the experimenter).

Experimental Procedure. Participants were shown pairs of grayscale pictures of female faces, and were given the evaluative task of choosing which face in each pair they found most attractive. In addition, on some trials, immediately after the choice, they were asked to verbally describe the reasons for choosing the way they did. Participants had been informed in advance that we would solicit verbal reports about their intentions during the experiment, but not the specific trials for which this was the case. Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended.

The experiment employed a three by two factorial design, with deliberation time and similarity of the face-pairs as factors. For time, three choice conditions were included: one with two seconds of deliberation time, one with five, and a final condition where participants could take as much time as they liked.

For similarity, we created two sets of target faces, a high similarity (HS) and a low similarity set (LS). Using an interval scale from 1–10 where 1 represents "very dissimilar" and 10 "very similar", the HS set had a mean similarity of 5.7 (SD=2.08), and the LS a mean similarity of 3.4 (SD=2.00). The face pictures were collected from the The Psychological Image Collection at Stirling (PICS), online face database (http://pics.psych.stir.ac.uk/). We used pictures from the Nottingham and the Stirling collection, and 15 face-pairs were constructed on the basis of a rough matching of the photos (position of the head, background luminance, background color, attractiveness, etc.). After this, a group of independent raters (n=15) coded all pairs for similarity, and six pairs were selected for the HS and LS set.

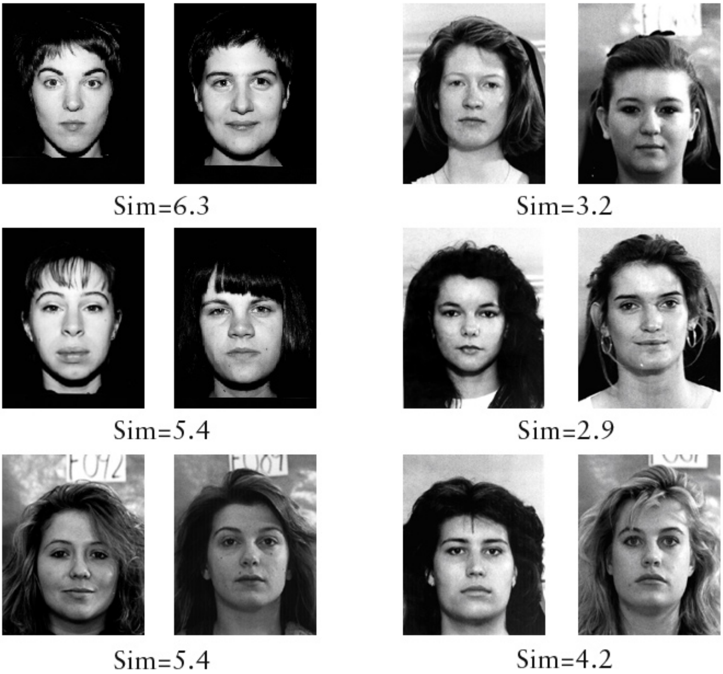


Figure S1. The face-pairs used for the manipulated trials in the experiment, with similarity scores displayed below each pair. The High-Similarity group is shown on the left, and the Low-Similarity group on the right.

Each participant completed a sequence of 15 face-pairs, three of which were manipulated. The manipulated face-pairs always appeared at the same position in the sequence (7, 10, 14), and in the same order. For all of these pairs participants were asked to state the reasons behind their choice. All reports were recorded and later transcribed. To provide a comparison class, verbal reports were also solicited for three trials of non-manipulated pairs. The non-manipulated (NM) and manipulated (M) pairs were counterbalanced during the experiment (with the LS set serving as non-manipulated control in the HS-groups, and equally the other way around).

Using standard change blindness terminology, this task would be described as involving incidental change detection, one-shot stimulus presentation, and occlusion-contingent change (*1*). The period the hidden picture remained unseen on the table during the switch was approximately 2s from drop-down to pick-up (with some variations due to natural arm movements).

Detection rates for the manipulated pictures were measured both concurrently and retrospectively, with three graded levels of detection being used for our analysis. A trial was classified as concurrently detected if participants showed any signs of detection during the switch (such as explicitly reporting that the faces had been switched, or indicating that something went wrong with their choice). After the experiment all participants were asked a series of increasingly specific questions in a post-test interview to investigate whether they had any inkling that something had gone wrong ("What did you think about the experiment?", "Did you find anything odd with the experiment?" and "Did you notice anything strange with the stimuli presented in the experiment?"). Participants that revealed no signs of detection in this procedure were then presented with a hypothetical scenario describing an experiment in which the faces they choose between are surreptitiously switched (i.e. the very experiment they had just participated in), and asked whether they thought they would have noticed such a change. This question was included to determine the folk-psychological status of our design (i.e. whether it would be perceived as counter-intuitive or not). Finally, all participants were debriefed about the true nature of the design, and asked if they had noticed anything in the experiment resembling the switches that we had just described. If they answered "no" to this question, we concluded that they did not consciously notice any of the manipulations made during the experiment. All other participants were then given an opportunity to sort through their chosen pictures and indicate which faces they felt could have been manipulated. A trial was classified as retrospectively detected if participants picked out the corresponding manipulated picture in the set. If participants did so, but also indicated any number of false positives, those trials were classified in a category called possible retrospective detection. The inclusion of this category in the analysis was meant to compensate for the possibility of underreporting due to unknown social factors present in the interview.

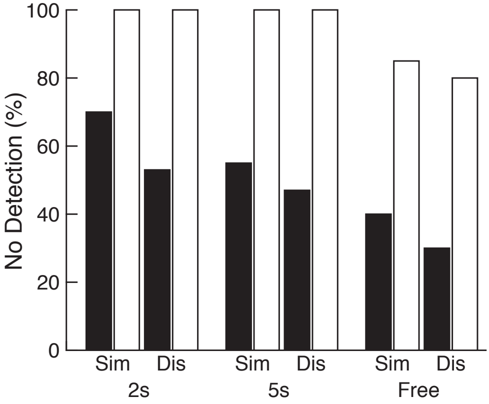


Figure S2. Percentage of subjects across the different conditions failing to detect all manipulations (black bars), and at least one manipulation (white bars).

Supporting Text

Detection criteria

Taken together we believe the three categories of detection in our experiment gave the participants a fair chance to voice their concerns, and that they go a long way towards ensuring that no conscious detections were left out. In devising a cued-procedure (i.e. allowing participants to sort through their chosen faces) for the retrospective detection test, and the inclusion of participants that even named false positives in the possible retrospective detection category, we tried to err on the side of being too liberal about what to count (for example, if we had terminated our post-test interview after the initial question about whether participants experienced anything odd during the experiment, only a single retrospective detection would have been registered).

However, when discussing detection criteria it is very difficult to remain neutral with respect to different theories of consciousness. For our concurrent detection criterion we relied on spontaneous verbal report by the participants (even if we did not demand an articulate response). But why should we give special status to verbal reports? According to a prominent tradition in the field of implicit learning we should always be looking for the most exhaustive measure of conscious processing (*2-4*), otherwise we might end up establishing false dissociations between differentially sensitive measures of the same conscious resource. This methodological principle has been dubbed the *sensitivity criterion* (*4*).

The customary way of adhering to the sensitivity criterion is to use concurrent forced-choice to measure conscious detection (*5*). Applied to the current experiment this method would probably have resulted in more instances of detected manipulations than the spontaneous reporting we relied on. However, as we see it, there is a substantial difference between being unaware of a specific influence in a natural context, and being similarly unaware of some stimuli, influence, or process under the most penetrating probe (i.e. what the sensitivity criterion prescribes). The experiment was meant to simulate a choice situation in which no prior evidence indicates that a high level of monitoring is needed, and it is only very rarely that natural conversations are accompanied by clever simultaneous forced choice questions and reaction time measures to exhaustively probe our conscious knowledge.

Of course, any attempt at an ecological explanation of decision making would have to accommodate both non-vigilant (relaxed, non-suspicious), as well as vigilant (guarded, suspicious) choice. Depending on whether the correction for prior detection is applied in our experiment it can be seen to occupy different positions along this dimension, with the uncorrected version situated further towards the suspicious pole. Had our experiment been framed as an explicit detection task, we have no doubt that most participants would have been able to spot the manipulations immediately.

Previous Studies

Before implementing our main experiment we ran a series of basic studies exploring the phenomenon of choice blindness. These studies add to the evidential base of the current experiment by demonstrating the effect in a different medium and with a different design, and with different types of stimuli.

First, we created an experiment in which participants had to choose which one of two abstract patterns presented on a computer screen they found most aesthetically appealing (the patterns were collected from various websites containing 'artistic' computer wallpaper for non-commercial use). Each trial began when the participants clicked on a left-aligned start-icon that made the two patterns appear on the right side of the screen. Participants were given 1500ms to consider their choice, then an alerting sound was played, and they had to move the cursor to the preferred pattern. In addition, we required the cursor trajectory to the target pattern to pass through one of two small, color-coded, intermediate squares corresponding to either the upper or the lower pattern on the right. When the participants passed through one of these squares, the entire screen flashed in matching color for 50ms. Similarly to the current experiment, on some trials, a mismatch between choice and outcome was created. On a manipulated trial, the attention-grabbing properties of the midway square and the

50ms screen flash were used to conceal the fact that the two choice alternatives switched places while the participants were moving the cursor across the screen. The full experiment consisted of 15 trials, three of which were manipulated. Twenty participants (12 female) were tested. In total, counting both concurrent and retrospective detections (and using data uncorrected for prior detections), 19% of the manipulated trials were detected.

In a subsequent experiment we used the same decision paradigm, but instead of abstract patterns we used female faces to choose between (as with the current experiment, the PICS online face database was used for the selection). In addition, immediately after their choice participants were asked to state their reasons for choosing the way they did. The experiment consisted of 30 trials, five of which were manipulated. Twenty-two participants (14 female) were tested, and the total detection rate was 32%. However, with five manipulated trials used rather than three, prior detection made a larger impact on the detection rate. Using corrected data, detection rate drops to 20% (this can also be seen in the fact that 9 out of 22 participants did not detect any of the five manipulations). Analysis of the verbal reports revealed similar patterns as in our main experiment, with no clear differentiation between the NM and M-reports.

Finally, we used the same setup as in the previous experiment, but with a set of male faces to choose between (again, the faces were collected from the PICS database). In addition, eye-tracking was used to verify that participants attended to the pictures both during the deliberation phase, and when giving their verbal reports. Eighteen participants (12 female) were tested, and total detection rate was 37% (29%, when corrected for prior detection). Analysis of the eye-tracking data revealed that participants attended to the pictures both before and after their choice. Again, analysis of the verbal reports revealed no differences between the NM and M-trials.

Throughout the whole series of studies, and in pilot controls, we conducted post-experiment interviews to determine the subjective confidence participants felt about their choices. While opinion about whether the task was difficult or not fluctuated somewhat, a great majority of the participants believed 1500ms was enough time to make a proper choice.

Choice Blindness Blindness

When we claim that it is hard to believe how a choice between the face-pairs in our study could be confused, we are not simply asking our readers to inspect the pairs in Fig. S1 and form their own opinions. During the post-test interview in the experiment we requested all participants that had not yet voiced any suspicion to consider a hypothetical choice-manipulation extension of our experiment (see above, *experimental procedure*) and asked them if they believed they would have noticed such a change. The result shows that of the participants in our study that failed to notice any of the manipulations, 84% believed that they would have been able to do so (a result comparable to similar metacognitive probes in the change blindness literature (*6, 7*). Accordingly, many participants also showed considerable surprise, even disbelief at times, when we debriefed them about the true nature of the design. This effect of "choice blindness blindness" was also evident in our earlier computer-based experiments, with roughly 87% percent of participants claiming that they would have noticed if the outcome of their choice had been manipulated in the hypothetical experiment we described.

Analysis of the introspective reports

Analysis of verbal reports often proceeds in several iterations, where the early rating results are used to distill a more distinct and consistent categorization (*8-10*). The contrastive analysis we employed to analyze potential differences between the NM and M-trials, were based on a two-stage classification of the verbal reports of our participants. As the NM-reports stem from a situation common to everyday life, while the M-reports are produced in response to a truly anomalous experimental probe, it would be natural to suspect that the two types of reports would differ in many ways. To investigate this, we identified four simple variables, based on 'surface' features of the reports (empty reports, laughter, the length of the reports, and the tense of the reports), and four promising psychological dimensions (emotionality, specificity, certainty, and dynamic self-reference). For all of these items common sense would suggest that the NM- and M-reports ought to be differentiated: participants in the M-trials ought to be more likely to say “I don’t know”, or “I have no idea”, when asked to state the reasons behind a choice they did not make (*empty reports*); they ought to give shorter reports (*length of report*); they ought to produce more nervous laughter or giggle in response to the unfamiliarity of the situation (*laughter*); and they ought to make more references to past tense in their reports, talking about what they thought in relation to the original context of choice, rather than what they think about the picture they are seeing now (*tense of report*); participants in the M-trials ought also to show less emotional engagement, as the M-reports are given in response to the alternative they did not prefer (*emotionality*); they ought to make less specific and detailed reports, as no prior reasons have been formulated for the manipulated alternative (*specificity*); they ought to express less certainty about their choice (*certainty*), and they ought to reflect more about the current choice situation, and engage in more dynamic self-commentary, typically by questioning their own prior motives (*dynamic report*).

Independent raters first made untrained judgments for the classifications and dimensions we had identified (except length of report, which we calculated using the spreadsheet software). Each rater coded the whole set of reports. Three raters coded the four simple variables, and we used another three raters for the more complex scales. Next, we consulted with the group of raters, and used their input to sharpen our criteria and to calibrate our scales. Then a second group of (3+3) independent raters was given the same task. Before the new rating procedure each rater was provided with a training kit containing definitions and examples (available upon request from the authors). The approximate amount of training and instruction given to the raters ranged from 15 minutes for the simple categories, to approximately 45 minutes for the psychological dimensions. This procedure resulted in good interrater agreement (see discussion below).

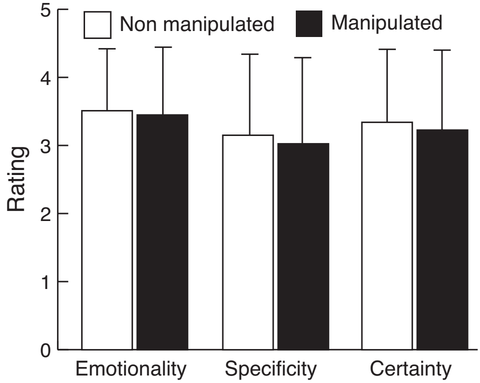


Figure S3. The content of the verbal reports rated along the dimensions of (A) emotionality, (B) specificity, and (C) certainty. As can be gleaned from the figure, no significant differences between the non-manipulated and the manipulated reports were found with respect to these three dimensions.

The final contextual analysis proceeded somewhat differently. Here, we were interested in investigating the *relation* between the content of the M-reports and the picture they were presented with at the time of the report. More specifically, raters were given the task of classifying whether the reports contained references to unique or distinguishing features of one of the two faces in each pair – i.e. whether the report was *about* a particular face. As with the other categorizations, this task was first given to three independent raters, then calibrated, and then given to another three raters for a final classification. However, as we wanted the classification to be unquestionable, we only included instances of reports in the final analysis for which the raters had absolute agreement.

The introspective reports collected in our experiment are rich and varied, and it is important not only to search for differences between the NM- and M-reports, but also to provide a descriptive representation of the content of these reports. In Fig. 3 we plot the frequency of eight different categories for the M-reports, laid out in a rough continuum between confabulatory and truthful report. The figure is built around epistemic ‘anchor points’ at each end (i.e. the categories ‘specific confabulation’ and ‘original choice’, for which we can be certain that the reports are either confabulatory or truthful), and then reports are collated according to the degree to which they are likely candidates to be confabulations. For example, a report saying “[I chose her] because she has a nice face” is placed at the center of the continuum. A report of this kind contains no information that allows us to assign it to either of the two choice alternatives (i.e. everybody has a face; it is not a distinguishing feature). Also, it has no additional interesting properties, like a strong emotional component, or a high degree of specificity. There are good reasons to believe this report in fact *is* a confabulation (after all, it is produced in direct response to a face the participant manifestly did not choose), but the content of the report gives no further clues about whether this is the case. In contrast, a report that is highly emotional, like “I simply love this girl”, represents a more severe mismatch between the actual choice and the manipulated outcome, and is placed closer to the confabulatory pole. On the other hand, a report that is devoid of any content, like “I don’t know”, or “I can’t tell”, is marked by uncertainty, and is therefore placed further towards the truthful pole.

As mentioned above, a great strength of our methodology is that it allows for us to detect categories of reports in the M-trials that undoubtedly refer to the manipulated picture (“specific confabulation”) or the original context of choice (“original choice”). But currently there is no way to make these distinctions for the NM-reports, which preclude any comparisons between NM- and M-reports for these two categories. The categorization in Figure 3 is mutually exclusive, and weighted by proximity to the two poles. Reports were first placed in the two outmost categories, then in the category of dynamic report, then in detailed confabulation, then according to emotionality, then uncertainty, and finally the rest of the reports were divided into the simple and relational categories. As we see it, the resulting distribution gives a highly interesting impression of the contents of the M-reports, revealing the variable nature of, and the varying tendencies for, truthful and confabulatory report by our participants.

To measure interrater reliability (IRR) we used Pearson’s product moment correlation as our main index. Table S1 shows the IRR levels for all variables and dimensions used in our analysis. The IRR is based on the average of the pair-wise Pearson product moment correlations between the three raters. Pearson’s *r* is a well-established index that measures internal consistency and covariation between raters. As we were mainly interested in investigating potential differences between the two classes of reports (NM and M), a covariation index is appropriate to use. However, it should be noted that estimates of IRR may fluctuate between different measurements. In the words of (*9*): "Despite all the effort that scholars, methodologists, and statisticians have devoted to developing and testing indices of intercoder reliability, there is no consensus on a single 'best' index" (p. 593). As (*9*) contend, it is advisable to calculate IRR using more than one measure, and to demonstrate consistency across measures. Thus, although r is a commonly applied statistic for estimating the IRR, we have also chosen to include calculations based on Intra Class Correlation (ICC), and Krippendorff's Alpha (see Table S1). The ICC is a measure widely endorsed to estimate IRR when ratings from more than two judges are considered (*11, 12*). We based our ICC on a two-way ANOVA, treating both the targets (verbal reports) and the raters as the random factors. Because systematic differences among levels of ratings were considered relevant, a measure of absolute agreement was chosen. In the terminology proposed by Shrout and Fleiss (*13*), we computed a case 2 model with three raters ($ICC_{2,3}$). Krippendorff's Alpha is a chance corrected index of absolute agreement, which generally is considered to be a 'conservative' measurement of IRR (*9, 14*). As with the methods used to calculate IRR, there are no absolute standards about what constitutes acceptable levels of reliability (*9*), but as a result primarily intended for research purposes, our IRR levels must be considered high (*14-16*).

Variable	Pearson's r	ICC	Krippendorff
Laughter	0.95	0.98	0.95
Empty Reports	0.82	0.92	0.80
Tense	0.92	0.97	0.93
Emotionality	0.79	0.91	0.78
Specificity	0.88	0.96	0.88
Certainty	0.78	0.89	0.73
Dynamic Report	0.80	0.92	0.80
Specific Conf.	0.78	0.91	0.78
Original Choice	0.82	0.93	0.82

Table S1. Three alternative interrater reliability (IRR) measures for each variable used in our analysis. For each variable, the rating was performed by three independent raters. Listed in the left column are the average pair-wise Pearson product moment correlations between the three raters. The center column contains values for a case 2 model Intraclass Correlation (ICC) of absolute agreement, treating both the verbal reports and the raters as random factors. The right column contains values for Krippendorff's Alpha, a chance corrected index of absolute agreement, which is generally considered to be a conservative measurement of IRR. As can be seen in the figure, the IRR levels are uniformly high, with good consistency between measures.

From Change Blindness to Choice Blindness

It has been known for a long time that human participants are inept at noticing changes in a visual scene when the transients accompanying that change no longer convey information about its location, a phenomenon that has been termed *change blindness* (*17*). During the last decade the phenomenon of change blindness has generated an extraordinary amount of interest among researchers interested in the workings of the human visual system (*1*), particularly with reference to the mechanisms of attention (*18*), and the nature of visual consciousness (*19*). But despite this, the full potential of change blindness as a tool for studying the human mind is far from realized. Why should change blindness only be used to study distinctly *visual* aspects of human cognition? (*1*) writes: "the study of change detection has evolved over many years, proceeding through phases that have emphasized different types of stimuli and different types of tasks. *All stud-*

ies, however, rely on the same basic design. An observer is initially shown a stimulus… a change of some kind is made to this stimulus… and the [visual detection] response of the observer is then measured" (p. 251, our emphasis). We were interested in the possibility of modifying this basic design to incorporate other non-perceptual elements of cognition. In particular, we wanted to investigate the relationship between intention, choice, and introspection. Our approach involves embedding different forms of change-manipulations in simple decision tasks and concurrently probe participants about the reasons for their choice. We see three main reasons for why this constitutes a novel and significant extension of the change blindness literature.

Firstly, choice blindness brings the conceptual tools of change blindness from the basic study of perception into a new domain of inquiry. Research on change blindness has occasionally contained elements of interaction (most notably, the real-person interactions in *20, 21*), and at least one task in which the actions of the participants have functional relevance has been investigated (*22*), but ours is the first study to incorporate meaningful decision making in an evaluative task. In change blindness experiments participants are usually more likely to notice changes when they concern features of particular relevance to the scene, or if they are of central interest to the participants, or if the participants are particularly knowledgeable about them (*1, 23*). For choices it would almost seem to be a defining feature that they concern properties of high relevance and interest, or things we are very knowledgeable about. But in the current experiment, in the great majority of trials, our participants were blind to the mismatch between choice and outcome. While intending to choose X (a central-interest, non-peripheral, valenced stimuli), they failed to take notice when ending up with Y. This is a result that ought to be surprising even to the most seasoned change blindness researcher. On a more general level, we believe decision making to be domain with immediate intuitive appeal. There can be no doubt that we often care deeply about what we choose. The fact that we may be blind to the outcome of these choices is a finding that potentially could change our most intimate conceptions of ourselves as decision makers.

Secondly, choice blindness can be used to study introspection and preference change. Looking at the wider methodological aspects of our work, we believe choice blindness opens up exiting new opportunities for research. During the course of a normal day humans make countless choices: some slow and deliberate, some rapid and intuitive, some that carry only minor significance, and some that impact greatly on our lives. But for all the intimate familiarity we have with everyday decision making, it is very difficult to probe the representations underlying this process, or to determine what we can know about them from the 'inside', by reflection and introspection (*24-26*). The greatest barrier for scientific research in this domain is the nature of subjectivity. How can researchers ever corroborate the reports of the participants involved, when they have no means of challenging them? As philosophers have long noted, incorrigibility is a mark of the mental (*27*). Who are *they* to say what *my* reasons are? But as we have shown in the current analysis, choice blindness can be used to investigate the properties of introspective report. Beyond the exploratory work reported here, we envisage the collection and construction of large scale databases of reports given in relation to NM- and M-trials. By varying stimulus, personality and situational dimensions within the body of reports, powerful systematic comparisons between NM- and M-reports will become possible (both hypothesis-based and of a more data-driven nature). It is our belief that this will allow researchers to find patterns of reporting that will enable them to say something about the general properties of introspective reports, something no other current method is able to reveal. However, this is not the only methodological possibility afforded by the phenomenon of choice blindness. For example, by extending our basic design to incorporate repeated decisions in longer series of trials, choice blindness can be used to gain insight into the interplay between decision and feedback, choice and report, attitude and outcome. In this vein we have shown how feedback from M-trials can induce *preference change*, and how this bias of future choices relates to the introspective reports given in the experimental situation (*28*).

Thirdly, different mechanisms may underlie choice blindness and change blindness. Given that the current behavioral study was not designed to address the neuro-cognitive underpinnings of either choice or change blindness, it would be premature to offer any speculations whether they indeed are identical. However, as we see it, our experiment is perfectly positioned to bridge the disconnected research areas of choice/intentionality and change blindness, and to create some productive friction between the two. This can be seen clearly by a brief exposition of what intentional choice is supposed to entail. (*29*) write: "*voluntary action implies a subjective experience of the decision and the intention to act… For willed action to be a functional behavior*, *the brain must have a mechanism for matching the consequences of the motor act against the prior intention*" (p. 80, our emphasis, see also *30-32*). But if this is the case, how can it be that the participants in our study often failed to detect the glaring discrepancy between the prior intention and the outcome of their choice? Matching this question with the most common explanations for change blindness offered in the literature does not seem to produce any satisfactory answers. In fact, in our view, given the almost complete lack of reference to mechanisms of decision making and intentionality in the change blindness literature, choice blindness would be an even more remarkable phenomenon if it turned out to be qualitatively identical to change blindness.

For example, the prevalence of choice blindness in our experiment might be due to a failure to sufficiently encode the choice alternatives during the deliberation phase (*33*). But from the perspective of a decision researcher it would amount to a strangely maladaptive decision process not to encode the features that are supposed to be the very basis of the choice, or the gross identity of the two alternatives (at the very least, this should hold for the condition with free viewing time, where the participants themselves set the criteria for when to terminate the deliberation). Another option is that the intentions simply are forgotten during the two second interval when the card is switched. But intentions are not supposed to be instantly forgotten. As (*29*) contend, they are supposed to be the guiding structures behind our actions (and phenomenologically speaking, this is what many people claim them to be), which makes this option equally unattractive to decision theorists. Similar things can be said for the other common explanations for change blindness: that initial representations might be disrupted or overwritten by the feedback (*34*), that change blindness results from a failure to compare pre- and post-change information (*35, 36*), or that explicit change detection is impossible because the representations are in a format inaccessible to consciousness (*37*). They are all viable candidates to explain choice blindness, but also more or less incompatible with popular theories of choice and intentionality. If our task can be seen as a good example of willed action, involving perfectly standard intentions and choices (and currently we can see no reason why this should not be the case), but the outcome of the experiment could be fully explained by the conceptual apparatus of change blindness research, then something would seem to be seriously amiss in current theories of decision making and cognitive control.

Supporting References and Notes

1. R. A. Rensink, *Annu. Rev. Psychol.* **53**, 245 (2002).
2. P. M. Merikle, E. M. Reingold, *J. Exp. Psychol. Gen.* **127**, 304 (1998).
3. P. M. Merikle, M. Daneman, in *The New Cognitive Neuroscience*, M. S. Gazzaniga, Ed. (MIT Press, Cambridge, MA, ed. 2, 2000), pp. 1295-1303.
4. D. R. Shanks, M. F. St. John, **17**, 367 (1994).
5. P. F. Lovibond, D. R. Shanks, *J. Exp. Psychol. Anim. Behav. Processes* **28**, 3 (2002).
6. D. T. Levin, N. Momen, S. B. Drivdahl, D. J. Simons, *Visual Cogn.* **7**, 397 (2000).
7. B. J. Scholl, D. J. Simons, D. T. Levin, in *Thinking About Seeing: Visual Metacognition in Adults and Children*, D. T. Levin, Ed. (MIT Press, Cambridge, MA, 2004), pp. 145-164.
8. K. A. Neuendorf. *The content analysis guidebook*. (Sage, Thousand Oaks, CA, 2002).
9. M. Lombard, J. Snyder-Duch, C. C. Bracken. *Human Com. Res.* **28***,* (2002).
10. S. E. Stemler. *Practical Asses., Res. & Eval.* **9**, 4 (2004).
11. K. O. McGraw, S. P. Wong. *Psych. Methods*. **1**, 4. (1996).
12. J. Uebersax. *Statistical methods for rater agreement*. (Retrieved July 20, 2005, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm).
13. P. E. Shrout, J. L. Fleiss. *Psychological Bulletin* **86**, 420 (1979).
14. K. Krippendorf. *Human Commun. Res.* **30**, (2004).
15. K. Krippendorf. *Content Analysis: an Introduction to Its Methodology* (Sage, Thousand Oaks, CA, 2003).
16. J. L. Fleiss. *Statistical methods for rates and proportions* (2nd ed.) (John Wiley and Sons, NY, 1981).
17. D. J. Simons, D. T. Levin, *Trends Cogn. Sci.* **1**, 261 (1997).
18. P. U. Tse, D. L. Sheinberg, N. K. Logothetis, *Psychol. Sci.* **14**, 91 (2003).
19. A. Noe, *J. Conscious Studies* **9**, 5 (2002).
20. D. J. Simons, D. T. Levin. Psychon. *Bull. & Rev.* **5**. (1998).
21. D. T. Levin, D. J. Simons, B. Angleone, C. F. Chabris. *Brit. J. Psychol.* **93**, 289 (2002).

22. J. Triesch, D. Ballard, M. Hayhoe, B. Sullivan. *J. of. Vision.* **3** (2003).
23. S. Werner, B. Thies. *Visual. Cogn.* **7**. (2000).
24. R. E. Nisbett, T. D. Wilson, *Psychol. Rev.* **84**, 231 (1977).
25. T. D. Wilson, E. Dunn, *Annu. Rev. Psychol.* **55**, 493 (2004).
26. A. I. Jack, A. Roepstorff, *J. Conscious. Studies* **11**, 7 (2004).
27. R. Rorty, *J. Philos* **67**, 399 (1970).
28. B. Tärning, L. Hall, P. Johansson, S. Sikström, A. Olsson. *Unpublished Manuscript.* (2005).
29. A. Sigiru et al. *Nat. Neurosci.* **4**. (2004).
30. M. Ullsperger, D. Y. Cramon, *Cortex* **40**, 593 (2004).
31. K. R. Ridderinkhof, W. P. M. van den Wildenberg, S. J. Segalowitz, C. S. Carter, *Brain Cogn.* **56**, 129 (2004).
32. P. Haggard. *TRENDS in Cog. Sci.* **9**, 6 (2005).
33. J. K. O'Regan, A. Noe. *Behavioral and Brain Sci.* **24**. (2001).
34. M. R. Beck, D. T. Levin. *Percept. & Psychophys.* **65** (2003).
35. S. R. Mitroff, D. J. Simons, D. T. Levin. *Percept. & Psychophys.* **66**, 8 (2004).
36. A. Hollingworth. *J. Exp. Psychol. Hum. Perc. & Perf.* **29** (2003).
37. D. J. Simons, M. Silverman, in. *The Visual Neurosciences,* L. M. Chalupa & J. S. Werner, Eds. (MIT Press, Cambridge, MA, 2004).

PAPER THREE

How Something Can Be Said About Telling More Than We Can Know

On choice blindness and introspection

Petter Johansson, Lars Hall, Sverker Sikström,
Betty Tärning & Andreas Lind

Abstract: The legacy of Nisbett and Wilson's classic article, *Telling More Than We Can Know*: *Verbal Reports on Mental Processes* **(1977), is mixed. It is perhaps the most cited article in the recent history of consciousness studies, yet no empirical research program currently exists that continues the work presented in the article. To remedy this, we have introduced an experimental paradigm we call choice blindness (Johansson, Hall, Sikström, & Olsson, 2005). In the choice blindness paradigm participants fail to notice mismatches between their intended choice and the outcome they are presented with, while nevertheless offering introspectively derived reasons for why they chose the way they did. In this article, we use word-frequency and latent semantic analysis (LSA) to investigate a corpus of introspective reports collected within the choice blindness paradigm. We contrast the introspective reasons given in non-manipulated vs. manipulated trials, but find very few differences between these two groups of reports.**

1. Introduction

Nearly thirty years have passed since the publication of Nisbett and Wilson's seminal article *Telling More Than We Can Know: Verbal Reports on Mental Processes* (1977). Arguably, this article is one of the most widely spread and cited works on the nature of introspection ever to be published. As of May 2006, according to the ISI Web of Science Index, Nisbett and Wilson (N&W) (1977) have been cited an astonishing 2,633 times[1].

No doubt there are many reasons for these extraordinary citation numbers. The comprehensive and accessible review of N&W has long held an attraction for applied researchers dealing with different forms of verbal report. These citations come from the most diverse fields of research: nursing studies, human-computer interface design, demography, psychotherapy, sports psychology, etc.[2] More specifically, N&W has become part of the "checks and balances" of survey and consumer research, as a basic item that must be considered, like experimental demand effects, or the possibility of sampling error (Schwarz & Oyserman 2001).

Yet, despite this, no systematic empirical research program exists that carry on the pioneering work of N&W. It is a piece everybody seems to return to, but hardly anybody tries to improve upon. Buried in the mass of citations one can find a group of articles from the eighties that strove to advance the methodology of N&W (see, e.g., Guerin, 1981; Sabani & Silver, 1981; Morris, 1981; Sprangers, Vandenbrink, Vanheerden, & Hoogstraten, 1987; Quatrone, 1985), but the output from this initiative is all but invisible in the current debate. Despite the prolific work of Wilson himself, who has taken the general idea of lack of introspective access in several new directions (e.g., Wilson, 2002; Wilson & Kraft, 1993; Wilson, Laser, & Stone, 1982; Wilson, Lindsey, & Schooler, 2000), the empirical debate about N&W soon came to a standstill, with multiple layers of inconclusiveness confusing just about everyone involved (as meticulously summarized by White (1988) in his tenth anniversary review of N&W).

Consequently, then, when a scholarly reviewer like Goldman (2004) discusses the epistemic status of introspective reports, he feels the need to address (and refute) the 27-year-old "*challenge*

[1] To put these numbers in perspective it is more than five time as many citations as that gathered by Thomas Nagel's classic essay "*What is it like to be a bat?*" (1974), nearly ten times as many as that given to any of Benjamin Libet's famous articles on the subjective timing of conscious will, and more than twice as many as the combined cites given to *all* the articles that have appeared in the *Journal of Consciousness Studies* and in *Consciousness and Cognition* during the last ten years.

[2] See for example Higuchi and Donald 2002; Jorgensen 1990; Sandberg 2005; Jopling 2001; and Brewer, Linder, Vanraalte, and Vanraalte 1991.

from Nisbett and Wilson," rather than some red-hot contemporary alternative.

It is ironic that the exemplary structure of the original article might be partly to blame for this lack of development. N&W not only tried to show experimentally that "there may be little or no direct access to higher order cognitive processes" (1977, p. 231), but they also tried to present an explicit framework for future studies, and a fully-fledged alternative theory about the origins of introspective reports (thereby taking upon themselves a burden of explanation that most researchers would shun like the plague)[3]. Their basic idea was that the accuracy of introspective reports could be determined by comparing the reports of participants in the experiments to those of a control group who were given a general description of the situation and asked to predict how the participants would react—the so-called *actor-observer* paradigm (Nisbett & Bellows, 1977). If actors consistently gave more accurate reports about the reasons for their behavior than observers did, then this would indicate privileged sources of information underlying these reports. If not, then the position of N&W would be further supported.

Unfortunately, as is shown by the contributions of White (1988) and others (e.g., Gavanski & Hoffman, 1986; Kraut & Lewis, 1982; Wright & Rip, 1981; Wilson & Stone, 1985), it is an exceedingly complex task to unravel all the possible influences on report in an actor-observer paradigm (and this was *before* the whole simulation vs. theory-theory debate got started, which complicates things even further, see Rakover (1983) for an early hint of this debate to come). White (1987) writes:

> In [its] original form the proposal [of N&W] foundered, largely because it is at present untestable. It is difficult if not impossible to ascertain the nature and extent of involvement of "introspective access," whatever that is, in the generation of causal reports, and one cannot assume a straightforward relationship between "introspective access" and report accuracy. In addition, a valid distinction between "process" and "content" or "product" has yet to be pinned

[3] It would seem incumbent on one who takes a position that denies the possibility of introspective access to higher order processes to account for these reports by specifying their source. If it is not direct introspective access to a memory of the processes involved, what is the source of such verbal reports? (Nisbett & Wilson, 1977, p. 232)

> down, despite some attempts to do so. Given these problems, the proposal effectively degenerated into a simpler hypothesis that causal report accuracy cannot be significantly enhanced by information about relevant mental activity between stimulus and response. As we have seen, tests of this hypothesis have so far proved inconclusive. But to continue refining such tests with the aspiration of good internal validity is likely to prove an empty methodological exercise (p. 313).

Thus, with an initially promising but ultimately too narrow conception of how to refine the N&W approach, this line of empirical investigation of introspection ground to a halt. While the disillusioned quote from White might suggest a more general point, that empirical studies of introspection will always be subjected to wildly differing conceptual analyses (of "content", "access", "process", etc.), and that no amount of empirical tinkering is likely to satisfy the proponents of the different consciousness camps (Rorty, 1993), we do not share this gloomy outlook. In our view, the lacuna left in the literature after the collapse of the actor-observer paradigm ought to be seen as a challenge and an invitation. After almost thirty years of intensive research on human cognition, it really *ought* to be possible to improve upon the experimental design of Nisbett and Wilson (1977).

2. Choice Blindness and Introspective Report

In Johansson, Hall, Sikström, and Olsson (2005) we showed that participants may fail to notice mismatches between intention and outcome when deciding which face they prefer the most. In this study participants were shown pairs of pictures of female faces, and were given the task of choosing which face in each pair they found most attractive. In addition, on some trials, immediately after the choice, they were asked to verbally describe the reasons for choosing the way they did (the participants had been informed in advance that we would solicit verbal reports about their intentions during the experiment, but not the specific trials for which this was the case). Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended.

We registered both concurrently and in post-test interviews whether the participants noticed that anything went wrong with their choice. Tallying across all the different conditions of the experiment, no more than 26% of all manipulation trials (M-trials) were exposed. We call this effect *choice blindness* (for details, see Johansson, Hall, Sikström, & Olsson, 2005).

To solicit the verbal reports we simply asked the participants to state *why* they chose they way they did. As Nisbett and Wilson (1977) remarked in the opening lines of their article: "In our daily life we answer many such questions about the cognitive processes underlying our choices, evaluations, judgments and behavior" (p. 231). Thus, for the non-manipulated trials (NM-trials) we expected straightforward answers in reply. For the M-trials, on the other hand, the situation was very different. Here, we asked the participants to describe the reasons behind a choice they did not in fact make. Intuitively, it is difficult to envisage how one would respond to such an anomaly (i.e., we simply do not know what it is like to say why we prefer a particular picture, when we in fact we chose the opposite one). But based on common sense alone, one would suspect that the reports given for NM- and M-trials would differ in many ways.

To explore this contrast, we identified three main psychological dimensions that we believed could be used to differentiate between the reports given in response to NM- and M-trials. These dimensions concerned the *emotionality*, *specificity*, and the *certainty* of the reports. Our reasoning was that participants responding to a manipulated face ought to show less emotional engagement, as this was actually the alternative they did not prefer (*emotionality*); they also ought to make less specific and detailed reports, as no prior reasons have been formulated for this alternative (*specificity*); and they ought to express less certainty about their choice (*certainty*). As detailed in Johansson, Hall, Sikström, and Olsson (2005), we found no differences between the NM- and M-reports on these three dimensions.

In our view, these unexpected commonalities between NM- and M-reports raise many interesting questions about the nature of introspection. However, before any attempts to relate this result to current theories of consciousness are made, we believe the contrastive methodology as such needs to be further discussed and refined.

Debates about the validity and reliability of introspective report often involve lots of back and forth on clinical syndromes where confabulation is likely to be found (such as split-brain, hemineglect, hysterical blindness, or Korsakoff's syndrome, e.g. see Hirstein, 2005). What is striking about these cases is that the patients say things that are severely disconnected from everyday reality. The reports may not always be fantastic or incoherent, but we can easily check the state of the world and conclude that they are implausible as candidate explanations of their behavior. However, as confabulation is defined in contrast to normality, we run into problems when trying to investigate the mechanisms behind the phenomenon. As the confusion and stalemate on Nisbett and Wilson's actor-observer paradigm demonstrates, without the benefit of good contrast cases to work from, discussions of the possibility of confabulatory reporting in normal human populations tend to take on a distressingly nebulous form. The position of N&W was essentially that there are elements of confabulation in all introspective reports, but that these confabulations nevertheless are plausible and reasoned (based on either shared cultural beliefs or idiosyncratic theorizing). But how do we go about testing this interesting proposition, if we cannot even determine what a "genuine" introspective report should look like?

It is our hope that the analysis of introspective reports in our choice-blindness paradigm can contribute toward the goal of establishing a better grip on what constitutes truthful and confabulatory report, and to discern interesting patterns of responding along this dimension with respect to both individual variation and the context of choice.

In Johansson, Hall, Sikström, and Olsson (2005), to compare and contrast the NM- and M-conditions we used blind independent raters to evaluate each of the reports (thus following the natural instinct of experimental psychologists to ground any exploratory measurements by the concept of interrater agreement). But this is not the only way to conduct such an investigation. An obvious weakness of relying on naïve raters to refine the categories used is that they might fail to discern possible differences in the material that could have been revealed by expert analysis. In addition, on the flip side, there is a problem of potential bias in our original choice of categories. Who are *we* to decide what constraints that can be made on the potential contrasts between the NM- and the M-reports?

Thus, in this article, using a new corpus of introspective reports, we present two additional approaches to the same task. Firstly, we carry out an expert-driven linguistic analysis based on word-frequency counts. This analysis covers a great range of linguistic markers known to be important for contrasting different text corpuses, and functions as a complementary top-down way of capturing and recreating the psychological dimensions used in Johansson, Hall, Sikström, and Olsson (2005) (see description above). But while these dimensions are bound to be a reflection of the folk-psychological invariance of everyday life (i.e., everybody has experienced differing degrees of uncertainty and emotionality, etc.), we should be open to the possibility that a computational cognitive perspective might settle on far less intuitive contrasts as being the most productive for analyzing this type of material. To this end, as a more exploratory and data-driven approach, we introduce a novel implementation of Latent Semantic Analysis (LSA). As LSA creates a multidimensional semantic space using very few theoretical assumptions, it is perfectly suited to investigate possible similarities and differences between the NM- and M-reports that cannot easily be captured with the standard toolkit of linguistic and psychological analysis.

3. The Corpus of Reports

The corpus of introspective reports used for our analysis was collected in a recent study extending our previous choice blindness results (Hall et al., in prep.). As in Johansson, Hall, Sikström, and Olsson (2005), participants in this study were shown pairs of pictures of female faces, and were asked to choose which face in each pair they found most attractive. We constructed the face pairs in order to vary the discrepancy of attractiveness within each pair, while an attempt was made to keep similarity constant at an intermediate level (i.e. clearly different, but not drastically so, see Hall et al, in prep.).

Each participant completed a series of fifteen face-pairs, with four seconds of deliberation time given for each choice. As in the previous study, six of the pairs were designated as verbal report-pairs, and any three of these six were in turn manipulated for each participant. Eighty participants (49 female) took part in the study (mean age 24.1, SD 4.1), which gives a total of 480 reports collected.

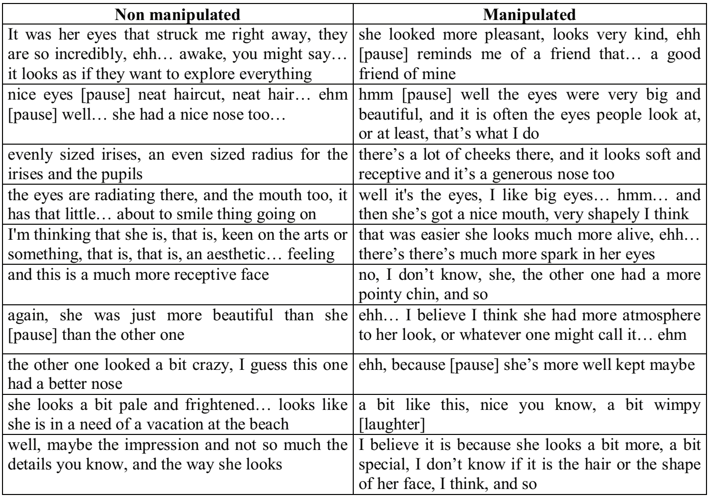
The collection of introspective reports is rich and varied. For the reader to be able to get a descriptive feel for the contents of the reports, Table 1 shows an illustrative selection of statements from both the NM- and M-trials.

To find out the opinion of the participants about the study, we conducted a semi-structured post-test interview. The interview sessions revealed that a great majority of participants felt that the given task was interesting, and that four seconds was enough time to make a meaningful choice (however, there was also a great range and natural variability within the reports, with both self-assured enthusiasm, and concerned caution at times).

The overall detection rate for the manipulated trials was roughly equivalent to our prior results, with 27.5% of the trials detected (for details, see Hall, Johansson, et al., in prep.). Adjusting for detections left 414 reports, and for technical reasons (mishap with the recorder, indecipherable talk, etc.) another 23 were omitted, which leaves 228 NM- and 163 M-reports for the final analysis.

In addition, the study was divided into two different conditions for the introspective reports. The first condition mirrored our previous setup, where we simply asked the participants to state the reasons for choosing the way they did. Here, interaction with the experimenter was kept at an absolute minimum, and no attempts were made to further prompt the participants once they spontaneously seceded in their talk. In the second condition, the same question was posed, but the experimenter encouraged the participants to elaborate their answers up to one full minute of talking time. This was done both by the use of positive non-verbal signals, such as nodding and smiling, and by their linguistic equivalents (such as saying "yes, yes"), and by interjecting simple follow-up questions (such as "what's more?", or "what else did you think of?"). The reason we included the second condition was to see whether longer reports would produce a clearer differentiation between

NM- and M-trials.[4] The reports elicited in the first condition are referred to as short reports and reports from the second condition are referred to as long reports. The average length of the reports was 20 words for the short ones and 97 words for the long ones. All reports were recorded digitally, and later transcribed. The utterances of the experimenter were transcribed, but removed from the corpus before analysis. Pauses, filled hesitations, laughter, and interjections are included in the corpus, but were not counted as words when establishing relative word frequencies between the reports. The final number of reports included in the analysis calculated by condition was 111 (NM-short), 117 (NM-long), 81 (M-short), and 82 (M-long).



Non manipulated	Manipulated
It was her eyes that struck me right away, they are so incredibly, ehh… awake, you might say… it looks as if they want to explore everything	she looked more pleasant, looks very kind, ehh [pause] reminds me of a friend that… a good friend of mine
nice eyes [pause] neat haircut, neat hair… ehm [pause] well… she had a nice nose too…	hmm [pause] well the eyes were very big and beautiful, and it is often the eyes people look at, or at least, that's what I do
evenly sized irises, an even sized radius for the irises and the pupils	there's a lot of cheeks there, and it looks soft and receptive and it's a generous nose too
the eyes are radiating there, and the mouth too, it has that little… about to smile thing going on	well it's the eyes, I like big eyes… hmm… and then she's got a nice mouth, very shapely I think
I'm thinking that she is, that is, keen on the arts or something, that is, that is, an aesthetic… feeling	that was easier she looks much more alive, ehh… there's there's much more spark in her eyes
and this is a much more receptive face	no, I don't know, she, the other one had a more pointy chin, and so
again, she was just more beautiful than she [pause] than the other one	ehh… I believe I think she had more atmosphere to her look, or whatever one might call it… ehm
the other one looked a bit crazy, I guess this one had a better nose	ehh, because [pause] she's more well kept maybe
she looks a bit pale and frightened… looks like she is in a need of a vacation at the beach	a bit like this, nice you know, a bit wimpy [laughter]
well, maybe the impression and not so much the details you know, and the way she looks	I believe it is because she looks a bit more, a bit special, I don't know if it is the hair or the shape of her face, I think, and so

Table 1. Extracts from the NM- and M-reports. The statements were chosen to display the range of responses present in the corpus, with examples taken from reports both high and low on one or more of the dimensions *specificity*, *emotionality,* and *complexity*. The extracts are taken from both the short and the long reports, with a rough matching on the three previously mentioned dimensions being made across the NM- and M-columns.

[4] This can be read both in the sense that the inclusion of more words in the study would increase the statistical power of the analysis, and that potentially confabulatory elements would be more prominent, making a possible contrast between the two types of report more vivid. It should be noted that this condition also served a role in the second focus of the study, which was to investigate whether choice might influence preference change (see Hall, Johansson, et al, in prep.).

4. Comparative Linguistic Analysis

In linguistics, research is often concerned with examining structural differences between different corpora of spoken or written text. Typical examples include comparing different stages in the language development of children (Durán, Malvern, Richards, & Chipere, 2004), contrasting spoken and written text (Biber, 1988), or attempting to authenticate all the works named as Shakespeare's (Elliot & Valenza, to appear).

The methods used to establish such contrasts are diverse, but they all strive to find distinctive markers, a linguistic "fingerprint" that says something interesting about the text under study (Biber, 1988; Labov, 1972). When investigating psychological aspects of language use, emphasis is normally placed on contextual factors influencing the situation, such as the relative status between the speakers, the conversational demands inherent in the situation, and obviously the history and personality of the speakers involved (Brown & Yule, 1983; Norrby, 2004). But the pitfalls of this type of qualitative content analysis are well known (Krippendorff, 1980), and any form of interpretative approach becomes increasingly laborious and ungainly as the amount of text increases.

However, an accumulating body of evidence suggests that a great number of factors can be discerned by analyzing the overall frequency of words used in a text, even if it means ignoring the actual content of the sentences produced. Pennebaker and co-workers have developed a method to differentiate between two (or more) corpora by systematically counting the words used (Pennebaker, Mehl, & Niederhoffer, 2003). They have built a large-scale database consisting of weighted and validated categories, such as words related to cognition ("cause", "know"), emotion ("happy", "bitter"), space ("around", "above"), as well as standard linguistic types (articles, prepositions, pronouns). This database has then been implemented in a specialized program called Linguistic Inquiry and Word Counting (LIWC), which is capable of sifting and sorting all the words from a particular text into the above-mentioned categories, thereby creating a linguistic profile of the text under study (Pennebaker, Francis, & Booth, 2001). Using LIWC, they have managed to establish telling differences between texts for such diverse areas as suicidal and non-suicidal poets (Stirman & Pennebaker, 2001), Internet chat rooms the weeks before and after the death of Lady Diana (Stone & Pennebaker, 2002), and language change over the life span (Pennebaker & Stone, 2003).

While issues of translation from Swedish to English barred us from using the LIWC program on our corpus of reports, we were able to implement our own version of the same methodology using a combination of commercial programs (CLAN), and homemade scripts written to solve specific problems during the analysis. The basic procedure then, for most of our measures, was that we identified different types of words and categories of interest, and then established their relative frequency in the material. These relative frequencies (the occurrence of the target category divided by the total number of words for each report) are the main unit used when comparing NM- and M-reports. Unless otherwise stated, the statistic used is Mann-Whitney U-test. A non-paired non-parametric test is used as there is an unequal amount of NM and M trials (due to the removal of detected M-trials), and because most of the variables did not follow a normal distribution curve.

As we stressed in the introduction, the analysis performed in this article is largely exploratory. Choice blindness is a new experimental paradigm, and the best we have been able to get from the research literature is guiding hunches and intriguing leads about what factors should go into the analysis. Thus, the categorization of the results below should not be read as carving deep metaphysical divisions, but rather as an attempt at pedagogical clustering to highlight interesting patterns for the reader.

In the presentation the English translations always appear in italics, and the original Swedish sentences or words appear in the following parentheses. Unless specifically mentioned, all presented comparisons between the NM- and M-reports below include both the short and the long condition. For ease of reference we have included a summarizing table at the end of the section, with detailed numbers for all the measures used (see Table 2).

4.1. Uncertainty

The most obvious contrast to make between the NM- and M-reports concerns the degree of certainty expressed by the participants in their reports. In Johansson, Hall, Sikström, and Olsson (2005), our blind raters felt that this was the easiest dimension to discern, and the one most firmly represented in the material. But this is not something peculiar to our particular corpus. The study of certainty has a long history in contrastive linguistics. It has, for example, been argued that female language often contains more words expressing uncertainty, and that it often is more imprecise and non-committal (Lakoff, 1975). The argument is centered on distinctive markers of uncertainty, such as *sort of, I think,* and *you know*, a class of expressions and words called *hedges* (Holmes, 1995, 1997). Similarly, differences in expressed certainty have been found between different social classes, academic disciplines (Varttala, 2001), and even within the same research fields when different languages are used (Vold, 2006). An issue closely related to hedging is *epistemic modality*, which concerns how we express our level of commitment to the propositions we produce. What is examined here is not just uncertainty but the full spectrum of security in a statement—from *I know it's true* to *I guess it's true* (Frawley, 1992).

However, when looking for markers of uncertainty, it is important to note that there are several different aspects of uncertainty at play in our material. Firstly, the participants might be unsure about the decision, indicating that they do not know *why* they chose one face over the other. Secondly, they might be hesitant about the act of speaking itself, simply not knowing what to say next. Thirdly, the participant might feel uncomfortable and cautious about the situation as such, sensing that something is wrong, but just not knowing what it is. Following the literature, we created several different measures to try to capture a very broad sense of uncertainty.

For the epistemic aspect of uncertainty, we set up a list of words and phrases with an established function as hedges: *perhaps* (kanske), *you know* (ju), *I suppose* (väl), *probably* (nog), *don't know* (vet inte), *I think* (tror jag). These particular hedges were chosen because they were highly frequent in our corpus, thus making them good candidates for being able to differentiate between the NM- and M-reports. For the calculations we used a composite measure based on the relative frequency of the class of hedges compared to all words for each report. This was done both as a group and for each individual word or phrase. However, we found no statistical differences between the NM- and the M-reports for epistemic uncertainty, neither for the short nor for the long condition.[5]

As a measure of hesitance, we used both filled and unfilled pauses in the speech. An unfilled pause was defined as a silence within sentences lasting for more than 0.5 seconds. The filled pauses consisted of vocalizations filling the gaps between words, as well as words without content or function in the linguistic context (e.g. um, er, *na* (nä), *yeah* (jo)). As such, pauses have been hypothesized to be an instrument for the speaker to manage his or her own cognitive and communicative processes—i.e. to buy time while planning what to say next (Alwood, 1998). Given the intuitive assumption about the choice blindness situation that the entirety of the verbal explanation is constructed on the spot, an analysis of pauses seemed to us to be a very promising measure to use. But as was the case for the epistemic markers, we found no significant differences between NM- and M-reports for the amount of pauses used. As an independent category of filler activity, we also calculated the amount of laughter present in the NM- and M-reports (the hypothesis being that laughter can function as a signal of nervousness, distress, or surprise, see Glenn, 2003), but again, we found no significant differences with respect to laughter between the NM- and M-reports.

In summary, using several different linguistic measures, we found no evidence of differences in expressed uncertainty between the NM- and M-reports.

4.2. Specificity

The crux of the dilemma in the choice blindness paradigm is what sources the participants draw upon, or what mechanisms they

[5] We also calculated this contrast using a more inclusive set of words related to uncertainty, but no significant effects could be found with this measure either (see table 2).

use, when delivering their introspective reports in the NM- and the M-trials. Again, the common-sense assumption would be that the NM-reports reflect the actual intention that resulted from the deliberation phase (this being a natural source of information when stating their reason, such that the participants can divulge whatever level of detail they deem appropriate). For the NM-reports, as these are given in response to an outcome the participant did *not* choose, it is altogether unclear what the basis of the report is, and if indeed we should predict that the participant would have anything at all to say.

However, we found no significant differences with respect to absolute word count. Another way to measure specificity is to count the number of unique words (that is, words only used once, in total 761 in the corpus). This division cuts through all word classes as a measure of relative rarity. But no significant differences between the NM- and M-reports were found on this measure either.

An alternative and more complex measure of the specificity of the statements is to look at the entire report, and determine to what extent the participants actually are talking about the choice they have made, and how much they are just (plain) talking. Following the guidelines of Brown and Yule (1983) we cleaned the corpus from all parts of the reports that did not involve a chain of reasoning, or listing of details that the participants thought had influenced their choice, thus separating the text into *content* and *metalingual comments*. Overall, around 50% of all transcribed text was classified as not strictly being about the choice, but this number did not differ significantly between the NM- and M-reports. Thus, the participants seemed to have as much content to report on regardless on whether they talked about a choice they had actually made, or responded to a mismatched outcome in a choice blindness trial.

Yet another way to get a grip on potential differences in specificity is to focus only on the amount of nouns used. This class of words contains all the details and features that surface in the participants' descriptions, such as "the face", "the eyes", "the hair". For the short reports we found no differences, but for the long reports there was a significant difference (Mann-Whitney $U = 3859$, $p = 0.019 < 0.05$) between the NM- and M-reports. The direction of the difference was also in line with the initial hypothesis—i.e., the relative frequency of nouns was higher in the NM-reports (mean = 0.089) than in the M-reports (mean = 0.078).

This is an interesting finding that raises the question of whether the dimension of specificity can also be discerned *within* the class of nouns, or if it lies more in the use of nouns as such. To investigate this, we listed all nouns from the material, and let two independent raters divide them into two groups.[6] One category concerned *specific* nouns, with words describing detailed features of the presented faces, such as *eyebrows* (ögonbryn), *haircut* (frisyr), *earrings* (örhängen), and *smile* (leende). The other category contained more *general* nouns, like *face* (ansikte), *picture* (bilden), *girl* (tjej), and *shape* (form). We tested these two categories separately, for both the short and the long reports, but with this measurement we found no significant differences for any of the conditions or categories.[7]

As a final test for specificity, we examined the generality of the noun difference, by running the same kind of analysis on the corpus of verbal reports collected in the Johansson, Hall, Sikström, and Olsson (2005) study. Using the current analysis as a template, we created a corresponding list of nouns for that material, divided into specific and general nouns (again, using two independent raters). Here, we found no significant differences between the NM- and M-reports, neither for nouns as a word class, nor for the division between specific and general nouns.

In summary, as in Johansson, Hall, Sikström, and Olsson (2005), we could not find any significant differences on the gross features of specificity for the NM- and M-reports, but for the more precise measurement of number of nouns used, a significant difference could be found for the long reports only (however, this difference

[6] The interrater reliability for this task was very high, and for the few instances where the raters differed in their opinion, the disagreement was solved through further discussion among the raters. A similar procedure was used for all instances of independent rating mentioned in this article.

[7] If we glean at the mean value, we can see that there are 'unsignificantly' more specific *and* non-specific nouns in the long NM reports; a difference that in combination creates the overall significant difference for nouns. So the difference does consist in the NM reports being more specific per se, just that more descriptive nouns in general are used.

could not be pinpointed to the use of more specific nouns, and it did not generalize to our previous corpus of reports).

4.3. Emotionality

The level of emotional engagement (whether positive or negative) is another of the obvious candidates for analysis that we investigated in Johansson, Hall, Sikström, and Olsson (2005). It is an obvious dimension to investigate because it is supposed to be present in the task (i.e. we would simply not have been so keen to compare the NM- and M-reports if it concerned a choice that the participants believed to be pointless). It is also a dimension that ought to be resistant to the manipulation, because even if the original reasons and intentions of the participants might be lost in the murky depths of their minds, at least they ought to still *prefer* the face they originally chose, and thereby show a more positive attitude toward the images in the NM-trials.

When looking for differences in emotionality, we proceeded in a similar fashion as we did with specificity. First we measured the amount of adjectives, having identified them as the word class with most relevance for the levels of emotional engagement that the participants displayed in their reports. For this overall measurement, we found no significant differences between the NM- and M-reports. Then, using two independent raters, we created two subdivisions of adjectives: positive words—*beautiful* (vacker), *happy* (glad), *cute* (söt)—and negative words—*tired* (trött), *boring* (tråkig), *sad* (sorgsen). For the negative adjectives we found no significant differences, but for the positive ones we found a significant difference for the long reports only (Mann-Whitney $U = 3837.5$, $p = 0.0164 < 0.05$), such that there were more positive adjectives in the NM-reports (with the mean = 0.0474 for NM-reports, and the mean = 0.0367 for the M-reports). As with the previous finding for nouns, this difference did not generalize to the corpus collected in Johansson, Hall, Sikström, and Olsson (2005).

As we discussed above, this is a difference that makes a lot of sense in terms of the situation. Participants ought to show a more positive attitude toward the face they actually chose. But as emotionality is such a salient feature of the choice situation, both at the time of the original deliberation and at the time when the verbal report is given, this finding is not the best option for a clean indicator of the distinction between truthful and confabulatory report. This is so because for the full minute of speech delivered in the long reports, there is ample time for the original preference to assert itself, and for the participants in both the NM- and M-trials to *add* features to their report (while this concerns only minute differences, on average the NM-trials ought to build up in a more positive direction than the M-trials would).

In summary, we found a significant difference in positive emotional adjectives used between the NM- and M-reports for the long condition only. However, this difference is of unclear origin, and we could not replicate the finding in the corpus used in our earlier study.

4.4. Deceit

One line of inquiry that could potentially be of great use in contrasting and understanding the NM- and M-reports is research on the linguistic markers of deceit and lying. Even though the (possibly) confabulatory reports given by the participants in the M-trials obviously cannot be equated with an act of conscious and deliberate lying, it could be argued that the two situations share many features; most importantly, that something with no grounding in actual experience is being talked about.

The idea that statements derived from memory of an actual experience differ in content and quality from statements based on invention or fantasy has been the basis for several different methods for detecting deceit, such as criteria-based content analysis (CBCA, originally developed as a technique to determine the credibility of children's witness testimonials, Steller & Köhnken, 1989), and Reality Monitoring (RM, originally a paradigm for studying false memory characteristics, see Johnson & Raye, 1981). More recently, with the advent of powerful computers for large-scale data mining, this concept has blossomed into a separate field of automated deception detection (for overview, see Zhou, Burgoon, et al., 2004a).

As an example of this development, Newman, Pennebaker, Berry, and Richards (2003) used Pennebaker's LIWC to distinguish between lies and truthful reports. In one of the conditions in this study, the participants were instructed to provide true and false descriptions of people they really liked or disliked. The deceptive element was thus to describe a person they really liked as if their feeling was very negative (and similarly, in the opposite direction for someone they disliked). Across all conditions, the software detected several persistent features that reliably predicted which statements were true and which were false. The variables they found to be primarily responsible for the differentiation were that liars used fewer first person references, fewer third person pronouns, fewer exclusive words ("except", "but", "without"), and more negative emotion words.

We were able to look directly at several of the critical variables identified by Newman, Pennebaker, Berry, and Richards (2003). In particular, as there ought to be no real sense of "me" having preferred the outcome presented to the participants in the M-trials, we deemed the "cognitive distance" effect for first person references to be a good candidate to be represented in our material (what also has been called *verbal immediacy*, see Zhou, Burgoon, et al., 2004b). We indexed all first person pronouns *I* (jag), *me* (mig), *mine* (min) in the corpus. These words were highly frequent, with *I* being the most frequent of all (with 1,406 instances in total). We also counted all third person pronouns as an index of third person references (dominated by *she/her* (hon, henne), but also including *it* (den, det), *they* (dom) and *her* (hennes). In our corpus, we were unable to find an equivalent to the "exclusive words" category used by Newman, Pennebaker, Berry and Richards (2003).

However, despite verbal immediacy being a reliable predictor of deception, we found no significant differences for first person vs. third person pronouns between the NM- and M-reports (or for the negatively toned adjectives, as reported in the previous section on emotionality).

In summary, we found no significant differences between the NM- and M-reports by measuring them against linguistic markers of deceit.

4.5. Complexity

Another more theoretically driven perspective on the potential for the detection of markers of deceit in linguistic corpora is the assumption that lying is a more cognitively taxing activity than truthful report. Here, what is normally seen as markers of deceit should rather be seen as markers of *cognitive load* (Vrij, Fisher, Mann, & Leal, 2006). Evidence for this position comes from the fact that when training interrogators to detect deceit, it is more effective to instruct them to look for signs of the subjects "thinking hard," rather than signs that they seem nervous or emotional (Vrij, 2004). But theories of cognitive load are obviously not confined to the field of deceit detection. It is one of the most widespread and most commonly used concepts in the cognitive sciences (and central to the whole idea of consciousness as a limited channel process, see Baars, 1997; Dehaene & Naccache, 2001). Translated to the task of introspective reporting in our choice-blindness paradigm, it lies close at hand to hypothesize that the participants in the M-trial would show a marked reduction in the *complexity* of the language used, as their resources ought to be taxed to a greater degree by the demands of reporting the reasons behind a choice they did not in fact make. For example, Butler, Egloff, et al., (2003) have reported a result close to this when showing that participants tend to use less complex language in a conversation task when they are simultaneously required to suppress a negative emotion.

The first and most simple way of measuring the complexity of NM- and M-reports is to look at the word length (e.g. Zhou, Burgoon, et al., 2004b), where longer words are believed to require more effort to use. We calculated the mean word length for each of the four conditions, but we found no significant differences on this measure (short mean NM = 4.3 M = 4.4, long mean NM = 5.2, M = 5.3).

Two more advanced approaches to sentence complexity are the sibling concepts of *lexical density* and *lexical diversity*. What is meant by lexical density is essentially how informationally "compact" a text is (measured as the number of content words in relation to the number of grammatical or function words, Halliday, 1985; Ure, 1977).[8] Lexical diversity, on the other hand, captures the uniqueness of the words used, i.e. how many different words there are in relation to the totality of the text (Malvern, Richards, Chipere, & Durán, 2004).

In our corpus we measured lexical density as the percentage of content words (nouns, verbs, adjectives, and adverbs) to all the words in a given text (content words plus grammatical words). Based on the hypothesized increase in cognitive load in the M-reports, it follows that they ought to have a lower lexical density. As we had already found differences in the base frequency of nouns and (positive) adjectives, it seemed as if this measure was a good candidate to reveal differences on a more structural level as well. However, we found no significant differences in lexical density between the NM- and M-reports.[9]

To measure lexical diversity we used the D algorithm from the CLAN software suite.[10] The sampling procedure used when calculating the measure D needs a minimum of 50 words for each entry. Given this constraint, we were only able to determine the lexical diversity for the long reports. But as was the case with lexical density, we found no significant differences between NM- and M reports for this measure.

One interesting possibility here is that potential differences between the NM- and M-reports on lexical diversity are masked by a *priming effect*, such that novel words introduced during the NM-trials remain in an active state, and carry over to the (suppos-

[8] A standard example of differing lexical density is written and spoken text, in which written text normally has a larger proportion of content words (Halliday, 1985).

[9] It is interesting to note that there were differences between the *short* and the *long* reports, with the short reports being significantly more dense ($p = 0.007$).

[10] Intuitively, we can sense that there is a difference between for example the lush and varied style of Isabel Allende, and the stern and compact prose of Hemingway. But how to best capture such differences quantitatively is somewhat disputed (Malvern, Richards, Chipere, & Durán, 2004). The standard way of measuring diversity is type/token ratio (TTR) (i.e., the sentence "I am what I am" has three types and five tokens). However, as is now known, this method has certain statistical weaknesses. The best current alternative is the measure D, which we use here (Durán, Malvern, Richards, & Chipere, 2004). So far, D has mainly been used to study language development, but it has also been put to some use in comparative studies on specific language impairment (SLI) and second language acquisition (Malvern, Richards, Chipere & Durán, 2004).

edly content-free) M-trials (i.e., this would be another way of stating the hypothesis that the cognitive load of the M-trials would reduce the complexity of the language used). We investigated this hypothesis by looking at the order in which the verbal reports were given for each participant, and calculating the number of new nouns introduced relative to what the participants had said before. However, the number of new nouns introduced did not significantly differ between the two conditions.

A final approach to unraveling the complexity of the introspective reports given by our participants would be to look at the *tense* and *themes* (i.e. structures of reasoning) they use to describe the chosen picture. There is no uniform way in which the participants use tense when explaining the reasons for the choices they have made. Sometimes they speak in the present tense, focusing on details in the preferred face ("she has such a round little nose"). But they can also refer back to the time of decision ("I liked her eyes and mouth"), or use comparative statements, in both past and present tense ("she had darker hair and she has so clear and pretty eyes"). The reasoning behind this measurement is again based on the concept of cognitive load. With less resources to spare in the M-trials, features of the current situation ought to have a greater impact on the report given (this could also be stated more intuitively as the idea that participants ought to refer more to present tense in the M-reports because they have no reason to refer back to from the moment the decision was made).

To investigate tense and themes we first created a basic index of all words related to tense (is/was, has/had, etc.), but we found no differences between the NM- and M-reports using this measurement. Next, to get a more precise measurement, we used the division between content parts and metalingual comments discussed in section 4.2 above, and indexed the content part of the reports into either *positive reasons* for choosing the way they did, or *comparative reasons* why they preferred one face over the other one. Then these two categories were in turn divided into past and present tense.[11]

[11] As it is very hard to divide spoken text into discrete chunks we did not count the relative number of statements in past or present tense, but only measured whether it occurred or not in each verbal report. The mean values presented in table 2 are to be understood as the number of reports in which *some* parts were in past or present tense (and Why- or Comparative statements).

But again, we found no significant differences between the NM- and M-reports.

	Short NM	**Short M**	***p***	**Long NM**	**Long M**	***p***
6 Words marking uncertainty	0.060 (0.007)	0.065 (0.010)	0.999	0.036 (0.002)	0.039 (0.002)	0.438
Extended measure of uncertainty	0.096 (0.009)	0.101 (0.011)	0.728	0.071 (0.007)	0.077 (0.008)	0.105
Filled pauses	0.047 (0.006)	0.047 (0.006)	0.452	0.048 (0.003)	0.054 (0.004)	0.228
Unfilled pauses	0.018 (0.005)	0.036 (0.015)	0.135	0.032 (0.003)	0.041 (0.005)	0.262
Laughter	0.010 (0.003)	0.019 (0.005)	0.343	0.008 (0.001)	0.010 (0.002)	0.590
Metalingual comments	0.493 (0.032)	0.544 (0.035)	0.296	0.543 (0.017)	0.544 (0.019)	0.745
Nouns	0.091 (0.009)	0.078 (0.009)	0.348	0.089 (0.003)	0.078 (0.004)	0.019
Specific nouns	0.055 (0.008)	0.043 (0.008)	0.320	0.052 (0.003)	0.046 (0.003)	0.178
Non-specific nouns	0.029 (0.005)	0.022 (0.004)	0.604	0.025 (0.002)	0.020 (0.002)	0.103
Nouns (Johansson et al 2005)	0.105 (0.009)	0.113 (0.011)	0.543	*	*	*
Specific nouns (Johansson et al 2005)	0.056 (0.007)	0.069 (0.011)	0.310	*	*	*
Non-specific nouns (Johansson et al 2005)	0.049 (0.007)	0.044 (0.006)	0.543	*	*	*
Adjectives	0.121 (0.009)	0.121 (0.009)	0.155	0.115 (0.004)	0.105 (0.004)	0.284
Adjectives (positive)	0.054 (0.008)	0.047 (0.007)	0.853	0.047 (0.003)	0.037 (0.003)	0.016
Adjectives (negative)	0.004 (0.002)	0.009 (0.003)	0.472	0.012 (0.001)	0.013 (0.002)	0.729
Adjectives (Johansson et al 2005)	0.116 (0.008)	0.108 (0.008)	0.511	*	*	*
Adjectives (positive) (Johansson et al 2005)	0.094 (0.008)	0.087 (0.008)	0.557	*	*	*
Adjectives (negative) (Johansson et al 2005)	0.022 (0.003)	0.021 (0.003)	0.849	*	*	*
Word length	4.288 (0.745)	4.403 (0.916)	0.339	5.215 (0.614)	5.265 (0.579)	0.557
Lexical density	0.331 (0.014)	0.317 (0.013)	0.453	0.303 (0.005)	0.290 (0.006)	0.130
Lexical diversity	*	*	*	D=53.015 (2.308)	D=49.528 (2.089)	0.369
Priming. new nouns	1.144 (0.111)	1.086 (0.140)	0.483	3.701 (0.211)	3.744 (0.322)	0.424
WHY present	0.225 (0.040)	0.173 (0.042)	0.376	0.838 (0.034)	0.927 (0.029)	0.062
WHY past	0.162 (0.035)	0.086 (0.031)	0.125	0.393 (0.045)	0.317 (0.052)	0.274
COMP present	0.108 (0.030)	0.037 (0.021)	0.071	0.137 (0.032)	0.085 (0.031)	0.267
COMP past	0.315 (0.044)	0.407 (0.055)	0.190	0.453 (0.046)	0.585 (0.055)	0.066
First-person pronouns	0.071 (0.007)	0.081 (0.009)	0.676	0.047 (0.003)	0.053 (0.004)	0.191
Third-person pronouns	0.123 (0.006)	0.116 (0.009)	0.800	0.108 (0.003)	0.112 (0.004)	0.646
Tense. verbforms present	0.107 (0.008)	0.115 (0.013)	0.599	0.104 (0.004)	0.111 (0.004)	0.281
Tense. verbforms past	0.080 (0.008)	0.077 (0.009)	0.746	0.053 (0.003)	0.051 (0.004)	0.612

Table 2. Summary of the results from the contrastive linguistic analysis. The number in parentheses is the standard deviation of the mean. The shaded sections represent the significant differences found between the NM- and M-reports.

In summary, using the concept of cognitive load and language complexity, we were unable to find any significant differences between the NM- and M-reports.

5. Latent Semantic Analysis

The differences we have found so far between the NM- and M-reports, using a whole battery of potential linguistic markers identified from the literature, have been small and very hard to interpret. But it is easy to envision that our search has been overly constrained by a limited theoretical outlook, or that is has been hampered because we lack crucial knowledge about some aspects of the relevant field of linguistics. Also, it could be argued that the "atomic" approach of word-frequency analysis is ill suited to capture differences of a more abstract semantic nature.

To allay these worries we decided to approach the corpus using a complementary bottom-up approach. Recent advances in computational cognitive analysis have opened up the intriguing possibility of quantifying semantics by applying advanced statistical techniques to huge text corpuses. These techniques are based on the postulate that semantics is carried by co-occurrences—that is, if two words frequently occur together in the same context (e.g. *love-like*), then this will be taken as evidence that the words have a similar meaning, or lie near each other in the semantic space.

Semantic spaces that include the semantic relationships of words from an entire language can be constructed using a method called *Latent Semantic Analysis* (LSA) (Landauer & Dumais, 1997). The way LSA works is that first a table for co-occurrence is created, where rows represent unique words and columns represent the contexts (e.g., sentences, paragraphs, or documents) from which the words are taken. Words that co-occur in the same context are marked with their frequency, otherwise a zero is marked. This table is then rescaled to account for differences in frequency by the logarithm of the frequency, and by dividing by the entropy across context. Finally, a semantic space is constructed by applying a mathematical technique called singular value decomposition (SVD) to reduce the large number of contexts to a moderate number of dimensions, all the while maintaining the maximal possible amount of the original information. The dimensions obtained correspond to the psychological concept of features that describe semantic entities in the words. The quality of the resulting semantic space can then be verified by applying a synonym test (and this information can in turn be used to further optimize the technique after optimization the number of dimensions left is typically found to be in the order of a few hundred, see, e.g., Landauer and Dumais 1997)

Semantic spaces have successfully been applied in a number of linguistic and memory settings. Semantic spaces based on LSA have been shown to perform comparably to students in multiple-choice vocabulary tests, and in textbook final exams (Landauer, Foltz, & Laham, 1998). By measuring coherence, semantic spaces have also been used to predict human comprehension equally well as sophisticated psycholinguistic analysis (Landauer, Laham, & Foltz, 2003). In the domain of information search, LSA has also been found to improve retrieval by 10–30% compared to standard retrieval measure techniques (Dumais, 1994). Similarly, LSA has been used successfully to differentiate documents. As an example, Landauer, Laham, and Derr (2004) used sophisticated projection techniques to visualize scientific articles from different fields by projecting the high-dimensional semantic space to two-dimensional maps.

Taken together, these results indicate that LSA is an extremely promising tool for analyzing the semantic aspects of texts. However, currently there are no methods available for quantitatively comparing the semantics of two different classes of verbal report data, and for visualizing the results in a clear and convincing manner. Here, we introduce a new implementation of LSA specifically developed for this purpose, and apply it to the corpus of reports collected in the choice-blindness paradigm.

5.1. Method

As a base corpus, the Stockholm-Umeå Corpus (SUC, Ejerhed & Källgren, 1997) consisting of one million Swedish words was selected. This corpus is balanced according to genre, following the principles used in the Brown and LOB corpora. Infomap (http://infomap.stanford.edu/), a natural language software that implements LSA, was then used to create a semantic space. Context was defined as 15 words before, or after, the current word in the present document. Following initial testing, we settled for a space consisting of 150 dimensions. The length of the vector describing each word was normalized to one.

The semantic spaces were processed in LSALAB,[12] a program specifically developed by one of authors to analyze semantic spaces. Each verbal justification for choosing a particular face was summarized to one point in the semantic space by averaging the semantic location of all the words included in the statement. To be sure that the semantic representations were stable and reliable, we included only the 4,152 most common words from the SUC corpus (words with lower frequency were ignored).

As we are unaware of any other studies applying statistical methods to compare conditions within a semantic space, we developed the following technique to handle the issue. The semantic point describing each condition (e.g., NM- and M-trials) was summarized as the average of the semantic points of all statements included in the condition. The Euclidean distance was then used as a measure of distance between the conditions (μ_1). After this, a bootstrap technique was applied to estimate the variability in distance. Statements were randomly placed in either of the two conditions (using the same number of trials), and the distance was calculated. To achieve a reliable estimate this was repeated for 200 trials. A one-tailed t-test was calculated by subtracting the mean distance of the random trials (μ_0) from the distance between the conditions (μ_1), and this was then divided by the estimated standard deviation of distance for the random trials (σ).

As LSA deals with a multidimensional space, graphic illustration is essential to understanding the results. However, the plotting of such high-dimensionality spaces is problematic, as it typically requires a projection to only two dimensions.[13] To deal with this

[12] For details, see www.lucs.lu.se/people/sverker.sikstrom/lsalab_intro.html

[13] Landauer et al. (2004) argue for the visualization of semantic spaces as a powerful tool for understanding, viewing, and exploring semantic data. They were able to plot the semantic representation of more than 16,000 scientific articles from *Proceedings of the National Academy of Sciences* (PNAS) using the Gobi software (Swayne, Cook, & Buja, 1998). In this case, dimensionality reduction was conducted by a combination of mathematical tools and visual inspection.

problem we propose the use of a two-dimensional separation-typicality map. These maps are obtained by the following method.

We base both of the axes on the Euclidean distance, where the x-axis represent *separation* and the y-axis *typicality*. Separation on the x-axis is based on a distance measure that maximally differentiates between the two conditions. The natural choice is the distance from a statement to the prototype of one of the conditions. To separate condition 1 and 2, we simply plot the difference in distance (DID), which is the Euclidean distance from a statement to the prototype of condition 1 minus the Euclidean distance from same statement to the prototype of condition 2. However, the DID measure is subject to a statistical artifact. Because the instances are compared with the prototype, the separation between the conditions will be inflated. This artifact can be removed by a bootstrapping technique whereby the statements are randomly placed into the two conditions. To obtain sufficient statistics we repeated this 200 times. We then subtracted the average DID obtained from the random samplings from the DID of each statement. The resulting corrected DID value, which we label DID´, is free from statistical artifacts, so that the expected value of the separation from randomly generated populations is zero. DID´ is a measure of the separation between the conditions. If the two prototypes are identical then the value will always be zero.

On the y-axis we plot the typicality of the statements. This is simply the Euclidean distance between the statement and the prototype of all statements. This measure is bounded between zero and two in our semantic representation. A zero value indicates that the statement is identical to the prototype of all statements.

Although this procedure was successful in separating and finding sub-cluster in the data space, it has several problematic aspects to it. Firstly, the choice of a projection to a low-dimensional space can be made in an almost infinite number of ways, so the resulting conclusion becomes highly dependent on this choice. Secondly, while choosing projections, statistical artifacts may bias the separation between conditions so they appear to be larger than they actually are. For example, separating two conditions sampled from the same population for 100 dimensions will results in an expected value of 5 statistically different dimensions due to chance. Plotting these dimensions will amount to a form of data fishing, and the separations will only be statistical artifacts. Thirdly, when using the Landauer et al. (2004) methodology, the axes on the plot are not immediately available for interpretation.

A value of two indicates that the statement is maximally different from the prototype. A value of one indicates that the statement is unrelated to prototype (i.e., the expected value of a randomly generated statement). Most often the values will fall in the range 0 to 1, where low values indicate statements that are typical and high values indicate semantically atypical statements.

5.2. Results

There was no statistical difference in semantic content between the NM- and M-reports (t (388) = –0.91; p = 0.82 > 0.05). Thus, the result of the statistical analysis of the semantic space indicates that the participants justify their choice using the same semantic content for both the NM- and the M-trials.

To visualize these results we use the separation-typicality map described above. Figure 1 plots the separation between the statements on the x-axis, and the typicality (low values indicate high typicality) on the y-axis. Each dot represents a NM-report, and each cross an M-report. The large dot and cross represent the average values over all statements in each condition. The curves in the lower part of the graph are the densities of the respective condition. As is apparent from figure 1, the overlap between the NM- and M-reports is almost complete. The typicality of statements ranges from approximately 0.35 (high typicality) to 1.2 (low typicality), with a mean around 0.6, where 1 represents statements that are unrelated to the prototype of all statements.

While LSA is a well-established and powerful technique for building semantic spaces, it has never before been used for significance testing in this type of contrastive methodology. Thus, a possible reason for the lack of separation between NM- and M-reports could be that our proposed method is not sensitive enough to differentiate between the two conditions. In order to minimize this risk, it is important to demonstrate that the method indeed can detect meaningful differences under conditions where those differences are likely to emerge. To demonstrate this we ran the same kind of differentiation analysis using the gender of the participants as an input variable.[14] In contrast to what was the case for the NM- and M-reports, we found a highly significant difference between the introspective reports given by men and women (t (388) = 2.98; p = 0.002 < 0.05). Thus, it can be shown that the method we used is sufficiently sensitive to distinguish between the semantic content of statements produced by two contrast groups.

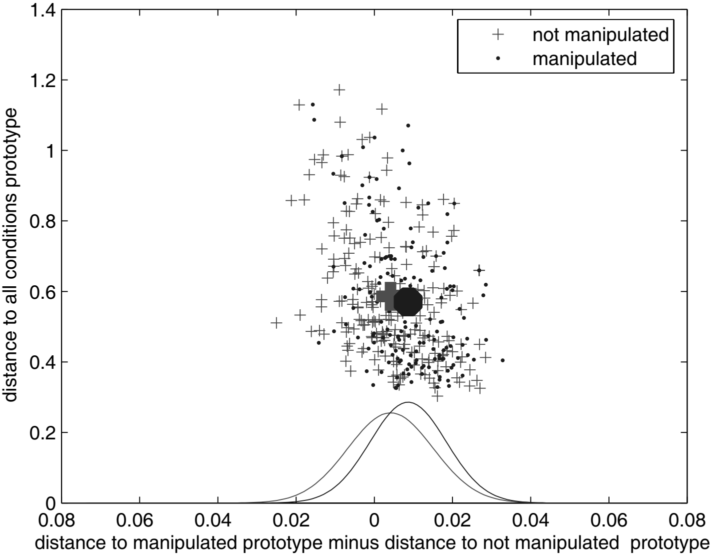


Figure 1. Separation-Typicality map for the NM- and M-reports. The y-axis plots the semantic distance to the prototype of all conditions as a function of difference in distance (DID) on the x-axis. Each cross and dot represent a manipulated or not manipulated statement respectively. The large dot and cross represent the average values over all statements in each condition. The expected distance between two randomly semantic locations is one, and the maximally possible distance is two, compared with the distance to all conditions prototype on the y-axis. The difference in distance between the conditions on the axis represents the difference between the conditions, so that if the two conditions' prototypes were identical then the distance would be zero. The two curves in the lower part of the graph show the density of statements for the two conditions.

[14] As the experiment collected very few personality variables, the age-spread of the participating student population was limited, and each image-pair contained too few reports to be entered into the analysis, gender emerged as the best candidate variable to work with.

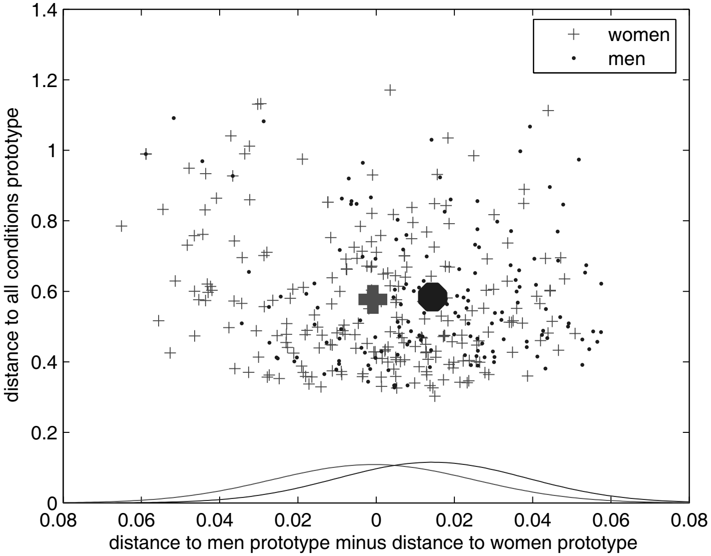


Figure 2. Separation-Typicality map for the Female and Male reports. Each cross and dot represent a male or female statement respectively. In all other regards, the figure is the same as Figure 1.

Figure 2 shows a separation-typicality map for female and male reports. As is evident from the figure, a clear separation between the two groups of report can be found. This is reflected in the large variability on the x-axis (compare with the low variability in Figure 1 showing the NM- and M-reports).

However, given that the statements made by men and woman differ in their semantic content, the question remains how best to characterize these differences. To try to capture the differences found, we listed all the words in the constructed semantic space that had the closest semantic location to the male and female prototypes respectively. These associates may be conceived of as a type of "keywords" that summarize something about all statements in the conditions. The first thing to notice is that the keywords for statements made by men and women are highly similar (e.g., see the first two columns in Table 3). The first seven associates are identical (with the exception of a single flip of the ordering). This demonstrates that the similarities between the male and female reports are great, yet we are still able to discern the subtle differences residing in the material. This point further strengthens the inference that had there been any semantic content differences between the NM- and M-reports, it is highly likely that our method would have picked them up.

As explained above, one of the virtues of LSA is that it embodies very few assumptions about the nature of the subject under study. In this way, there is a greatly diminished risk that the results are contaminated by either common-sense intuitions, or the particular theoretical outlook of the experimenters. To identify more clearly the difference in the reports made by males and females, we subtracted the male and female prototype vector from each other. The closest semantic associates to this vector are listed in column four in Table 3.

Associates		Differences	
Men	**Women**	**Men**	**Women**
It	It	Analysis	Hers
But	But	Interested	She
Not	Not	True	Face
I	Be	Democratic	Foot
Be	I	Doubt	And
To	To	Name	Down
Just	Just	Know	Fine
Have	Only	Pull	Hand
As	Have	Think	Out
Know	She	It	Skirt
Him	Accomplish	Hardly	Mouth
What	He	Starting-point	Kiss
Become	And	What	Sit
And	Become	Up	Arm

Table 3. The closest semantic associates to male and female prototypes. The first two columns show the fourteen closest semantic associates to statements made by men and women respectively, starting with the closest associates. The last two columns show semantic associates to the vector describing the difference between the two prototypes, where the column labeled men is the closest associate to the vector men minus women, and the column labeled women the vector women minus men. It is important to stress that none of the words displayed in the columns actually needs to be represented in the choice-blindness corpus (i.e., no male participant need ever have used the word "democratic" when describing why they choose one face over the other). In this case the associates instead come from the million word SUC corpus used to anchor the semantic space. All words in the table are translated from Swedish to English.

For females, out of the approximately four thousand possible words in our semantic space, the two highest associates were the female pronouns *her* and *she*. A large proportion of the remaining associates were body parts (*face*, *foot*, *hand*, *mouth*, *arm*). For males, the closest associates to this vector are shown in column three in Table 3. These associates tend to be more abstract (*analysis*, *democratic*), and revolve around the theme of knowing (*true*, *doubt*, *know*, *think*, *hardly*).

It is not possible to provide an exact summary of the semantic differences in associations between the gender specific statements, as there is no fully transparent mapping from the dimensions captured by LSA onto everyday concepts. But, as reported above, the outcome suggests a separation along a dimension of concreteness-abstractness, and into themes of knowing vs. body parts, and in the particular use of personal pronouns. However, these results are far from the end-point of the inquiry. They should rather be seen as a kind of *data-driven hypothesis generators*. For validation and translation into everyday concepts, additional work would be required that attempted to further quantify and test the identified dimensions.[15]

[15] For example, if we compare these results to the more than twenty significant differences that we found between the male and female reports using the categories previously reported for the word-frequency analysis, the complementary, but also partially overlapping, character of the LSA analysis becomes obvious. Regarding the female LSA associates for the female pronouns, a match can be found with the word-frequency analysis that indicated a higher degree of use of personal pronouns by women (short reports, Mann-Whitney U = 3447, $p = 0.026 < 0.05$). The LSA differences between females and males for the dimension of concreteness-abstractness also seems to be reflected in the word-frequency analysis, where we found females to be using more specific nouns (long reports, Mann-Whitney U = 3678.5, $p = 0.004 < 0.05$), and more non-specific nouns (short reports, Mann-Whitney U = 3379, $p = 0.016 < 0.05$). However, the knowing-theme from the LSA analysis does not seem to have an immediate counterpart among the epistemic measures used in the word-frequency analysis, and there are also several other significant differences from the contrastive linguistic analysis that did not emerge in our global LSA comparison (i.e. word length, high-low frequency words, present tense, pauses, prepositions, conjunctions, etc.).

6. How something can be said about telling more than we can know

It probably has not escaped the reader that this article has an unusual format for the presentation of the main results—i.e., we treat the failure to find distinguishing markers between the NM- and M-reports as an equally important finding as any of the potential differences found. We are aware that, from a textbook perspective, this logic is clearly flawed (i.e., with standard significance testing, the null hypothesis cannot be confirmed, only rejected), yet we cannot escape the conclusion that the overall pattern of findings indicates that the NM- and M-reports are surprisingly similar. To really appreciate this null-hypothesis blasphemy, we must go back to the sentiments we had, and the predictions we made (including those of our colleagues) before we conducted our first choice blindness experiment. Tentatively stating a hypothesis at this time, we predicted not just differences between the NM- and M-reports, but *huge* differences. As it stands now, not a single difference found in the current corpus would survive a standard Bonferroni correction.[16] This can be compared to the strong pattern of differences between male and female reports, which we were able to discern both with word-frequency analysis and with LSA.

Another way of framing the subtlety of the possible differences between NM- and M-reports existing in our material is by comparing them to the literature on automatic lie detection we briefly referenced in section 4.4. For detection of lies based on linguistic cues only, Newman, Pennebaker, Berry, and Richards (2003) and others (e.g. Zhou, Burgoon, et al., 2004b), have shown that

[16] Bonferroni correction is a commonly adhered-to guideline when doing exploratory research, a safeguard to prevent results arising from chance fluctuations when multiple tests of statistical significance are done on the same data set. It states that for multiple comparisons the p level should be equal to alpha-level/ number of observations (0.05/N). As more than 30 variables are measured in this article (for both short and long reports), even if not adhered to strictly, none of the seemingly significant results are firm enough to remain after a Bonferroni correction. The reason we did not include this calculation in the results section is that we prefer to err on the side of including non-existent differences, rather than the other way around. As this type of contrast has not been made before, we believe it to be of great importance to grasp every straw there is to generate further hypotheses about how the NM- and M-reports might relate to each other.

prediction models can be built that capture general differences between truths and lies using very similar dimensions to those measured in this article (i.e., certainty, emotionality, complexity, etc.). It is a telling point that the differences in the deceit literature are so small that untrained human observers basically predict at chance level, while finely calibrated software only reaches levels of predictability of about 60–65% (Newman, Pennebaker, Berry, & Richards, 2003). However, for the contrast between the NM- and M-reports in our material it is at present doubtful whether *any* such model can be built.

We believe we have conducted a thorough and revealing investigation of the introspective reports collected so far in our choice blindness paradigm. Including the analysis done in Johansson, Hall, Sikström, and Olsson (2005), we have used three complementary types of measurement (psychological rating, word-frequency analysis, and LSA), and all three have come out with very similar results.

But obviously, this is just a starting point. For example, the fact that the two tentative differences we found in the material (on specificity and emotionality) only could be found for the long reports might suggest that one should look more closely at *time* as a factor in future studies. However, the remarkable thing from our perspective is that the debate about the nature and validity of introspection is still conducted at a level where the introduction of a contrast class between (potentially) genuine, and (potentially) confabulatory reports seemingly can tell us a great deal about what introspection amounts to. A simple contrastive methodology is often derided by researchers from more mature fields of science, but it can still function as a springboard for other more penetrating approaches (as has been the case with lesion studies, studies of individual differences, cross-cultural comparisons, etc.). In this sense, Nisbett and Wilson (1977) were far ahead of their times when they introduced a methodology that required the experimenters to *know and control* the causes of the behavior of the participants for it to work. N&W strove admirably for ecological validity in their experiments, but 30 years later (notwithstanding the wet dreams of some marketers and retailers) this is still something the behavioral sciences are incapable of doing, save in the most circumscribed and controlled environments.

In this vein, it can be seen that the most famous of the experiments of N&W, the department-store stocking experiment, involved a rather strange and contrived task (e.g., Kraut & Lewis, 1982; Kellogg, 1982). It seems to us, had only the experimenters had a better grasp of what influenced the choice behavior of normal consumers, they would not have given them the artificial choice between *identical* stockings, but rather something that would have involved actual products of varying quality.

While we do not want to pretend that the task we have used here (and in Johansson, Hall, Sikström, & Olsson, 2005) involves an important choice for the participants, it is a very straightforward one, reflecting a type of judgment that people often make in their daily lives (and undoubtedly, many people have strong opinions about facial attractiveness). It has the virtue of being a simple and vivid manipulation that does not place the same exorbitant demands on the experimenters to be able to secretly influence the decision process of the participants. Like the hypothetical "intuition pumps" so often employed in debates about consciousness and introspection (see Dennett, 1991), this is an experiment where it is child's play to twiddle with the knobs (parameters) of the setup, and produce potentially very interesting results (by changing properties of the stimuli, deliberation time, questions asked, context of choice, personality variables, etc.).

Philosophically speaking, our choice blindness paradigm is of the same breed as the N&W experiments. We believe it to be an improvement over N&W in many regards, but at this point there are many opportunities for interpretations open for the wily theoretician. For example, the fact that we can hardly find any differences between the NM- and M-reports could stem from the participants actually reporting the very same thing in both conditions—i.e. the intentions they had for making their actual choice. But this is a strained interpretation to make when one sees how good the match between the given reports and the presented faces often are, and it creates outright absurdities in those cases where the reports refer to unique features of the manipulated face (e.g. "I chose her because I love blondes", when in fact the dark-haired one was the chosen one). Conversely, when differences between NM- and M-reports are found, they could have been created at the time of actual reporting, rather than being inherited from the deliberation phase. As we discussed briefly in the section on emotionality, the interaction of prior preferences and the outcome of the choice could possibly lead the two classes of reports to diverge (i.e., in the M-trials the participants are reacting to a face they did not prefer, no wonder then they are not exuberant about it now). It is also clear that the simplification we have made in this article, where we keep the analysis of the verbal reports more or less separate from the basic choice blindness effect, cannot be maintained in the long run. If we are to fully understand introspection, then we should be prepared to explain the whole architecture of a decision-making system in which one might fail to notice mismatches between intention and outcome, but yet give perfectly intelligible verbal reports in response to the manipulated choice. However, as we said in the introduction, we have an upbeat outlook on the prospects for development in this field. It seems to us that the simple contrast at the heart of our choice blindness paradigm is perfectly poised to be used in the kind of triangulation of subjective reports, behavioral responses, and brain imaging data that Roepstorff and Jack (2004) identify as the best route for future studies of introspection and consciousness to take.

In conclusion, we want to emphasize the potential of our method over the particularities of the results in this article. When Nisbett and Wilson (1977) took upon themselves not only to introduce a new experimental paradigm, but to formulate a theory of introspection in sharp contrast to the prevailing view, they set the research community up for a high-strung showdown, not unlike the archetypal movie scene where the protagonists suddenly find themselves locked at mutual gunpoint (the so-called "Mexican standoff"), and where the smallest twitch of the pen inevitably will release a hail of deadly arguments. In our minds, far too little has been said about telling more than we can know, for us to have reached a point where a standoff is called for. Instead, it is our hope that the effort put forward here will lead to a renewed interest in experimental approaches to the study of verbal report and introspection.[17]

[17] If we allow the visionary movie industry to lead our way, in contrast to the spaghetti westerns of the 70s, the B-movie thrillers of the 80s, and the bloody mayhem of Tarantino in the 90s, the recent movie *Munich* (2005), contains a scene with a friendly resolution of an incredibly tense Mexican standoff.

References

Allwood, J. (1998). *Some frequency based differences between spoken and written Swedish*. Paper presented at the XVIth Scandinavian Conference of Linguistics, Department of Linguistics, University of Turku.

Baars, B. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford: Oxford University Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Brewer, B. W., Linder, D. E., Vanraalte, J. L., & Vanraalte, N. S. (1991). Peak performance and the perils of retrospective introspection. *Journal of Sport & Exercise Psychology, 13*(3), 227–238.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Burgoon, J. K., Blair, J. P., Tiantian, Q., & Nunamaker, J., Jay. F. (2003). *Detecting deception through linguistic analysis*. Proceedings of the Symposium on Intelligence and Security Informatics (ISI-2003).

Burrows, J. F. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*, 267–287.

Butler, E. A., Egloff, B., Wilhelm, F. H., Smith, N. C., Erickson, E. A., & Gross, J. J. (2003). The social consequences of expressive suppression. *Emotion, 3*, 48–67.

Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown & Company.

Dixon, J. A., & Foster, D. H. (1997). Gender and hedging: From sex differences to situated practice. *Journal of Psycholinguistic Research, 26*(1), 89–107.

Dumais, S. T. (1994). *Latent semantic indexing (LSI) and TREC-2*. Paper presented at the Second Text Retrieval Conference (TREC2).

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.

Ejerhed, E., & Källgren, G. (1997). Stockholm Umeå corpus (Version 1.0): Department of Linguistics, Umeå.

Elliot, W., & Valenza, R. (to appear). Two tough nuts to crack: Did Shakespeare write the *Shakespeare* portions of Sir Thomas More and Edward III? In *Shakespeare yearbook*.

Frawley, W. (1992). *Linguistic semantics*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gavanski, I., & Hoffman, C. (1986). Assessing influences on one's own judgments: Is there greater accuracy for either subjectively important or objectively influential variables. *Social Psychology Quarterly, 49*(1), 33–44.

Glenn, P. (2003). *Laughter in interaction*. Cambridge: Cambridge University Press.

Goldman, A. I. (2004). Epistemology and the evidential status of introspective reports. *Journal of Consciousness Studies, 11*(7–8), 1–16.

Guerin, B., & Innes, J. M. (1981). Awareness of cognitive-processes—replications and revisions. *Journal of General Psychology, 104*(2), 173–189.

Hall, L., Johansson, P., Tärning, B., Sikström, S. (in preparation). Choice Blindness and Preference Change: Lund University Cognitive Science.

Halliday, M. A. K. (1985). *Spoken and written language*. Deakin: Deakin University Press.

Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science, 6*, 293–340.

Higuchi, K. A. S., & Donald, J. G. (2002). Thinking processes used by nurses in clinical decision making. *Journal of Nursing Education, 41*(4), 145–153.

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., et al. (2005). Distinguishing deceptive from non-deceptive speech. *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005).*

Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.

Holmes, J. (1995). *Women, men and politeness*. London: Longman.

Holmes, J. (1997). Women, language and identity. *Journal of Sociolinguistics, 1*(2), 195–223.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*(5745), 116–119.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67–85.

Jopling, D. A. (2001). Placebo insight: The rationality of insight-oriented psychotherapy. *Journal of Clinical Psychology, 57*(1), 19–36.

Jorgensen, A. H. (1990). Thinking-aloud in user interface design: A method promoting cognitive ergonomics. *Ergonomics, 33*(4), 501–507.

Kellogg, R. T. (1982). When can we introspect accurately about mental processes. *Memory & Cognition, 10*(2), 141–144.

Kraut, R. E., & Lewis, S. H. (1982). Person perception and self-awareness knowledge of influences on ones own judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

Kraut, R. E., & Lewis, S. H. (1982). Person perception and self-awareness: Knowledge of influences on ones own judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford: Blackwell.

Lakoff, R. (1975). *Language and woman's place*. New York: Harper Colophon Books.

Landauer, K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *PNAS, 101*, 5214–5219.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2)*, 211–240.

Landauer, T. K., Foltz, P., & Laham, D. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Morris, P. E. (1981). The cognitive psychology of self-reports. In C. Antaki (Ed.). *The psychology of ordinary explanations of social behaviour*. London: Academic Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665–675.

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology, 35*(9), 613–624.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Norrby, C. (2004). *Så gör vi när vi pratar med varandra* (2nd ed.). Lund: Studentlitteratur.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC2001. New Jersey: Lawrence Erlbaum Associates.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*. *54*, 547–577.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use across the lifespan. *Journal of Personality and Social Psychology, 85*(2), 291–301.

Quattrone, G. A. (1985). On the congruity between internal states and action. *Psychological Bulletin, 98*(1), 3–40.

Rakover, S. S. (1983). Hypothesizing from introspections—a model for the role of mental entities in psychological explanation. *Journal for the Theory of Social Behaviour, 13*(2), 211–230.

Rakover, S. S. (1983). Hypothesizing from introspections: A model for the role of mental entities in psychological explanation. *Journal for the Theory of Social Behaviour, 13*(2), 211–230.

Richards, J. M. (2004). The cognitive consequences of concealing feelings. *Current Directions in Psychological Science, 13*(4), 131–134.

Roepstorff, A., & Jack, A. I. (2004). Trust or interaction? Editorial introduction. *Journal of Consciousness Studies, 11*(7–8), V–XXII.

Rorty, R. (1993). Holism, intrinsicality, and the ambition of transcendence. In B. Dahlbom (Ed.), *Dennett and his critics: Demystifying mind* (pp. 184–202). Oxford: Blackwell.

Sabini, J., & Silver, M. (1981). Introspection and causal accounts. *Journal of Personality and Social Psychology, 40*(1), 171–179.

Sandberg, J. (2005). The influence of network mortality experience on nonnumeric response concerning expected family size: Evidence from a Nepalese mountain village. *Demography, 42*(4), 737–756.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*(2), 127–160.

Shuy, R. W. (1998). *The language of confession, interrogation, and deception*. Thousand Oaks, CA: Sage.

Sprangers, M., Vandenbrink, W., Vanheerden, J., & Hoogstraten, J. (1987). A constructive replication of white alleged refutation of Nisbett and Wilson and of Bem: limitations on verbal reports of internal events. *Journal of Experimental Social Psychology, 23*(4), 302–310.

Steller, M., & Köhnken, G. (1989). Criteria-based content analysis. In D. C. Raskin (Ed.). *Psychological methods in criminal investigation and evidence*. New York: Springer-Verlag. 217–245.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine, 63*, 517–522.

Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology, 24*, 172–182.

Swayne, D. F., Cook, D., & Buja, A. (1998). Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics, 7*, 113–130.

Ure, J., & Ellis, J. (Eds.). (1977). *Register in descriptive linguistics and linguistic sociology*. The Hague: Mouton Publishers.

Vartatala, T. (2001). *Hedging in the scientifically oriented discourse. Exploring variation according to discipline and intended audience*. Ph. D. Thesis.

Vold, E. T. (2006). Epistemic modality markers in research articles: A cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics, 16*(1), 61–87.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology, 9*, 159–183.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10*(4), 141–142.

White, P. A. (1987). Causal report accuracy: retrospect and prospect. *Journal of Experimental Social Psychology, 23*(4), 311–315.

White, P. A. (1988). Knowing more about what we can tell: Introspective access and causal report accuracy 10 years later. *British Journal of Psychology, 79*, 13–45.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. London: The Belknapp Press.

Wilson, T. D., & Kraft, D. (1993). Why do I love thee—effects of repeated introspections about a dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin, 19*(4), 409–418.

Wilson, T. D., Laser, P. S., & Stone, J. I. (1982). Judging the predictors of ones own mood: Accuracy and the use of shared theories. *Journal of Experimental Social Psychology, 18*(6), 537–556.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126.

Wilson, T. D., & Stone, J. I. (1985). Limitations of self-knowledge: More on telling more than we can know. In P. Shaver (Ed.), *Review of personality and social psychology: Self, situations, and social behavior* (Vol. 6, pp. 167–185).

Wright, P., & Rip, P. D. (1981). Retrospective reports on the causes of decisions. *Journal of Personality and Social Psychology, 40*(4), 601–614.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., Jay. F., & Nunamaker, J. (2004a). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139–165.

Zhou, L., Burgoon, J. K., Nunamaker, J., Jay. F., & Twitchell, D. P. (2004b). Automatic linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*, 81–106.

PAPER FOUR

Magic at the Marketplace

Lars Hall, Petter Johansson, Betty Tärning,
Thérèse Deutgen & Sverker Sikström

Abstract: We were interested in investigating whether the recently discovered phenomenon of *choice blindness* (Johansson, Hall, Sikström & Olsson 2005) would extend to decisions made in more naturalistic settings, and for the modalities of taste and olfaction. We set up a tasting venue at a local supermarket and invited passerby shoppers to sample two different varieties of jam and tea, and to decide which alternative in each pair they preferred the most. Immediately after the participants had made their choice, we asked them to again sample the chosen alternative, and to verbally explain why they chose they way they did. At this point we secretly switched the contents of the sample containers, so that the outcome of the choice became the opposite of what the participants intended. All in all, no more than a third of the manipulated trials were detected, thus demonstrating considerable levels of choice blindness for the taste and smell of two different consumer goods.

Introduction

In Johansson, Hall, Sikström and Olsson (2005) we demonstrated that participants may fail to notice mismatches between intention and outcome in a simple decision task. In the study we showed the participants pairs of pictures of female faces, and gave them the task of choosing which one they found most attractive. Unknown to the participants, on certain trials, we used a card magic trick to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended. We registered whether the participants noticed that anything went wrong with their choices. Counting across all the conditions of the experiment, no more than 26% of the manipulation trials were detected. We call this effect *choice blindness* (for details, see Johansson, Hall, Sikström & Olsson 2005).

The fact that processing of faces is of great importance in everyday life, stemming both from the evolutionary and social significance of facial recognition and evaluation (Rhodes, 2006; Bruce & Young, 1998; Schwaninger, Carbon & Leder, 2003), suggests to us that choice blindness will generalize widely to other visual stimuli, and even across modalities. However, we cannot rule out the possibility that there is something about the hypothesized "holistic" processing of human faces (e.g. Tanaka & Farah, 1993; Tanaka & Sengco, 1997; but see Gauthier, Curran, Curby, & Collins, 2003) that prevented our participants from properly categorizing and verbalizing the mismatch between their original choice and the manipulated outcome. Moreover, while it is clear that lasting judgments of attractiveness for human faces can be made within a split second (Olson & Marshuetz, 2005; Willis & Todorov, 2006), a span far shorter than the deliberation time we gave the participants in Hall, Sikström and Olsson (2005), it is always possible that a less constrained procedure would have generated a different result.

For these reasons we were interested in investigating whether the phenomenon of choice blindness would extend to choices made in more naturalistic settings. As we see it, consumer choice is a perfect domain in which to test this paradigm. The modern marketplace is an arena where the tug of explicit and implicit influences on the behavior and opinions of consumers is played out in a particularly fierce manner. Recently, psychologist have weighed in heavily on the side of non conscious influences on consumer choice, both as a general framework of analysis (Dijksterhuis, Smith, Van Baaren, & Wigboldus, 2005; Chartrand 2005), and with the discovery of various implicit effects, such as those arising from preference fluency (Novemsky, Dhar, & Schwarz, in press), placebo effects of marketing (Shiv, Carmon & Ariely, 2005; Irmak, Block & Fitzsimons, 2005), name-letter branding (Brendl Chattopadhyay, Pelham, & Carvallo, 2005), and from incidental brand exposure in minimal social interactions (Ferraro, Bettman & Chartrand, in press). Even the age old claim about subliminal influences on choice behavior has been revitalized in recent developments (Winkielman, Berridge & Wilbarger, 2005; Fitzimons, Chartrand & Fitzimons, in press).

But at the same time the marketplace is an arena of remarkable vividness and explicitness, where everything is written on the sleeve (or at least in the barcode) of the products on display. In modern societies people not only have a long history of consumption decisions to fall back upon, they also have en enormous repository of symbolic knowledge about the goods available (comparing the average person today to the most knowledgeable 16th century scientist, they probably ought to be considered as *scholars* of consumer brands and products). But not only this, consumers often have firm opinions about marketing and branding of products as such, and they think and reflect about how these factors influences their own decisions. Thus, one cannot deny that there is validity to traditional forms of consumer surveys based on introspection, and to the methods of multidimensional sensory rating often used by industry researchers (for different perspectives on this debate, see Dijksterhuis, Smith, Van Baaren, & Wigboldus, 2005; Chartrand, 2005; Simonson, 2005; Strack, Werth & Deutsch, 2006; Woodside, 2004; Schwarz, 2003).

To investigate whether choice blindness would extend to modality specific choices between different consumer goods, we set up a sample stand at a local supermarket, where we invited passerby customers to participate in a blind test of two paired varieties of jam and tea. In a pretest, using a locally available assortment of jam and tea, we composed candidate pairings roughly matched on color and consistency, and allowed an independent group of participants to rate the similarity of the two alternatives in each pair. Pilot testing indicated very low levels of detection for the more similar pairs for both jam and tea, so for the main study we included one pair from the middle of the distribution, and to really test the limits of choice blindness, the two most dissimilar pairs from the comparison.[1]

[1] While we strived to create the most dissimilar product pair matching possible within the constraints of our budget, two studies from the US can illustrate the fact that our match-up of local brands only probed a tiny corner of the world market for jam. Wilson & Schooler (1991) investigated the effects of introspecting about reasons for choosing different brands of jam, and their selection of samples was based on a Consumer Report study comparing no less than 45 different types of strawberry jam alone. Similarly, a study by Iyengar & Lepper (2000) investigated whether the amount of choice alternatives would affect subsequent purchase decisions for jam, and this study was conducted at an upscale Californian supermarket which carried more than 300 varieties of jam.

In order to create a convincing covert exchange of the chosen samples, we created two sets of “magical” jars, lidded at both ends, and with a divider inside. These jars thus looked like normal containers, but were designed to hold one variety of jam or tea at each end, and could easily be flipped over to execute a switch (see Figure 1).

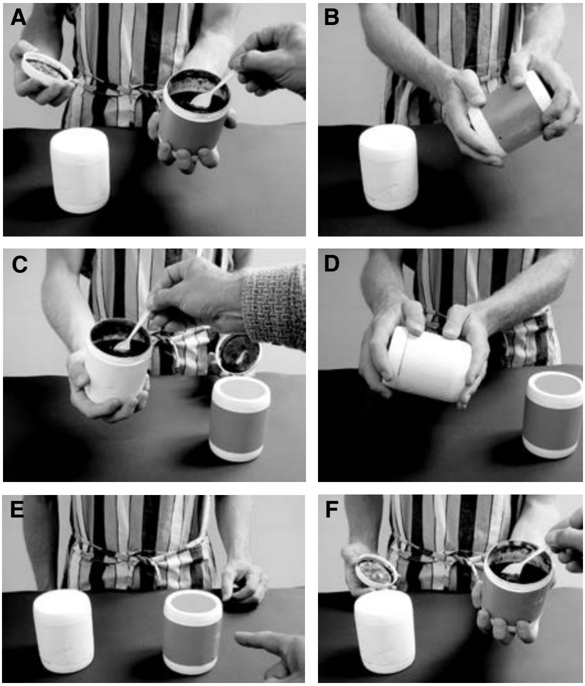


Figure 1. A step-by-step illustration of a manipulated choice trial in the jam condition. **A.** The participants sample the first jam. **B.** The experimenter secures the lid back on and flips the jar upside down whilst putting it back on the table. The jar looks normal, but it is lidded at both ends, and with a divider inside, containing one of the included samples at each end. **C.** The participants sample the second jam. **D.** The experimenter performs the same flipping maneuver for the second “magical” jar. **E.** The participants indicate which jam they prefer. **F.** The participants sample the chosen jam a second time, but since the containers have been flipped they now receive the alternative they did not prefer.

Based on the piloting and our previous studies of choice blindness we expected to find that participants would fail to notice the mismatch in many of the manipulated trials. Given the gap in similarity between the first pair and the other two, we also expected that a higher detection rate would be found for the less similar pairs. As a part of the choice procedure we instructed the participants to rate how much they liked each sampled alternative. We expected to find a relation between the discrepancy of these likeability scores and the level of detection, such that larger rated differences between the two samples would correlate with higher degrees of detection.

Furthermore, we were interested in studying the effect of incentives on the level of choice blindness. To this effect, half of the participants were offered the chosen sample (either a jar of jam or a package of tea) as a gift to bring home after the completion of the study. We expected that the provision of this incentive would motivate the participants further and increase their attention to the decision process (Hertwig & Ortman, 2001; 2003), which in turn would lead to a higher rate of detection for the manipulated gift trials. In addition, our setup permitted us to investigate possible indirect influences of the manipulated choices on subsequent behavior. After the participants hade made their selection we asked them to rate how difficult they felt it was to tell the two samples apart, and how confident they were about the choice they hade just made. Even if a manipulation was not overtly detected, it might still have recognizable effects on these following judgments. We reasoned that the second tasting of the manipulated sample might distort the original memory of the discrepancy of the two options, and that the participants would indicate that they found it more difficult to tell the two samples apart in the manipulated trials than in the control trials. Similarly, we hypothesized that if the participants had any lingering doubts from the experience of the manipulation, this ought to reveal itself as a lowered confidence in the choice.

Experiment

Method

Participants. A total of 180 consumers (118 female) at a supermarket in Lund, Sweden, participated in the study (three participants were removed due to recording problems). The age of the participants ranged from 16 to 80 years (mean=40.2; std=20.0). They were recruited as they passed by a tasting venue we had set up in the store. We presented ourselves as being independent consultants contracted to survey the quality of the jam and tea assortment in the shop. All participants were naïve to the actual purpose of the study. After the study, they gave their written consent to be included in the analysis. The study was approved by the Regional Swedish Ethics Board in Lund.

Material. As stimulus material, we used three pairs of jam and three pairs of tea. The pairs were selected from a pretest in which independent participants rated the similarity of 8 pairs of jam and 7 pairs of tea, for taste and smell respectively. The scale used ranged from 1 (very different) to 10 (very similar). To isolate the dimension of interest (taste for jam, and smell for tea) the pairs were roughly matched with regard to color and consistency. The average rated similarity for the included pairs ranged from 4.05 to 6.55 for the jam, and from 3.25 to 6.4 for the tea. As pilot testing indicated very low levels of detection for the more similar pairs for both tea and jam, in the main study we chose to include one pair from the middle of the distribution, and the two most dissimilar pairs from the match up. For jam the chosen pairs were *Black Currant* vs. *Blueberry* (mean=5.1; std=2.5), *Ginger vs. Lime* (mean=4.1; std=2.2), and *Cinnamon Apple vs. Grapefruit* (mean=4.0; std=2.7). For tea the chosen pairs were *Apple Pie vs. Honey* (mean=4.7; std=2.4), *Caramel & Cream vs. Cinnamon* (mean=3.6; std=1.8), *Pernod (Anise/Liquorice) vs. Mango* (mean=3.25; std=2.5).

For the choice manipulation, two small containers were glued together bottom-to-bottom, creating a single jar with two independent sections with separate screw-on lids. A paper wrapping was then applied over the mid-section to complete the illusion of a single unbroken container (color coded in red and blue to make it easier to distinguish among the alternatives). In each trial two of these containers were used, filled with either two different sorts of jam or tea (i.e. each jar was a mirror of the other one, expect for the colored label, and which compartment that was facing upwards at the beginning of the experiment).

The verbal reports of the participants, and their interaction with the experimenters during the study, was digitally recorded in MP3 format.

Procedure. The experiment took place at a local supermarket. We recruited the participants by asking them whether they were willing to take part in a "quality control" test of the jam and tea assortment at the store. At the start of the experiment we informed the participants the test was to be done with the product labels removed, focusing only on the taste of the jam, and the smell of the tea, and that they should indicate which sample they preferred the most in each pair. We also asked for their consent to be audio recorded during the experiment. All participants agreed to the recording. In addition, half of the participants were told that they would receive either a package of tea or a jar of jam as a gift at the completion of the test (the specific gift depended upon their choice in the designated manipulation trial, see description below). Two experimenters were present during the test. Experimenter 1 asked questions, took notes, and managed the recording device, while Experimenter 2 conducted the preference test. For each participant, either the tea or the jam condition was manipulated. The order of presentation, the type of manipulation, and which pairs of jam or tea that was included was randomized for each participant.

In a manipulated trial, the participants were presented with the two prepared jars. After tasting a spoon of jam from the first jar, or taking in the smell of the tea, they were asked to indicate how much they liked the sample on a 10-point scale from "not at all good" to "very good". While Experimenter 1 solicited the preference judgment, and interacted with the participants, Experimenter 2 screwed the lid back on the container that was used, and surreptitiously turned it upside down. After the participants had indicated how much they preferred the first option, they were offered the second sample, and once again rated how much they liked it. As with the first sample, Experimenter 2 covertly flipped the jar upside down while returning it to the table. Immediately after the participants completed their second rating, we then asked the them to sample the preferred option a second time (for those trials in which equal ratings had been given, the participants were forced to deliberate again, and pick one alternative), and to verbally motivate why they liked this jam or tea better than the other one. As both jars had been turned upside during the prior sampling, and the upper compartments thus were reversed, the participants were now given the opposite of what they actually chose. After the participants had finished the third (manipulated) sample, and explained their choice, they were asked to indicate on a 10-point scale how difficult they felt it was to discriminate between the two alternatives (from "very difficult" to "very easy"). Finally, they were asked to indicate on a 10 point scale how confident they were in their choice (from "very unsure" to "very certain").

The same procedure was used for the non-manipulated (NM) trials, with the only difference that in the NM trials no jars were turned. For each pair of jam or tea tested, 30 M and 30 NM trials were collected.

After the participant had completed both a jam and a tea pairing, we asked them whether they had felt that anything was odd or unusual with the setup of the tasting session, or with the sampled alternatives. This was done to see whether the participants would spontaneously indicate that some form of change or mismatch had taken place. After this, the participants were debriefed about the true nature of the experiment, and they were again given an opportunity to indicate weather they had registered or suspected that we had manipulated the choice alternatives. The experiment lasted between five to ten minutes.

We used three different criteria of detection for the manipulation trials. A manipulated trial was classified as a *concurrent* detection if the participants voiced any concerns immediately after tasting or smelling the manipulated jam or tea. A manipulation trial was classified as a *retrospective* detection if the participants at the end of the experiment (either before or after the debriefing) claimed to have noticed the manipulation. Finally, as a more implicit form of detection, even if the participants did not consciously report that something went wrong with their choice, we registered whether they for any reason described the taste or the smell of the chosen sample as somehow being different the second time around (i.e. tasting/smelling stronger, weaker, sweeter, etc.). We call this final category a *sensory-change* detection.

Results

To briefly recapitulate, we had the following hypotheses. Firstly, and most importantly, that most of the manipulations would not be detected. Secondly, that the detection rate would increase with the degree of dissimilarity between the sample pairs. Thirdly, that the greater the discrepancy in rated attractiveness between the two samples would be, the more likely it would be that the manipulation would be detected. Fourthly, that the incentive of receiving the chosen tea or jam as a gift would increase the detection rate. Fifthly, that the perceived ease of discrimination between the stimulus options would be greater for the non-manipulated than the manipulated trials. And finally, that the rated confidence in the choice would be higher for the non-manipulated compared to the manipulated trials.

Counting across all pairs, no more than 14.4% of the jam trials and 13.8% of the tea trials were detected concurrently. An additional 6.2% of the jam and 6.9% of the tea trials were detected retrospectively, and 12.4% of the jam and 11.5% of the tea trials were registered as a sensory-change type of detection. In total, 33.3% of the manipulated jam trials, and 32.2% of the manipulated tea trials were detected.

We found significant differences in detection rate between the most and least similar jam pairs ($\chi^2_1 = 4.16$, $p = 0.04<0.05$ and between the most and the least similar tea pairs ($\chi^2_1 = 8.85$, $p = 0.003<0.05$) (see Figure. 2). Contrary to our prediction, the participants that received the gift incentive had a lower detection rate in the tea condition ($\chi^2_1 = 7.12$, $p = 0.007<0.05$), but no difference was found for the jam condition. We found a correlation between the rated discrepancy of attractiveness within a pair, and detection frequency for jam (F(2, 84) = 5.08, p = 0.03<0.05), but not for tea. There was a difference in the perceived ease of distinguishing between the two samples when comparing the NM-trials and the non-detected M-trials for tea (F(2, 147) = 4.06, p = 0.046<0.05), but not for jam. There were no differences in rated confidence between the NM trials and the non-detected M-trials for either jam or tea.

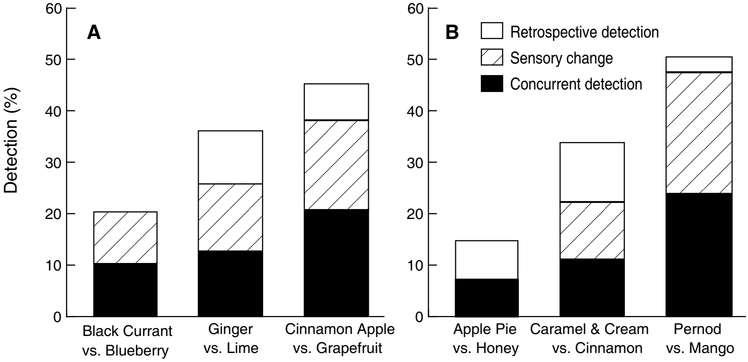


Figure 2. The data is divided into detection type (retrospect detection, sensory change, concurrent detection), pair (three stimuli pairs), and modality (**A** for jam and **B** for tea).

Discussion

In line with our main hypothesis, the results showed that no more than a third of all manipulation trials were detected by the participants. Thus, in the great majority of trials they were blind to the outcome of their choice. Moreover, in two thirds of the trials we classified as detected the participants showed no conscious reaction at the moment they received the manipulated outcome. Instead, they either made the claim at the end of the study that they had felt something was amiss about the situation, or they reported a sensory change without realizing that the product they were experiencing was not the one they previously preferred. Even for such remarkably different tastes as spicy Cinnamon-Apple and bitter Grapefruit, or for the sweet smell of Mango and the muscular Pernod (that variously evokes associations of liquorice candy, or cough-syrup, or strong aniseed spirits like Absinthe and Ouzo), was no more than a fifth of the manipulation trials detected concurrently, and less than half counting all forms of detection.

But what does this result mean? Why did the participants fail to notice so many of the mismatches? One obvious answer is that they did so because they simply did not care about the decisions made in our experiment. This is a reply with intuitive appeal. An experimental finding like choice blindness is naturally bound at the limits by choices we know to be of great importance in everyday life. While it lies close at hand to speculate about couples at the altar solemnly affirming their choice of partner, and then (after the minister pulls some unearthly sleight-of-hand!) bringing home a complete stranger, no one would fail to notice such a change (and this, we fear, includes even those involved in the most hasty of Las Vegas marriages).

Yet, we feel there is ample of territory to explore between our small-scale consumer survey, and the preposterous idea of covert spouse swapping. In the study we found evidence of a correlation between the discrepancy of rated attractiveness between the two samples, and the likelihood of detection (but only for the jam, and not for the tea part of the study). At the same time, it certainly did not seem as if the participants were indifferent about the decisions made in our test. As is evident from the mean attractiveness scores for the chosen items (6.8 for jam and 7.0 for tea on a scale from 1-10), people tend to *like* jam and tea.[2] Similarly, from a comparative standpoint, even gaps of rating that ran almost the full length of the scale did not always result in successful mismatch detection (i.e. for choices between samples that were described as "near perfect", and "plain horrible").

On the other hand, *if* choice blindness only occurred in situations where the participants do not really care about the outcome of their choices, we could instead use our experimental approach to measure the level of interest actually experienced by partici-

[2] Indeed, this was one of the reasons why we chose these particular products for the test; tea is one of the most drunk and celebrated beverages in the world, and the average European consumer gobbles down more than one kilo of jam every year, see EU Market Survey: Preserved fruit and vegetables (2003).

pants in psychological experiments. Decision tasks like the one we used in this experiment are exceedingly common in psychological research (frequently in conjunction with different forms of rating procedures), and often it is very difficult to appraise how engaged the participants are in the task at hand. Given that fewer instances of choice blindness ought to be expected for choices that participants care more about, then choice blindness manipulations would be perfectly suited as an implicit probe of the interest of the participants.

But interest cannot be the full story. We also found that the degree of choice blindness exhibited by the participants was modulated by the similarity of the choice pairs, with significantly higher rates of detection for the most dissimilar pairs compared to the most similar ones. As de Houwer (2006) argues, implicit measures used in experimental psychology typically are not as "clean" and unambiguous as researchers often like to think they are. Like the parent phenomenon of change blindness, choice blindness is likely to be sensitive to both motivational and attentional factors, to various encoding and retrieval demands, and to the particular nature of the external feedback used (e.g. see Mitroff, Simons & Franconeri, 2002; Rensink, 2002; Simons & Rensink, 2005).

Thus, while we find it overly optimistic to expect a perfect correlation between susceptibility to choice blindness and other (implicit or explicit) measures of interest or attitudes, it is an highly interesting possibility that choice blindness might be of help in mapping out the relationship between the abstract concepts of common sense psychology (wants, needs, reasons, intentions, etc.) and the functional architecture of decision making.

For our remaining three hypotheses, we got more mixed results. The gift incentive did not generate an increase in the detection rate, but instead resulted in a lowered level of detection for the tea condition. This goes against the thrust of evidence presented by Hertwig and Ortman (2001) about the effects of incentives in psychological research on decision making. However, it is not unheard of that incentives can generate diminished engagement in a test (e.g. Read, 2005), and we can speculate that we should have tied the incentives more clearly to a performance goal to generate a greater effect on effort and attention (Hertwig & Ortmann, 2003).

Our attempt to measure more indirect effects of the choice manipulation on the judgments of the participants did not result in any clear pattern of findings. We found that participants in the tea condition rated it as more difficult to discriminate between the two samples in non-detected manipulated trials, as compared to non-manipulated trials, but this effect was not found for the jam condition. Given the subtlety of the effect, in future studies the possible memory distortion found here would have to be targeted and isolated in a longer series of decisions. Finally, we found no effect of the undetected manipulated trials on the expressed confidence of the participants in their choice. This can either be interpreted as choice blindness being a very robust effect, with the mismatched outcome having no impact on confidence, or that the particular format we used for eliciting confidence judgments was not sensitive enough to register any awareness that the participants might have had of the manipulation. For example, Tunney (2005) have suggested that binary high-low confidence judgments might do a better job of capturing purported implicit influences upon judgments than a continuous scale like the one we used.

In summary, we have demonstrated considerable levels of choice blindness for decisions between samples of jam and tea at a local supermarket. This result extends the findings of Johansson, Hall, Sikström and Olsson (2005) using visual stimuli to the modalities of taste and olfaction, and further establishes choice blindness as a robust effect in the domain of decision making.

References

Brendl, C. M., Chattopadhyay, A., Pelham, B. W., & Carvallo, M. (2005). Name letter branding: Valence transfers when product specific needs are active. *Journal of Consumer Research, 32*(3), 405-415.

Bruce, V., & Young, A. (1998). *In the eye of the beholder: The science of face perception.* Oxford: Oxford University Press.

Chartrand, T. L. (2005). The role of conscious awareness in consumer behavior. *Journal of Consumer Psychology, 15*(3), 203-210.

de Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers and A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction*. Thousand Oaks, CA: Sage Publishers. 11-28.

Dijksterhuis, A., Smith, P. K., Van Baaren, R. B., & Wigboldus, D. H. J. (2005). The unconscious consumer: Effects of environment on consumer behavior. *Journal of Consumer Psychology, 15,* 193-202.

EU Market Survey (2003). *Preserved fruit and vegetables*. Centre for the promotion of imports from developing countries.

Ferraro, R., Bettman, J., & Chartrand, T. L. (in press). *Like ships passing in the night: The effect of minimal social interactions on brand choice.*

Fitzsimons, G. M., Chartrand, T. L., & Fitzsimons, G. J. (in press). *Automatic effects of brand exposure on behavior.* Unpublished manuscript.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience, 6*(4), 428-432.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Bhavioral and Brain Sciences, 24*(3), 383.

Hertwig, R., & Ortmann, A. (2003). Economists and psychologi sts' experimental practices: How they differ, why they differ, and how they could converge. In I. Brocas and J. D. Carrillo (Eds.). *The psychology of economic decisions.*Oxford: Oxford University Press.

Irmak, C., Block, L. G., & Fitzsimons, G. J. (2005). The placebo effect in marketing: Sometimes you just have to want it to work. *Journal of Marketing Research*, *42*, 406-409.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995-1006.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*(5745), 116-119.

Mitroff, S. R., Simons, D. J., & Franconeri, S. L. (2002). The siren song of implicit change detection. *Journal of Experimental Psychology: Human Perception and Performance, 28(4),* 798-815.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2004). The effect of preference fluency on consumer decision making, *Working Paper*.

Olsson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion, 5*(4), 498-502.

Read, D. (2005). Monetary Incentives, What Are They Good for?. *Journal of Economic Methodology, 12,* 265-76.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245-277.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199-226.

Schwaninger, A., Carbon, C.C., & Leder, H. (2003). Expert face processing: Specialization and constraints. In G. Schwarzer and H. Leder (Eds.). *Development of face processing.* Göttingen: Hogrefe. 81-97.

Schwarz, N. (2003). Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research, 29*, 588-594.

Shiv, B., Carmon, Z., & Ariely, D. (2005). Placebo effects of marketing actions: Consumers may get what they pay for. *Journal of Marketing Research, 42*(4), 383-393.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present and future. *Trends in Cognitive Sciences, 9*, 16-20.

Simonson, I. (2005). In defense of consciousness: The role of conscious and unconscious inputs in consumer choice. *Journal of Consumer Psychology, 15*(3), 211-217.

Strack, F., Werth, L., & Deutsch, R. (2006). Reflective and Impulsive Determinants of Consumer Behavior. *Journal of Consumer Psychology, 16*(3), 205-216.

Tanaka, J. W., & Farah, M. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46*, 225-245.

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory and Cognition, 25*, 583-592.

Tunney, R. J. (2005). Sources of confidence in implicit cognition. *Psychonomic Bulletin and Review, 12*, 367-373.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181-192.

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin, 31*, 121-135.

Woodside, A. G. (2004). Advancing from subjective to confirmatory personal introspection in consumer research. *Psychology and Marketing, 21*(12), 987-1010.

The term *Choice Blindness* refers to the experimental finding that people are prone to miss even dramatic mismatches between what they want and what they get. This effect is demonstrated in a series of experiments using different stimuli as well as different experimental methods.

Not only were the participants in these experiments blind to the manipulation of their choices, but they also offered introspectively derived reasons for preferring the alternatives they were given instead. The thesis also analyses the participants‘ verbal reports when explaining choices they had in fact not intended to make.

One conclusion drawn in the book is that we may know a lot less about the reasons for our actions than we think we do.

LUND UNIVERSITY COGNITIVE STUDIES 130