The malleability of political attitudes

Choice blindness, confabulation and attitude change

THOMAS STRANDBERG
COGNITIVE SCIENCE | LUND UNIVERSITY | 2020

The malleability of political attitudes

The malleability of political attitudes

Choice blindness, confabulation and attitude change

Thomas Strandberg

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Thesis advisors: Petter Johansson, Lars Hall and Fredrik Björklund
Faculty opponent: Daniel M. Oppenheimer

To be defended, with the permission of the Faculty of Humanities and Theology of Lund University, in LUX room C126,
on 24 August 2020 at 10:15

Lund University Cognitive Science
Department of Philosophy
LUND UNIVERSITY

DOCTORAL DISSERTATION

Date of issue 2020-08-24

Author(s): Thomas Strandberg

Sponsoring organization

Title and subtitle:
The malleability of political attitudes: Choice blindness, confabulation and attitude change

Abstract

This thesis is an empirical and theoretical investigation of choice blindness, in particular in the domain of political attitudes. Choice blindness is a cognitive phenomenon in which people do not notice dramatic mismatches between what they choose and what they get while still offering seemingly introspective arguments to explain their (putative) choice. In this thesis I demonstrate that the effect also applies to salient political attitudes and evaluations of political candidates. All studies took place in close connection to real elections, and I have developed new tools building of the underlying choice blindness methodology to collect the data. Further, I explore the potential downstream effects, such as influence on voting intentions, and lasting attitude changes. I also explore the potential mechanism behind this effect, and show that confabulatory reasoning plays an important part in facilitating the observed attitude changes.
This thesis comprises of four papes published in academic journals:

Paper 1 is a proof-of-concept exploring if the choice blindness effect also applies to political attitudes and voting intentions. During the experiment we developed a novel survey methodology that allowed us to switch out peoples' survey ratings and give false feedback about which political camp their survey score aligned with.

Paper 2 expanded on the findings from paper 1 and focused the 2016 U.S presidential election. That election was a heated affair, with Hillary Clinton and Donald Trump head-to-head in a tight race, and many worried that the US was becoming increasingly polarized and that too much of the debate focused on the candidates' characters and not the political issues. We ran two experiments and demonstrated that we could get US citizens to express less polarized and more open-minded views about the two presidential candidates.

Paper 3 is my flagship paper in which I show that when participants falsely believe that they have stated some attitude, it can lead to lasting changes in those attitude. Further, the effect is much more accentuated if the participants also provide confabulating arguments of why they stated that attitude. These findings highlights the influential power of confabulation in facilitating changes in political attitudes.

Paper 4 is an exploratory effort to better understand and pinpoint what determines whether a manipulated trial is accepted or corrected, and how participants experience the correction.

Key words: choice blindness, confabulation, self-perception, political psychology, attitude change

Classification system and/or index terms (if any)

Supplementary bbliographical information

Language: English

ISSN: 1101-8453 Lund University Cognitive Studies 179

ISBN: 978-91-89213-06-7 (print)
ISBN: 978-91-89213-07-4 (digitalt)

Recipient's notes

Number of pages 186

Price

Security classification

Signature 

Date 2020-05-18

The malleability of political attitudes

Choice blindness, confabulation and attitude change

Thomas Strandberg

LUND
UNIVERSITY

Cover art by Sofia Khwaja

Faculty of Humanities and Theology
Department of Philosophy
Cognitive Science

ISBN 978-91-89213-06-7 (print)
ISBN 978-91-89213-07-4 (digital)
ISSN 1101-8453

Lund University Cognitive Studies 179

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

Acknowledgements

First, I would like to express my gratitude to the entire choice blindness dream team for helping me start and finish this project. I could not have pulled this off without them. And Anders Lindén who built the applications used in paper 3 and 4. Plus all the smart and tireless students at LU and McGill, that has made data collection so much easier (and more fun). Then I would like to thank my supervisors. Petter Johansson has truly been the best possible supervisor – always there, always so pragmatic and solution-oriented (plus kind, patient, understanding and genuinely helpful). Lars Hall, my co-supervisor, has from the very first moment added that spicy mix of thought-provoking and brilliant ideas that has motivated me to keep going and to test new things. I would not have made it without my co-supervisor Fredrik Björklund who helped get in to the PhD program, helped me start my first projects, and has been the perfect link to the world of social psychology. I would also like to thank Philip Pärnamets for being so genuinely curious and passionate about the human mind (and other related and unrelated phenomena) and, perhaps more importantly in the context of academia, being the ultimate conference/travel buddy. I have also had the pleasure to work Jay Olson, the magician-gone-psychologist. Some of my best research memories I share with this guy, and I look forward creating new ones in the future. I am also both lucky and proud to have worked at LUCS, with all the incredible LUCS people. I doubt there is a more easy-going, innovative, and friendly research environment out there. Special shout-out to some of the LUCS superheroes: Tomas Persson, Anna Östberg, Eva Sjöstrand and the almighty Anna Cagnan Enhörning – words are superfluous. I would like to thank my dad for always asking the two most important questions in the life of a phd student: "so what was it you were doing now again?" and "are you finishing soon?" Last but not least, I could not have done this without +100% of support from my family – Josefine, for always being super-understanding, even when it has been tough and felt unfair; my kids Kerstin and Eskil, for helping me keep my mind where it should be.

Contents

List of original papers

Paper 1

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLOS ONE, 8*(4): e60554.

Paper 2

Strandberg, T., Olson, J. A., Hall, L., Woods, A., & Johansson, P. (2020). Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *PLOS ONE, 15*(2): e0226799.

Paper 3

Strandberg, T., Sivén, D., Hall, L., Johansson, P., Pärnamets, P. (2018) False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General, 147*(9), 1382–99.

Paper 4

Strandberg, T., Hall, L., Johansson, P., Björklund, F., & Pärnamets, P. (2019). Correction of manipulated responses in the choice blindness paradigm: What are the predictors? In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Montreal, CA: Cognitive Science Society.

Some definitions

Choice blindness as both an effect and as a methodological tool

Note that in the text, I am treating choice blindness both as an effect or cognitive phenomenon as well as a methodological tool to study attitudes and self-reports.

Manipulation

Throughout the thesis I use different terms to describe the choice blindness manipulation, such as 'manipulation', 'false feedback', 'switch', 'shift' and 'change'.

Choices and propositional attitudes

When describing and discussing the different research, I switch interchangeable between words such as choices, decisions, judgments, preferences, attitudes, opinions etc. depending on the context. Usually I am referring to an action or the outcome of an action, or a proposition towards something (e.g. 'I believe that…', 'I think that…').

Part I

Introduction

The study of self-reports: do we know what we think we know?

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 extremely difficult to probe the representations underlying this process, or to determine what we can know about them from the 'inside', by reflection and introspection (Nisbett & Wilson, 1977; Jack & Roepstorff, 2004; Gilovich, 1991; Dennett, 1987; 1991a). One problem for researchers interested in experimental investigations of decision-making is that they cannot take the reports of the participants involved at face value when it is the very terms used in these reports that they want to study (i.e. what participants claim to 'intend' and 'decide', what their purported 'reasons' are, etc.). At the same time, self-reports about choice is an indispensable tool for academic research in the humanities and social sciences.

In their seminal paper, Nisbett and Wilson (1977) ran a series of studies to illustrate people's limited knowledge of the causes and processes that influence their attitudes and behaviors. In one of their experiments, two groups of participants watched the same movie. One group watched the movie in a quiet and comfortable room with no interruptions, and the other group watched the movie in a room with loud construction work going on just outside. Afterwards, participants in both groups were asked to describe what they thought about the movie. In addition, participants that were exposed to the construction work were also asked how much the noise influenced their enjoyment in the movie. They reported that the construction noise greatly disrupted the movie experience and that they would have liked the movie better had it been watched without the construction noise. However, when comparing the overall ratings of the movie,

both groups found it equally good. In another experiment, the researchers asked participants to choose between a variety of nightgowns or pantyhose. The items were placed in a row and participants tended to prefer the garment to the right. Participants were then asked to explain why they preferred a particular item. However, there was a twist: all items were identical. Still, participants rationalized their choice by saying that the quality was better or that they preferred one particular type of knitting over the others etc. None of the participants mentioned that they chose the garment because of its placement which was the only true feature separating the items.

Nisbett and Wilson concluded that people have a strong motivation to know how things work and to explain their environments, and to accomplish this humans have evolved to instantaneously see patterns and logical connections everywhere (Dennett, 1991b). For example, when children play and interact with others it comes natural to them to experiment, observe others' actions, and then interpret statistics and patterns to understand how the surroundings work (Gopnik, 2004; Busch & Lagare, 2019). This is how they so quickly acquire knowledge and learn new abilities. Interpreting the external world is immensely important to humans. As pointed out by Nisbett and Wilson, even our choices and the reasons for making those choices are strongly influenced by situational factors. In the first example, participants are confident that external factors – the construction noise – affected their enjoyment of the movie, although it did not. In the second example, participants are confident that their choice was based on personal preference and deliberation, which it was not. In both examples participants rationalize their judgments based on plausible and seemingly coherent reasons – but these were objectively unrelated to the true underlying reasons. This shows that also mental states such as beliefs and preferences tracks patterns in the world, and is thereby inferred and shaped by available evidence and not given by introspection as is commonly assumed by folk psychology (Dennett, 1991b; 1991c; Zawidzki, 2008). This is an effective, convenient and adaptive way to process information; however it also highlights the perils in trusting peoples' self-reports. In both of the experiments described above, we know that the reasons participants give are false because we are aware of the design of the experiment. But without knowledge about the experimental context, we would have no reasons to doubt what the participants said. Even more importantly, the participants are also unaware that their reasons are strongly influenced by the context. To them, the construction noise affected their

enjoyment, and the fabric texture really made them prefer a specific pantyhose. This shows how difficult it is to separate a self-report that is based on introspection with one that is inferred from context. Although at times questioned based on methodological grounds (e.g. White, 1980) the experiments done by Nisbett and Wilson are historically important for highlighting the unreliability of self-reports, and inspired how decision-making and introspection could be studied.

The genesis of choice blindness

Nisbett's and Wilson's underlying methodology needed an update, in which not just the situation but the actual outcome of the choice could be controlled. To address this issue, Johansson and colleagues created a methodological wedge that allowed researchers to get in between the choice and the self-reported arguments for making that choice. Nisbett and Wilson used between-subject experimental design, so they could only compare the differences between groups. Instead, Johansson and colleagues wanted a within-subject design where effects could be more directly measured at participant level. Further, this wedge made it possible to be certain whether participants' gave reasons that were constructed after the choice was already made. So they introduced the phenomenon of choice blindness, analogous with the change blindness phenomenon which shares some traits (Rensink, O'Regan & Clark, 1997). Change blindness is a visual attention phenomenon in which people tend not to notice changes that occurs in a visual scene right in front of them. While change blindness is strictly perceptual, choice blindness on the other hand means that people tend not to notice changes about their own choices. In this choice paradigm the experimental method was originally inspired by techniques used in close-up card magic, which allowed the researchers to surreptitiously manipulate the relationship between the choice and the outcome that the participants experienced. In their first study (Johansson, Hall, Sikström & Olsson, 2005), participants were shown pairs of pictures of female faces and asked to choose the most attractive face of each pair. In some trials, immediately after their choice, they were asked to verbally describe the underlying reasons for their choice. 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 (see Figure 1). That is, the participants were given false feedback about their choice,

and were then prompted to explain why the choice was made. From a common sense perspective it would seem that everyone immediately would notice such a change – but this was not what the researchers found. Instead, in a majority of the trials the participants accepted the opposite face as the one they preferred, while also offering seemingly introspective reasons for their choice (Johansson, Hall, Sikström, Tärning & Lind, 2006).

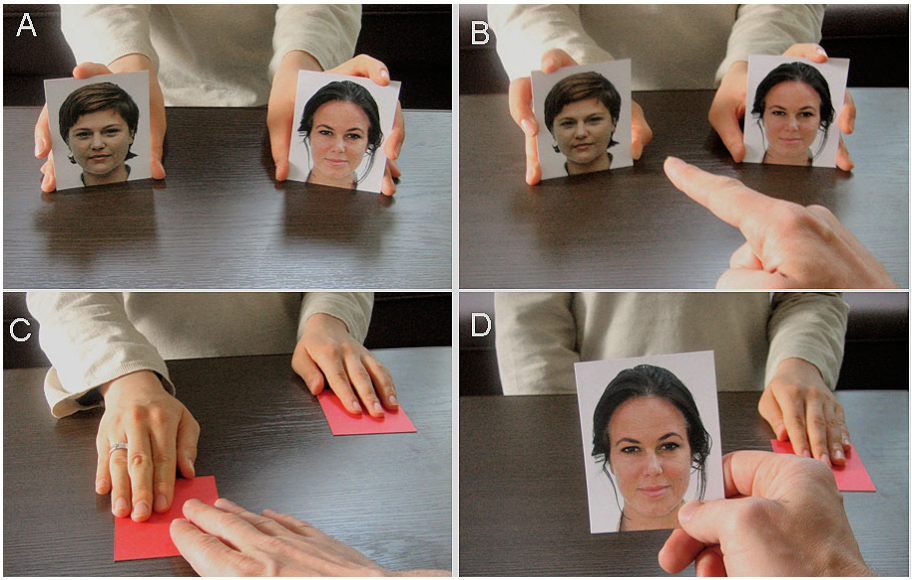

Figure 1 – An illustration of a manipulated trial in the first choice blindness experiment. (A) Participants are shown two pictures of female faces and asked to choose which one they find most attractive. Unknown to the participants, a second card depicting the opposite face is concealed behind the visible 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 asked to explain their choice.

That is, the participants seemed to experience the non-chosen alternative as their preferred face and since the explanations matched the non-chosen face these were likely rationalizations made post-hoc. Importantly, without choice blindness as a wedge, it would not have been possible for the researchers to know that these rationalizations were made up at the time participants were asked to explain their choices.

At the end of the experiment, but before debriefing participants about the study purpose, participants were asked if they thought they would notice if someone

had suddenly exchanged their preferred alternative for the non-preferred. Almost everyone answered "yes", highlighting that most people assume to have access to their mental states and control over their decisions (Johansson, Hall, Sikström & Olsson, 2005, supplemental material).

Since the 2005 study, choice blindness has been replicated in a variety of domains, for different modalities, and with a myriad of method and design variations. For example, choice blindness has been found for male and female faces, both when presented by hand (as in the original study) and on a computer screen (Johansson, Hall & Sikström, 2008) as well as by a virtual agent (Johansson, Hall, Gulz, Haake & Watanabe, 2007). Recently, Wang, Zhao, Zhang and Feng (2018) found that the effect was larger for sad faces (compared to happy or neutral). Choice blindness has also been demonstrated for consumer choice, such as for the taste of jam and the smell of tea (Hall, Johansson, Tärning, Sikström & Deutgen, 2010), and for the attention to product ingredients (Cheung, Junghans, Dijksterhuis, Kroese, Johansson, Hall & De ridder, 2013). Further, choice blindness has been observed in the clinical domain such as for malingering and psychiatric self-diagnoses (Merckelbach, Jelicic & Pieters, 2011) as well as for evaluation of Obsessive-Compulsion Disorder (Aardema, Johansson, Hall, Paradisis, Zidani and Roberts, 2014); for financial decisions (McLaughlin & Sommerville, 2013); and risk assessment (Chater, Johansson & Hall, 2011). Sommerville and McGowan (2016) found that also children exhibited the effect – although detection increased to 80 % when the stimulus was chocolate with brand names! In one particularly relevant (and remarkable) branch of choice blindness research, the effect has also been found to affect witness testimonies and processing of legal information (e.g. Sauerland, Sagana & Otgaar, 2013; Sagana, Sauerland & Merckelbach, 2013; 2014; 2016; 2017). In one study, Sauerland, Sagana and Otgaar (2013) had participants listen to pairs of voices and decide which one they thought sounded most sympathetic or criminal. In a follow up task, less than 30 % detected when their choice of voice had been switched for the non-chosen alternative, which shows that choice blindness also extends to auditory stimuli and is relevant for so called ear-witness testimony. In another study, Cochran, Greenspan, Bogart and Loftus (2018) asked participants to view pictures of crimes taking place and then asked them to either recall details about the crime or identify a suspect from a line up. Cochran and colleagues found that when participants accepted manipulations to their eye-witness reports this affected their memories about the crime. These latter studies highlights that

choice blindness can have significant legal implications as well, and can potentially lead to wrongful convictions and obstructed justice.

The list of research presented about shows that choice blindness is a robust and widely replicated phenomenon that has been studied across many domains and modalities. However, choice blindness as a phenomenon is to a large extent still unexplained, and I aim to expand on choice blindness research from methodological, empirical, and theoretical perspectives. The foundational idea of the studies that I present in this thesis originated in the dawn of choice blindness, before it became such a widely studied cognitive phenomenon.

Choice blindness and moral issues

The fact that the first choice blindness findings were based on choices between two pictures of faces initially raised some doubts about the extension of the effect. One argument could be that the task was not motivating enough to the participants, another argument that the findings could perhaps not be generalized to other more important topics. In conjunction with the publication of the second choice blindness paper: 'How something can be said about telling more than we can know' (Johansson, Hall, Sikström, Tärning & Lind, 2006), James Moore and Patrick Haggard (2006) commented that:

> "…in the [choice blindness paradigm] the choice that is made is decidedly unimportant; it is unlikely that people profoundly care whether or not a face is attractive or not… a convincing refutation of this criticism would be a demonstration of the [choice blindness] effect for decisions regarding moral issues, for example. These would be decisions that are presumably less fallible and more resistant to confabulation."

We accepted this challenge and designed a novel tool that would allow us to use the underlying choice blindness methodology and give false feedback about peoples' moral attitudes.

In the first ever experiment on choice blindness and attitude responses (Hall, Johansson & Strandberg, 2012), participants answered a survey covering moral issues and their task was to rate to what extent they agreed or disagreed with them on a bi-polar scale. The experiment consisted of two conditions, one covering foundational moral principles (such as whether or not it can be justifiable to harm

another person) and the other covering concrete moral examples (for example if it can be justifiable to hide immigrants awaiting deportation).

The self-transforming survey

To accomplish this, we needed to expand on the original experiment and invent a way to apply the underlying choice blindness methodology on text and scale responses (which is the typical format in moral decision-making). Further, we wanted an experimental design that was easy to use and allowed us to get out of the lab and reach a more diverse population. After testing several different manipulation techniques, that included the use of tiny magnets, vanishing ink, bleed-thru paper, and a swami sharpie concealed in a fake thumb, we finally settled for a simple paper-pen questionnaire design. We invented the self-transforming survey – a questionnaire building on a stage magic routine called out-to-lunch (Robinson, 1898; Bowyers, 1928). During a typical out-to-lunch act, the participant writes some personal information onto a piece of paper (or playing card), folds it, and then keeps it safely in her pocket away from the eyes and hands of the magician. At the end of the act, when the magician asks the participant to reveal what she wrote on the note, it now reads something entirely different (originally "out to lunch"). Similarly, the self-transforming survey allowed participants to fill out the survey without any experimenter interference. Due to the design of the questionnaire, the false feedback was automatically generated after all the questions had been answered but before participants were asked to explain them. But instead of a note saying "out-to-lunch" the outcome or implications of the described moral issues that the participants' rated had been magically reversed.

The questionnaire was attached to a clipboard, with the moral statements distributed over two pages. After completing the survey, participants were asked to return to the first page, read three of the statements out loud, and explain why they responded the way they did. At this point two of the statements had now been manipulated so that the participants' responses indicated the opposite positions.

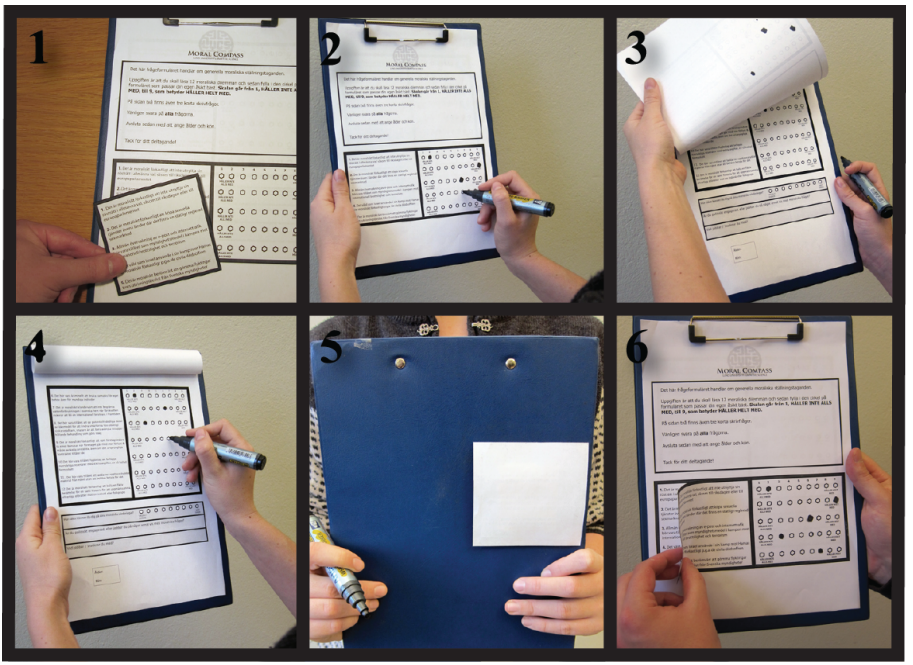

Figure 2 – An illustration of a manipulated trial in the moral experiment. (1) The questionnaire is attached to a clipboard, with the questions distributed over two pages. A paper slip with moral statements is attached to the first page of the questionnaire to conceal the same, but negated set of statements printed on the page. (2) The participants rate their agreement with the statements on the first page of the questionnaire and (3) they turn to the second page, and (4) rate their agreement with a second set of principles. (5) When the participants are asked to flip back the survey to the first page to discuss their opinions, the add-on paper slip from (1) now sticks to a patch of stronger glue on the backside of the clipboard, and remains attached there. This reveals the altered set of issues on the first page, and when the participants now read the manipulated statements the meaning has been reversed (in effect, the equivalent of moving the actual rating score to the mirror side of the scale). (6) During the debriefing, the experimenter demonstrates the workings of the paper slip to the participants, and explains how the manipulation led to the reversal of their position. See http://www.lucs.lu.se/cbq/for a video illustration of the method.

After the participants had read a statement, we interjected and summarized their attitude in a question by saying "so you don't agree that [statement]?" or "so you do agree that [statement]?" to avoid any misunderstanding of what the rating implied. The reversal was achieved by attaching a lightly glued paper-slip on the first page of the questionnaire, containing opposite versions of some of the statements.

The layout and shape of the attached slip allowed it to blend in perfectly with the background sheet (for example, the edges of the slip were blackened). When the participants folded the first page over the back of the clipboard, the paper-slip stuck on an even stickier patch on the backside of the questionnaire. Thus, when

participants flipped back to the first page in order to explain those responses, this revealed two statements that now had the opposite meaning compared to what the participants had originally responded to. With this procedure, the magic trick behind the manipulation was seamless and almost impossible for the participants to notice.

Importantly, all the magic and manipulation happened in the hands of the participants themselves. During the typical out-to-lunch act the participants are fully aware that they are in the middle of a magic trick and expect the things they write down to change. However, in the context of a moral survey participants have no reason to believe that their responses will suddenly change. The questionnaire format also made it easy and intuitive to correct any errors; in general people are familiar with occasionally misreading or ticking the wrong box on a form or survey.

One instant advantage of the clipboard design was that the portability of the "device" made it possible to take the experiment out of the lab and into the streets and thereby reach a more diverse population than the typical university student. Further, the questionnaire design was intuitive so it did not need any extensive instructions, and it had a very ecological procedure since most people have answered some kind of survey before and have bumped in to pollsters in the streets. We also tried to make the interaction with the participants as naturalistic and non-invasive as possible: we stressed that there was no time constraint, and that we had no moral or political motives and would therefore not judge or argue their opinions in any way.

Summary of the moral experiment

Just as in the original experiment, participants often accepted the false feedback made to their moral attitude responses. This meant that when participants accepted and explained a manipulated response their initial rating was the same but the actual moral statement had been changed. It could look something like this: if a participant first strongly agreed that "It is morally reprehensible to purchase sexual services in democratic societies where prostitution is legal a regulated by the government" the manipulation changed the response to strongly agreed that "It is morally defensible to purchase sexual services in democratic societies where prostitution is legal an regulated by the government". In other words: "reprehensible" had been switched to "defensible" while the rating was

kept the same. This technique made it possible to investigate whether participants would exhibit similar blindness as in Johansson et al. (2005), but using moral issues as stimuli. Studies have shown that people are capable of both expressing strong convictions in a moral position while at the same time incapable of explaining why that position is held (Haidt et al. 2001). Thus, it was possible that our participants as well might express strong ratings but not really being able to defend these. However, this is not what we found, and instead participants gave long and vivid explanations for both non-manipulated and manipulated trials. This made us wonder what role confabulatory reasoning could play in the context of a choice blindness situation, which we further explored in paper 3 of this thesis. Another immediate observation concerning the difference between the moral experiment and the original study was that there was a difference in how people experienced the false feedback during detected trials. Instead of detecting the sleight of hand, participants rather corrected themselves, by claiming to have made a mistake (one feature we shall return to in paper 4).

To conclude: Moore and Haggard (2006) raised doubts about whether the effects of choice blindness would be found in domains that are more relevant to people – such as their moral attitudes. In Hall, Johansson and Strandberg (2012) we showed that people indeed often do not detect when their survey responses to moral issues have been manipulated – instead, they accepted the manipulated responses as their own and gave reasons supporting them. This raised some new research inquires which has formed the scope of this thesis.

This thesis project: the malleability of political attitudes

The insights from the original face study and the moral study serve as the backdrop for the research presented in this thesis. The scope of this research can be divided into three main objectives:

I. *Investigate if choice blindness extends to political attitudes.* The finding in Hall, Johansson and Strandberg (2012) was an important step – both theoretically and methodologically – to study the limitations of the choice blindness effect. A natural next step was to move into political attitudes. In particular during elections, the measurement of political attitudes have direct real-world implications as they influence voting strategies, all kinds of political decisions, as well as define much of our social lives (Taber & Lodge, 2013).

II. *Investigate if choice blindness effects last over time.* That choice blindness can have consequences and lead to downstream effects has been shown in repeated choice tasks (Johansson, Hall, Tärning, Sikström & Chater, 2014; see also see also Taya, Gupta, Farber & Mullett-Gilman, 2014). While long-term effects have not been explored. Establishing if (and to what extent) the false feedback can have a lasting effect on peoples' political attitudes will be an important contribution to the choice blindness paradigm, and to the literature on attitude change and political psychology.

III. *Explore determining factors of why participants accept or correct the false feedback.* In order to understand choice blindness as a phenomenon it is important to study the mechanisms and determining factors behind the effect, such as what makes participants correct or accept the manipulation. In paper 4, I specifically address this issue by testing and discussing some potentially relevant factors.

Prelude to the papers

From here, I will connect each of these objectives with the four studies that I have conducted. In turn, the studies will be broken down into relevant parts and these will be discussed mixing theory from cognitive science and social psychology. At the end I close with some final words on how these findings can inspire future research on both choice blindness and political psychology.

To sum up my position before we set out, my overarching interpretation of the results and findings in these studies (and other similar or related experiments) is through a self-perceptive inferential framework, in which people observe their

own actions and behaviors and infer or interpret what the reasons for making those actions must have been. As such, I see choice blindness primarily being caused by lack of introspective access mixed with interpretation of action and situational information. Given the cross-disciplinary nature of choice blindness there are of course many other possible explanations involving information processing, attention, perception, encoding and retrieval of memory as well as various motivational and contextual factors. Some will be brought up here and some will be left for another time.

Part II

Choice blindness and political attitudes

Summary of paper 1: proof-of-concept

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLOS ONE, 8*(4): e60554.

Background and rationale

In paper 1 we wanted to expand on the findings from Hall, Johansson and Strandberg (2012) and apply the choice blindness methodology on political attitudes. The study was conducted during the final month of a Swedish general election, and polling of public opinion was a prevalent phenomenon. Virtually every media outlet had its own "election compass", where people could answer a political survey and instantly get a summary of which political alternative that best represented their views. At that time there was also debates in Sweden and the US if politicians should focus all their effort on 'swing voters' – i.e. the roughly 10% of undecided voters. We wondered if we could experimentally reveal a higher percentage of potential voting flexibility.

Methods, design and experimental procedure

A total of 162 participants in Malmö and Lund answered our election compass. First, they stated which of the two contending coalitions they planned to vote for and then rated to what extent they agreed on 12 political issues where the two coalitions had different opinions. These issues were selected together with leading

opinion researchers in Sweden, and represented the most critical issues of the election. The election compass had a new way of giving participants false feedback: instead of changing the words in the issues, we directly changed participants' ratings. To achieve this, we discreetly observed participants as they were filling out the survey, and filled out an identical cut-out of the scales but replaced some of the participants' ratings with new manipulated ratings. The cut-out scales had glue on the back, and when participants were finished with answering the survey, we momentarily took it to review and pasted the new set of ratings on top of theirs. After participants had completed the survey and we had changed some of their ratings, they were asked to explain some of them. Afterwards, we overlay a semi-transparent template categorizing the responses into either of the two political coalitions, and tallied the response into a summary score. As a consequence of the manipulations, their score had been shifted to favor their non-preferred coalition as measured by a voting intention question at the beginning. Participants were then asked to explain their overall score and then state their voting intention again (see Figure 1 on pp. 80).

Key findings and methodological advancement

Only 22% of the total amount of manipulated responses was corrected, and 92% of the participants endorsed the reversed summary score as their own. Voting intention measured before and after manipulation show that 45% of the participants were influenced by the false feedback and shifted their voting intention in the manipulation direction, out of which 10% shifted from voting for one of the coalitions to voting the other; and 19% shifted from clear coalition support to being undecided (see paper 1 for the full results).

Importantly, for the first time we show that choice blindness also applies to political attitudes. The issues used were the most important issues of the election and was constantly discussed in the media and by the public. We also found and interesting downstream effects, as the reversed summary score affected participants' voting intention. Further, we invented a new incarnation of the self-transforming survey, which increased the versatility of the experimental procedure.

Summary of paper 2: expanding the methodology

Strandberg, T., Olson, J. A., Hall, L., Woods, A., & Johansson, P. (2020). Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *PLOS ONE, 15*(2): e0226799.

Background and rationale

The 2016 US presidential election was a heated affair, with Hilary Clinton and Donald Trump head-to-head in a tight race, and many worried that the US was becoming increasingly polarized and that too much of the debate focused on the candidates' characters and not the political issues (e.g. Statler-Throckmorton, 2016; Waldman, 2016; Gleckman, 2016). To address this, we developed an experiment to investigate if we could make US citizens express less polarized and more open-minded views about the two presidential candidates.

Methods, design and experimental procedure

We took the overall principles from paper 1, but re-designed the survey to be about comparing the two candidates on their leadership abilities instead. Participants rated Clinton and Trump on 12 leadership traits, and the poles of the scales had pictures of faces. Using an overlay, we then segmented participants' ratings into three categories: favoring Clinton, favoring Trump, or open-minded. In paper 1 the ratings were "repolarized" from one side of the scale to the opposite side. Here, we wanted to participants' more extreme ratings less extreme (i.e. moved from favoring one of the candidates into open-minded).We also wanted to replicate it in an online setting, and thereby reach a larger and more diverse population. We therefore built an online version of the experiment which replicated most of the paper-and-pen features of the first experiment.

In experiment 1, 136 people attending the first presidential debate between Trump and Clinton (held in Hampstead, NY prior to the election), as well as people in the streets of New York City, participated in the study. Participants rated the two candidates on 12 leadership traits (such as trustworthy and analytic), and they were then given false feedback on a majority of their most polarized ratings, which shifted these into the open-minded category. After going over and explaining their ratings, we summarized their score and ask them to explain why

a majority of their responses fell into the open-minded category. At the end of the experiment, participants rated the two candidates again but now on their overall competency as leaders (see Figure 1 on pp. 98). In experiment 2, we replicated the study online one week before the election with 498 participants, but instead of rating their competency at the end participants rated how much they favored the two candidates (see Figure 3 on pp. 105).

Key findings and methodological advancement

Only 12% of the manipulations in experiment 1 and 41% in experiment 2 were corrected (the potential reasons behind this difference is discussed in Part 3). More importantly, a vast majority of participants (94% in experiment 1 and 72% in experiment 2) rationalized their now more "open-minded" or moderate views. For example, one Trump supporter claimed, "I guess I fall somewhere in the middle — I'd like to think I'm a little moderate. I think at this point it's important to be open-minded." Others claimed that their parents raised them this way, or that they needed to be unbiased in their line of work. There was no difference in either correction rate or rationalization rate between participants that originally had most ratings favoring Clinton with those that had most ratings favoring Trump. There was no difference in overall competency rating (experiment 1) or favorability rating (experiment 2) between experimental and control groups. Thus, the downstream effects found in paper 1 did not generalize to the relationship between evaluation of leadership traits and overall competency/-favorability.

These findings show that choice blindness also applies to contentious topics such as character evaluation, and generalizes to a US context involving liberal and conservative participants. Importantly, we show that a pen-and-paper design can be automatized and scaled to an online format, which enables fast and flexible data collection potentially reaching thousands of participants globally.

Summary of paper 3: lasting attitude change

Strandberg, T., Sivén, D., Hall, L., Johansson, P., Pärnamets, P. (2018) False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General, 147*(9), 1382–99.

Background and rationale

One potentially interesting downstream effect is whether choice blindness influences future expressions of political attitudes. In Johansson, Hall, Tärning, Sikström & Chater (2013; see also Taya, Gupta, Farber & Mullett-Gilman, 2014) participants rated pictures of faces before a choice blindness task, and after being given false feedback about their choices they rated the pictures again. Participants gave much higher ratings to the accepted pictures, showing that accepted false feedback can affect choices in the short term. However, no studies have explored if choice blindness has a more lasting effect.

Secondly, this enabled us to test the role of confabulation in facilitating attitude change. The underlying idea was that confabulating about falsely made attitude responses will reinforce the self-perceptive processes and therefor accentuate the attitude change (inspired by classical work from e.g. Janis & King, 1954; 1956; Broockman & Kalla, 2016, which reported that counter-attitudinal argumentation can influence future preferences).

Methods, design and experimental procedure

First, we introduced a new medium: the initial rating, manipulation, presentation and exposure to the false feedback all occurred on a digital tablet. It was designed to resemble the paper surveys and participants responded by drawing and X using a touch-sensitive tablet pen. If participants wanted to change a response, they would simply click the change icon placed next to each scale (which deleted their current response), and then draw another X.

We ran two experiments, the first on 140 participants and the second on another 232 participants. As target items, we used six issues about the environment and school politics. In experiment 2, we also included the Cognitive Reflection Test (CRT, Frederick, 2005), which is a short reasoning test which

probes peoples' capacity to override intuitive (yet wrong) gut responses and instead use more careful analytic thinking and arrive at the correct answer.

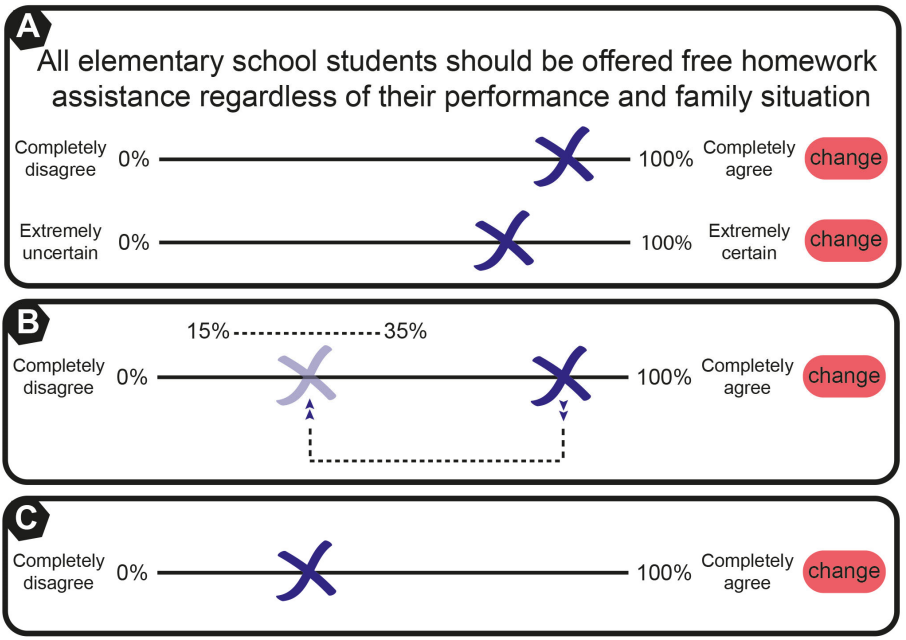

Figure 3 – The manipulation mechanism in paper 3. Participants rate to what extent they agree with a political statement as well as their level of confidence on a visual-analog scale ranging from 0% to 100% (A). After responding to all 12 statements, participants are asked to go over four of the responses together with the experimenter. At this stage, the application has moved two of their responses to the opposite side of the scale. The manipulation moves the responses across the midline and randomly places them between 15% and 35%, or 65% and 85% (B). In the acknowledge condition, participants are asked to just verify their responses. In the confabulation condition, they are also asked to explain the reasons behind each response (C). Participants can always change a response by clicking the change button (A–C).

To test the potential lasting effect, we set up an experiment which first included a typical choice blindness study, however participants were not debriefed about the study purposes immediately after the experiment. Instead, they stated their attitudes again in two follow up sessions conducted five minutes after the experiment and one week later (see paper 3 for details). To test the potential effect of confabulation on attitude change, during the first part of the study (that is, the choice blindness part) we divided the participants into two conditions: Acknowledge and Confabulation. In the Acknowledge condition, participants were instructed to go through some of their responses (by toggling between them

in the tablet survey). They would read the statement out loud, tell where on the scale their X was, whether this meant that they agreed or disagreed with the issues and about to what extent (such as "I strongly disagree with that"). Participants in the confabulation condition would do all of this but in addition also explain each response. This allowed us to compare what effect confabulation would have on correction rates as well as the responses in the two follow ups.

Key findings and methodological advancement

Overall, 50% of the manipulations were corrected. The correction rate was higher in the confabulation condition (54.5%) compared to the acknowledge condition (44.6%) indicating that the added reasoning make participants think more about their responses or the issues at hand.

Importantly, we found large effects of attitude change in the two follow up sessions, in which participants that accepted the false feedback shifted their attitude significantly in the direction of the manipulation. This effect was much larger in the confabulation condition; where participants shifted their responses in the five minute follow up by 30mm on the 100mm scale, and 22mm when measured on week later. A Bayesian regression model with data from both experiments shows that just accepting a manipulation affects participants' future responses on average 16.6mm. This is increased with another 9.8mm if participants also confabulate.

Additionally, the CRT score correlated with correction, meaning that participants with a high score were more likely to correct than those with a low score.

The digital version of the self-transforming survey increased the scalability and flexibility of the experimental procedure while still maintaining much of the look from the paper surveys. Further, it allowed us to setup and control all the experimental parameters, and measure every interaction participants had with the survey (such as time spent, how many times they changed or moved a response, etc.).

Summary of paper 4: exploring determining factors of accepting or correcting the manipulations

Strandberg, T., Hall, L., Johansson, P., Björklund, F., & Pärnamets, P. (2019). Correction of manipulated responses in the choice blindness paradigm: What are the predictors? In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Montreal, CA: Cognitive Science Society.

Background and rationale

In the final paper, the focus was on what may dispose the participants to correct the choice blindness manipulations. Given that the choice blindness tasks presented in this thesis involves expressing and discussing political attitudes, we primarily explored factors related to the concept of attitude strength. Attitude strength is typically defined as an attitude's resistance to change, persuasion, and contextual influence, and its effect on thought and behavior (e.g. Krosnick & Petty, 1995). Since choice blindness taps in to several of these defining features, we thought that attitude strength and choice blindness must be closely associated. We identified several candidate measures that we thought would contribute to predict whether a manipulation is corrected or not. These were two so called meta-attitudes (i.e. psychological impressions of the attitudes), two measures of cognitive style, and two measures of political awareness. All of these have received considerable attention in both psychology and political science where the concept of attitude strength is central (Fazio, 1995; 2007).

Methods, design and experimental procedure

The Participants completed a two-part experiment. One week prior to the main experiment, we assessed two of their cognitive styles: Preference for consistency (PFC) and Need for cognition (NC). PFC has been used to measure peoples' tendency to have consistent cognitions – such as behaving consistently with ones' attitudes (Cialdini, Trost & Newsom, 1995). PFC has also been shown to predict attitude changes resulting from social- or contextual influence (Bator & Cialdini, 2006). NC has been extensively used in attitude change research, where it has

been found that high NC individuals have stronger and more resistant attitudes compared to low NC individuals (Cacioppo, Petty & Kao, 1984; Haugtvedt & Petty, 1992). We also measured participants' political awareness by asking them to state how interested and involved they were in politics. In the experiment, in addition to stating how strongly they agreed/disagreed with a political issue, participants also stated how central each issue was to them (the meta-attitude centrality) and how strong conviction they had in each particular attitude (the meta-attitude commitment). Centrality was measured using three items such as "how important is this issue to you?", and commitment was measured with three items such as "how certain are you about your attitude towards this issue?" The experiment ran on a digital version of the self-transforming survey, and since the main theme of the experiment was attitude strength, we wanted to ensure that we got many extreme responses as well. To achieve this, we used succinct and often divisive wedge issues, such as "the gasoline price should be lower" and "the Swedish monarchy should be abolished". Further, we deployed a much more liberal manipulation rule, in which the application would choose a random position on the opposite side of the scale. The implication of this was that some participants gave a response of 99% which was changed to 1%.

Key findings and methodological advancement

Despite the use of succinct wedge issues and a manipulation rule which often generated extreme shifts, only 58.4% of the manipulations were corrected. The strength of the attitude was correlated with correction, however none of the additional measures (meta-attitudes, cognitive style, political awareness) were. This is noteworthy given the extensive amount of literature discussing their relation to the formation of strong and resilient attitudes (this is further discussed in part 3 as well as in paper 4).

Part III

What happens in a choice blindness experiment?

In all four studies presented in this thesis, participants receive false feedback about their political survey responses. In roughly half of the experimental trials participants accept the manipulated responses and rationalize them as if they were their very own. That is, participants justify a political position that is different, sometimes even opposite, of what they answered just a few minutes earlier. Importantly, as seen in paper 3, this can have significant lasting effects as accepting and confabulating about manipulated responses sometimes leads to attitude changes. How can the different processes that are at play here be understood? There are a lot of different things happening during a choice blindness task and here I present a first synthesis of some of the main psychological mechanisms and discuss them in the light of previous research and existing theoretical frameworks. The individual processes and theories discussed below are often relevant for understanding several separate aspects of the choice blindness situation. However, for clarity, each process and theory is discussed where I find it most appropriate.

1. Acceptance

One fundamental part in a choice blindness experiment is that participants sometimes accept the manipulations. To reiterate, a trial is categorized as 'accepted' when the participants are exposed to a manipulated response and acknowledge that this is what they intended to respond. In such a case, participants seem to accept the manipulated response as their own attitude.

One interesting aspect here is that even in trials when the participants correct the response when presented with the manipulation, they still think that the manipulated rating was their own original response. For example, they say things

like they feel as if "their" response does not match their attitude and they now correct it, claiming to have made a mistake or having changed their minds. Therefore, very few of the participants actually suspect that anything has been altered during the experiment, instead attributing any perceived "error" as coming from themselves. In paper 4, we categorized all the reasons participants gave for wanting to correct and found that almost no one detected the manipulation or suspected foul-play (as they sometimes did in for example Johansson et al. 2005).

But in the accepted trials participants believe that the false feedback response is what they intended to answer and that it represented their attitude in the matter. And from a common sense perspective, it may seem unintuitive that when people are given false feedback about their answers to a political survey they just filled out they do not notice this. However, as has been demonstrated many times before, people do not have perfect access to observe and assess their attitudes through introspection (Nisbett & Wilson, 1977; Gilovich, 1991; Dennett, 1987; 1991; Johansson et al. 2005). Instead, what seems to be going on is that when participants are exposed to the survey answers and asked to explain these, they often infer what their reasons must have been given all the available evidence – in this case their (putative) rating on the scale. This self-perception framework (e.g. Bem, 1965; Carruthers, 2011) will be further explored below in relation to the shifts in attitudes observed as a consequence of accepting the choice blindness manipulation.

That participants can be lead to falsely believe that they have stated an attitude, and then explain this as if it was their very own, highlights the post-hoc nature of reasoning. In these cases, the verbal reports that are given are strongly anchored in the belief that the participant has stated this attitude. But given the experimental setup, we know that these reports are to some extent confabulations.

2. Confabulation

Confabulation has been defined as 'unintentional lying' and described as something people do when they subconsciously make up stories to conceal memory gaps (Bonhoeffer, 1904). Historically, confabulation research has focused primarily on psychiatric disorders and brain damage (Hirstein, 2005) and research on everyday non-clinical confabulation has been scarce (Hirstein, 2009). However, the relevance of confabulatory reasoning in everyday decision making

has been highlighted in some prominent research paradigms (French, Garry & Loftus, 2009). For example, Elizabeth Loftus' work on false memories and eyewitnesses' recollection of past events shows that peoples' memories are highly susceptible to external input such as descriptions of the event, images, interview techniques etc. (e.g. Loftus & Zanni, 1975; Loftus & Hoffman, 1989). The participants in these studies were found to falsely "remember" and confabulate about things that never happened. Importantly, the consequences can be detrimental as false memory have been observed in studies on eyewitness testimonies (Sagana, Sauerland & Merckelbach, 2015), decisions in criminal line-ups (Sagana, Sauerland & Merckelbach, 2015), and reported memories of sexual trauma (Loftus, Garry & Feldman, 1994), etc.

The false memory studies have interesting parallels to choice blindness research, as both include false information and confabulation as key components. In choice blindness experiments, participants express arguments which explain a response they did not make. What is puzzling is that the confabulations seem to be qualitatively equal to the 'veridical' arguments produced in non-manipulated trials. For example, in the analysis of the verbal reports in Hall, Johansson and Strandberg (2012) there was no (discernable) difference between reports from manipulated and control trials when independent raters estimated what the participants' rating must have been, purely based on the arguments they gave. Similarly, Johansson and colleagues (2005; 2006) found no difference when comparing both manipulated and non-manipulated reports using both qualitative and quantitative language analysis techniques. In paper 2 in the current thesis, participants were asked to explain their overall summary score after comparing two presidential candidates on various leadership traits, and independent raters found that both groups (experimental and control) justified their score to a similar degree. More anecdotally, during the debriefing and post-experimental interviews participants often express surprise, shock, amusement or confusion when being told that the two-minute argumentation they just had for example about tax on gasoline prices was in fact the based on a response directly opposite of what they had originally answered.

Below are two confabulation examples adopted from paper 4. When looking at the original ratings in these cases, participants had strong attitudes in favor of these issues. Yet when looking only at the verbal report it becomes obvious that what they say is in line with the manipulated ratings. Their explanations are perfectly valid, and without knowing about the manipulation there would be no

reasons from a listener's perspective to doubt that this was what the participants truly thought.

The first person gave an original rating of 93%, meaning she initially strongly agreed with the statement. Her response was then manipulated to 15%, indicated she instead disagreed. This is her explanation:

> "So the next question is a tax on the rich should be reinstated, and there I answered about 15% which means I don't agree at all with that, and that's because… well I was thinking that just because you earn more money you shouldn't get punished for that, because often the reason you earn more is because you've studied a lot and invested a lot of money to get where you are today. So in one way you should rather get rewarded. And I believe, but maybe this is just me, that if you have more money you're also willing to share more in other ways."

The second person gave an original rating of 85%, which was then manipulated to 10%, and this is how he explained it:

> "Ok, it should be legal to download copyright-protected material from the Internet for personal use, and I answered 10% so I don't agree with that. The thing is that I'm downloading stuff all the time, so maybe it's strange to say that I'm against free download. However, when I think about all the people working in the entertainment industry it is obvious that I contribute to their income loss and then it feels really wrong. So from that perspective... I don't know... I feel very split."

These examples illustrate that the accepted false feedback did not only influence participants' perceived attitude responses, but also their underlying arguments for having made those responses. This highlights the flexibility in peoples' attitudes, but also how difficult it is to trust verbal reports. Without using the choice blindness method as a wedge, it would not be possible to study this form of attitude flexibility and confabulation.

3. Downstream effects and lasting attitude change

But the confabulations are not only affecting the most immediate attitude responses, they can also lead to significant downstream effects. In paper 1 participants' voting intentions were shifted in the false feedback direction, and in

paper 2 participants justified a summary score indicating that they rated two presidential candidates equally in terms of leadership. In paper 3 participants' attitudes – when measured both 5 minutes after the experiment as well as one week later – were changed.

So how can we better understand the mechanisms behind the downstream effects of accepting and confabulating about a manipulated response?

First of all, it shows that the confabulation was not just 'empty talk' and that this act can manifest itself in real changes, shaping both future verbal and non-verbal behavior. As such, confabulations are more than momentary glitches in the participants' self-awareness (Bortolotti & Sullivan-Bissett, 2019; Bortolotti, 2018). The effect of confabulation was particularly striking in paper 3, in which participants in the confabulation condition exhibited a larger attitude change, indicating that the mere act of confabulating increased the effect of choice blindness. Somewhat puzzling, there was no difference in attitude change as a function of confabulation length (measured in seconds). That is, long and short confabulations lead to the same amount of attitude change. This particular finding indicate that the effect of confabulation does not necessarily stem from the quantitative production of arguments, which has been indicated in other attitude change paradigms (e.g. Janis & King, 1954; 1956; Broockman & Kalla, 2016; Clarkson, Tormala & Leone, 2011; Barden & Tormala, 2014). Instead, in this context, it seems that just verbally committing to the attitude accounts for much of the difference in attitude change between the two conditions. Related to this, we also found that participants who started to confabulate, but then changed their minds and corrected the manipulation, still exhibited some attitude change. This was not found in participants that immediately corrected, indicating that even small and seemingly innocuous amounts of confabulation can influence future attitudes expressions.

How can we begin to understand these findings theoretically? Below they will be discussed from the perspective of various accounts of self-generated attitude change, meaning attitude changes that occur without direct persuasion or influence from an external agent.

Self-generated attitude change

Historically, there have been two main strands of self-generated attitude change paradigms. The first is cognitive dissonance theory, one of the most widely used

and popular models of attitude change (Brehm, 1956; Festinger, 1957; Harmon-Jones & Mills, 2019). Dissonance theory suggests that cognitive inconsistency generates an aversive psychological state that promotes regulation, which comes mainly through a change of attitudes or behaviors. Dissonance theory could potentially be used to first explain why some manipulations are accepted whereas other corrected by the participants, and then how accepting the false feedback can lead to the lasting attitude changes found in paper 3. The lines of reasoning would be that when a participant have an attitude A but is asked to explain why she answered in a way that indicates attitude B, she will experience dissonance and resolve this by either adjusting the reasoning and perceived attitude in line with the manipulation (accept), or change the response in line with their original attitude and its underlying reasons (correct). Presumably, higher degrees of dissonance would then predict if a trial is detected or not. However, in our experiments, we have no reason to believe that the participants, in every manipulated trial, experience a psychologically aversive state that would explain the two types of behavior. This is not something that has come up during the hundreds of experimental interviews, or in the debriefing discussions afterwards. In relation to accepted manipulated trials, there is simply no data that would indicate a widespread negative experience that the participants need to resolve.

The other popular model of self-generated attitude change is the self-perception theory, which suggests that people adjust their attitude to their own behavior through observations and self-inference (Bem, 1965; 1967; 1972; Carruthers, 2011; 2009). Skinner used self-perception as explanatory framework in his research on interpersonal perception, and suggested that people are trained to understand and describe the internal states of other people in their social environments (Skinner, 1953; 1957). Skinner argued that while we cannot directly observe other peoples' cognitive processes, we are capable of inferring them from external cues, and talk about these in everyday conversations. Self-perception was then introduced as an alternative explanation to the attitude changes observed in the dissonance experiments (Bem, 1965). Bem expanded on Skinner's ideas to also include an inferential self-perception framework that could be used to understand and explain how attitudes changes when we make observations about our own behavior. These self-observations are similar to how we make conclusions about other peoples' attitudes based on how they behave. Because we are trained to interpret other peoples' attitudes based on their behavior, we can use the same mechanisms to infer our own attitudes as well.

Recent versions of the self-perception theory, such as the interpretive sensory-access theory (ISA, Carruthers, 2011) suggest that the interpretative mind-reading system is strongly influenced by sensory information and working memory, and is fed with input such as sense of agency or authority, auditory feedback, visual imagery etc. According to Carruthers, this is what gives the self-observations their sense if introspection. Other extensions of the self-perception theory suggest that the interpretive system should not be seen as mind-reading, but rather as a mind-shaping (Zawidski, 2013; 2008). By this view, the self-perception's main task is not to let us reflect on of past behavior, but rather to prepare our future behavior for social interactions.

Although the self-perception theory may not explain every aspect of choice blindness, it is the theory I believe can be best used to understand what differentiates correction from acceptance, and how accepting participants come to change their attitudes in line with the false feedback. Presumably, in the corrected trials the participants have some memory of their previous action. Bem argues that when the information from internal cues is weak, ambiguous, or uninterpretable, the individual is functionally in the same position as an outside observer of his behavior. An observer who, necessarily, must rely upon those same external cues to infer attitudes by asking: *"what must my attitude be if I am willing to behave in this fashion in this situation?"* (Bem, 1970, pp. 29). That is, if the internal representation of the expressed attitude is stronger than the external information from the rating, participants correct. However, if the internal representation is weaker than the context then, the rating is used as evidence of what the attitude must have been. Perhaps, participants can experience themselves as having a strong attitude, but that this is disrupted when they get confronted with the external evidence. Or maybe they do not store that process, and rely on the external world to be stable and then provide them with the evidence of what they must have thought before. That is, the strength of the context is a factor, not just the strength of the internal signal. At that point, participants will infer the reasons behind the attitude; reasons that, maybe because they are verbalized and committed to, shapes future expressions of that attitude. As I see it, one of the clearest theoretical contributions of these studies is that the false feedback induced by choice blindness creates a self-perception situation where the participants truly believe the manipulated attitudes to be their own, thus creating much stronger grounds for consequential self-inference. This should not be interpreted as an irrational, or worse, even pathological, process, but instead as a reasonable

inferential response to a peculiar array of evidence. From this standpoint, the difference between the Acknowledge and the Confabulation conditions in paper 3 is one of degree, where confabulation simply adds another layer of evidence to the self-inferences. However, more speculatively, some self-perception theories have suggested that there might be a special relationship between attitudes and first-person authority, such that attitudes we endorse also creates a special sense of agency or ownership of that attitude (see Carruthers, 2011; Martin & Pacherie, 2013; Moran, 2001). Potentially, the inference induced by choice blindness might also include our beliefs and expectations about other people, and their reactions to our opinions — that is, part of the difference between the two conditions might reside in the confabulations functioning as a public commitment (as has been explored in the literature on conversational norms, Brandom, 1994; Grice, 1975).

4. Determining factors of acceptance and correction

To fully understand choice blindness as a phenomenon, we also need to start looking at possible factors that may determine if a manipulated trial is corrected or not. A manipulated trial was categorized as 'corrected' if participants explicitly stated that they wanted to change their response, or if they in any way expressed that a response did not correspond with their views, or indicated that something was wrong. In such a situation, the experimenter would tell them that they could change their response if they wanted to, after which they could base their explanation on that response instead (in the online condition in paper 2, the participants were instructed to look over the ratings again to see if they wanted to change any of them). In all experiments, correction was operationalized as actively changing a manipulated response back towards the original position.

When looking for factors that may influence if a manipulated trial is corrected or not, we can begin by dividing these into three broader categories: individual-, attitude-specific-, and contextual factors. The individual- and attitude-specific factors are what we explicitly measures in these studies. Individual factors are for example demographical information (age, gender, education etc.), cognitive style (e.g. reflective thinking), political awareness (e.g. political interest, political involvement). The attitude-specific factors are for example the strength of the

attitude (sometimes called extremity or intensity), meta-attitudinal judgments (so called meta-attitudes, such as confidence and importance). Contextual factors are things like the medium and the design of the experiment, such as paper versus digital survey, face-to-face versus online interaction, and written versus spoken confabulation. We have not systematically studied the effect of the contextual factors, but it is very likely that some of them would have an effect. It is possible that yet other factors are involved in determining what makes people correct or accept as well, however here I present a first synthesis which can then be expanded in future research.

No demographical factors such as age or education are determining whether a participant corrects or accepts the false feedback in these studies. Taken together, only one individual factor has been found to predict correction, and this is the score to the Cognitive Reflection Test (CRT; Frederick, 2005). None of the two cognitive styles (Preference for consistency and Need for cognition) contributed to predict correction, and neither did any of the political awareness factors. Looking at attitude-specific factors, the strength of the stated attitude correlated with correction, however none of the two meta-attitude constructs did. While we have not yet made any extensive comparisons between different contextual factors, this may affect participants and dispose whether they correct or accept the false feedback. The most noteworthy findings regarding these three categories will be discussed below.

Estimation of attitude strength

In both survey- and folk psychology (not to mention political science, consumer panels etc.) it is assumed that the strength of an attitude can be estimated through self-reported measures on a scale. This general concept is also prevalent in democratic societies, in which the people's voice is often mediated via polls. Yet, every so often someone suggests that we should consider abandoning the whole attitude concept (e.g. Converse, 1964; Wicker, 1971; Bassili, 2014). The reason for this often boils down to the view that attitude strength is both deceptively easy and impossible to measure. Easy, in that all that is needed is to ask people to numerically indicate their attitudes on a bi-directional scale. Impossible, in that we do not know what we just measured (Bassili, 2014).

In the attitude strength literature, it is common to think of the attitude as falling on a continuum based on an object-evaluation association (Fazio, 1995).

By this theory, the attitude is represented in memory as an association between an object and a summary evaluation of that object. The strength of the object-evaluation association determines the accessibility of the attitude, meaning the likelihood that the summary evaluation will come to mind in presence of the attitude object. Fazio's theory is popular since it makes it easy to conceptualize how an attitude's predictability, stability, and guidance on thought and behavior increases as the attitude moves outwards on the continuum, and decreases as it moves inwards towards the middle.

As stated above, in the studies reported in this thesis, the strength of the stated attitude correlates with correction. That is: the closer to the endpoint of the scale participants put their ratings, the more likely they are to correct them. From a common sense perspective, it would have been extremely surprising if this relationship was not found. We know that people can have strong and stable attitudes, and something of this must be reflected when they estimate the strength of their own attitude as "strong" on a scale. If you could turn the pope into an atheist by switching his answers on a questionnaire, the concept of a measurable "attitude" would never have emerged in the first place.

But what is surprising is that many participants with strong or extreme attitudes still fail to notice the switch in our experiments, and that many more fail to correct attitudes stated in the mid- to lower range of the scale. In addition, regardless of the strength of the attitude, the participants still experience the attitude stated as their own, and that it is genuinely representative of how they think about the issue at hand. And in terms of determinants of correction, if the same participant corrects an answer with an attitude strength of 70, but accepts a manipulation of an attitude with a strength of 80, what was the difference between these two trials? Whatever made them correct one but reject the other cannot have been just the strength of the stated attitude. And perhaps the fact that the participant did not correct the trial at 80 say something new about the strength of that attitude, if strength is understood as tendency to act in certain ways in the future, etc. Susceptibility to choice blindness, in the context of a survey, could then be considered an alternative or additive measure of attitude strength, a measure that captures something that eludes the subjective numerical rating of the attitude.

Future research could more systematically investigate this possible relationship. For example, by combining data from multiple previous choice blindness experiments using attitudes and scales we can start mapping the potential patterns between ratings, properties of the manipulation (such as how long the

manipulation is, whether it crosses the mid-line etc.), and whether the false feedback is accepted or corrected. It would also be possible to design studies with behavioral outcomes and determine the relationship between choice blindness and measures of attitude strength by how (if at all) it affects behavior.

Thus, the relationship between attitude strength and correction both supports and creates a problem for the everyday as well as for the theoretical concept of attitude strength. As such, it joins the past decades of empirical studies, meta-analyses and factor analysis that indicates that attitude strength is an extremely complex construct (if it exists at all; e.g. Bassili, 2014; Bassili & Brown, 2005; Schwarz & Bohner, 2001).

To increase the predictive capacity of measured attitude strength, multiple other measures and constructs of attitude strength has been developed. One category, which is of extra interest to the survey community, involves meta-attitudes. These are simple, easy-to-administer, self-report measures that address the psychological properties of an attitude. For example, if a person first state an attitude and then state how important that attitude (or issue) is to her, another dimension of information regarding that person's attitude is added to the mix (Krosnick, 1990). That is, the meta-attitudes' main task is to go beyond the stated attitude and retrieve more information that can be used to estimate the strength of the attitude. Given that attitude strength is commonly defined as the attitude's stability, resistance to persuasion / contextual influence, and guidance on thought and behavior, meta-attitudes should be able to predict these factors beyond the stated attitude (otherwise, it is redundant to add the extra measures, Visser, Krosnick & Simmons, 2003).

In paper 4, we included two meta-attitudinal constructs that has received much attention in the attitude strength literature: centrality and commitment. Each of these two meta-attitudes consisted of three items, one example of a centrality item was "how important is this issue to you?" and one example of a commitment item was "how certain are you about your attitude towards this issue?" Thus, centrality tapped more into concepts involving the actual issue, and commitment more towards how much conviction participants had of the attitude. However, none of these two measures contributed to the prediction of likelihood to correct the false feedback.

One reason why the meta-attitudes did not contribute to our understanding of correction is because they did not add anything beyond the stated attitude. That is, although they may have predicted some of the correction individually, the

stated attitude explained more of the correction variance than the meta-attitudes. This is noteworthy given that the meta-attitude constructs included multiple measurements, including confidence, which is a very popular measurement in psychology and cognitive science research (e.g. Petty, Brino & Tormala, 2002; Gwinn & Krajbich, 2020; Berger & Mitchell, 1989; Berger, 1992; Shanker-Krishnan & Smith, 1998). What our results indicate is that meta-attitudes are likely inferred from the stated attitude and may add very little information beyond that (see also Bassili, 2014; Wood, Rhodes & Biek, 1995 for similar sentiments).

Individual difference and cognitive style

It is also possible that individual differences may account for some of the variation in correction rate. In paper 3, correction was correlated with participants' score on the cognitive reflection test (CRT; Frederick, 2005), meaning that participants with a higher score were more likely to correct than those with a lower score. The CRT consist of three short reasoning tasks which probes peoples' capacity to override intuitive (yet wrong) gut responses and instead use more careful analytic thinking and arrive at the right answers. The CRT is interesting because it is a performance based measure and thereby not reliant on the participants' abilities to report on their own personalities and thinking styles (Pennycook, Ross, Koehler & Fugelsang, 2017). The CRT has been tested in multiple contexts relevant to research on political attitudes. For example, people with lower CRT tend to be more sensitive to various biases (Toplak, West & Stanovich, 2011), and they are also more likely to believe in fake news (Pennycook & Rand, 2018). The correlation between CRT and correction rate is interesting as it indicates that correction to some extent is linked to depth of information processing. However, it is unclear if the connection to CRT means that "more careful analytic thinking" leads to more stable set of attitudes, or that it makes the scale positions or ratings more memorable (or a combination of these two possibilities).

In paper 4, two self-report measures of cognitive style were also tested: Need for cognition and Preference for consistency. Need for cognition (NC; Carioppo, Petty & Kao, 1984; see Petty, Brinol, Loersch & McCaslin, 2009 for an overview) has been frequently used in attitude research, where it has been found that people with high NC tend to have attitudes that are more resistant to persuasion and context effects compared to people with low NC. This is, according to the researchers, because high NC individuals tend to form attitudes based on

deliberation and not intuition. Given this assumption, it is reasonable to expect that NC would correlate with correction rate. However, in paper 4 high NC participants did not correct more than low NC participants. Preference for consistency (PFC; Cialdini, Trost & Newsom 1995) measures individuals' propensity to behave consistently and their sensitivity to social influence. PFC is also associated with participants' tendency to change their attitudes due to social pressure or external demand (Bator & Cialdini, 2006). If there was an element of social or contextual demand in the choice blindness situation, we would expect to see a connection between PFC and individual correction rate. However, there was no correlation between the PFC score and correction. This can be seen as yet another piece of evidence that the choice blindness phenomenon is not due to a demand effect implicit in the experimental situation.

But as with NC, it could also be seen as evidence that these self-reported measures of cognitive style may not be as sensitive as is commonly assumed. If a person is defined as someone who has attitudes that are more resistant to persuasion and contextual effects, they really ought to have a higher degree of correction when their stated attitudes are reversed just a few minutes after having stated them.

At an individual level, it may also be that participants with an accentuated interest in politics tend to form more elaborated political attitudes, which in turn would lead to higher correction rate. This is both intuitive and in line with research on public opinion (Zaller, 1992). Further, ideology and political positions often both polarizes and biases peoples' political reasoning, and motivates them to defend their views at great length (Haidt, 2012; Kahan, 2013; Taber & Lodge, 2013). In the literature on public opinion, political awareness is often thought of as the most important predisposition when forming strong and resilient political attitudes (Zaller, 1992). Because of this, several 'political awareness' measures have been applied in all choice blindness experiments involving politics. In paper 1, participants rated how confident they felt about their political opinions and in paper 4 how interested they were in politics. In papers 1, 3 and 4 participants also answered whether or not they were involved in any political party or organization. However, the only study finding any connection between political awareness and level of correction is Hall, Johansson & Strandberg (2012), in which participants that rated themselves as politically active corrected more in one of two conditions. Thus, there seems to be little relationship between these political awareness variables and correction. Some of

the participants has even been politicians or political commentators, yet still failed to correct the manipulations. Further, there seems to be no relationship between ideology and susceptibility to choice blindness. In paper 2, there was no difference in correction rates between participants that favored Clinton and those that favored Trump, and this was also the case for supporters of opposing coalitions in Sweden (paper 1) and parties in Argentina (Rieznik et al. 2017). These results are noteworthy given that most studies took place in close proximity to real elections, and the questions were always chosen to reflect the most important issues of the time.

Taken together, it is interesting – given all prior research in political psychology and personality research – that we find almost no pattern of choice blindness at an individual level. The reason for this could be the same as why there was no relationship between correction and the meta-attitudes. Asking people about their political interest is essentially a meta-attitude, just at a more global level. Thus, even though political interest is measured on an individual level it should not necessarily add anything beyond each of the stated attitudes.

Medium and design

In the four studies presented in this thesis several variations in design, medium, stimulus material, and manipulation outcome were used. Although study format and design has received little attention – in particular in the individual papers presenter here – this is clearly a factor that may matter in terms of how participants react to the manipulations (e.g. Johansson, Hall, Gulz, Haake & Watanabe, 2007). If we take paper 2 as an example, in experiment 1 we used a paper survey and interviewed participants in the streets, and in paper 2 we ran a similar version of the same experiment online. Looking at the correction rate, those two experiments had different results: in experiment 1 the correction rate was only 12% and in experiment 2 it was 41%. Given the slight variation in both design and operationalization, it is difficult to isolate the exact factors contributing to this difference. For example, experiment 2 had more manipulations, written instead of verbal rationalizations, and included more detailed instructions on how to pay attention to the ratings. However, there are two potentially relevant factors that could be discussed.

Firstly, the default trust we place in the physical world to remain constant and reliable. This means that constant high-level monitoring of the constancy of the

external world might not be needed unless the context implies something different (see O'Regan, 1992). Thus, one factor is the prior (im)plausibility of the manipulation. In Experiment 1, the manipulations were performed using a magic trick, which is extremely improbable in the context of a political survey. Likely none of the participants had ever filled out a pen-and-paper survey that was altered mid-air. In contrast, in Experiment 2, even though we attempted to replicate the general procedure of the original trick, participants were faced with a far less magical procedure. People are familiar with malfunctioning computer programs and websites, and thus our participants may have had a different level of monitoring when tracking their external responses.

Secondly, the dimension of social trust may be relevant to the choice blindness experiments. With this I do not mean that the participants were aware of the manipulation but refrained from telling us, but rather that factors like tacit cultural norms, status relations, concerns for communicative (Gricean) relevance and so on may influence the level of monitoring the participants applied in this context. Using experiment 1 and experiment 2 in paper 2 again as examples, the presence and absence of an experimenter is an important difference that could have influenced the correction rate.

Another potentially relevant factor could be the difference between verbally explaining (experiment 1) versus silently revising (experiment 2) the responses. While participants in experiment 2 were also confronted with the manipulations, they did not have to engage in the mental task of generating arguments for the new position. One might expect this additional reasoning process to generate more corrections, presumably by helping participants think more deeply about the issue and discovering that they do not agree with the manipulations. On the other hand, if reasoning serves not as attitudinal fact-checking but as a way for participants to further commit to and defend their attitudes, the verbalization process might lead to fewer corrections (as we saw in paper 2). A final plausible factor – which is relevant to all studies concerning political attitudes – regards the saliency of the attitudes. As an illustration, experiment 2 was conducted closer to the election compared to experiment 1, and it is possible that more of the participants in experiment 2 had firmly decided who they would vote for.

Even if this aspect of choice blindness research has not been the focus of my thesis, it is clear that a lot of additional research is needed to differentiate the relative effect of all these potential factors.

Part IV

Looking ahead

In summary

In papers 1—4 it is shown that choice blindness applies to political attitudes, both involving salient political issues as well as presidential candidates' character traits. In papers 1—3 significant downstream effects are found, such as lasting attitude change. In all studies, and in particular in paper 3 and 4, some possible factors that may (or may not) determine whether participants accept or correct the false feedback is studied and discussed.

In short, the findings presented in this thesis demonstrate attitude flexibility as a consequence of accepting false feedback about previously stated attitudes, and how confabulatory reasoning facilitates shifts away from these original positions. These findings were obtained studying political attitudes, a domain of central importance to public life. Importantly, the attitude flexibility occurs outside of the participants' own awareness, and the only way we can know about it is because we control the experimental situation. As far as the participants are concerned, the [accepted manipulated] response is what they intended to answer, and the arguments they give are the true reasons behind the response. Further, the attitude flexibility can have downstream affects, as accepting and arguing for a manipulated response can influence participants' voting intention, and lead to attitude changes lasting at least one week.

Ideas for future research

One thing that has stood out in all four studies is the influential power of confabulation. As such, I think confabulation, in particular experimentally induced everyday confabulation, should receive more academic attention than it currently has (see Hirstein, 2005; 2009 for an abundance of research on clinical

confabulation). One direct way to explore this inquiry is by systematically transcribe, code, analyze and compare the verbal reports from all the interviews collected in already conducted choice blindness experiments. Johansson and colleagues did analyze the verbal reports of the original face study using both quantitative and qualitative analysis methods. However it was a rather small sample and the task involved a stimulus which generated quite short reports. A large-scale research effort which compares the argumentative and paralinguistic differences between 'veridical' and 'confabulated' reports would have immediate interdisciplinary value. Further, by taking data from various domains, modalities and mediums it would be possible to compare other features than strictly linguistic, such as how the experience of the situation influences the reports. One current limitation in manually classifying large quantity of verbal data is that it is very time-consuming work. Hopefully, the development of novel machine-learning and voice recognition tools that can be used to automatically transcribe and analyze interviews will open up a whole new way of working with choice blindness studies, as well as generating a revolution in cognitive science and social psychology in general.

Related to this, we found that across all studies very few participants reported that they suspected their responses had been changed, or that the manipulated rating was not their own answer. Rather, in the corrected trials, participants often experienced the manipulated response as their own but felt it did not represent their attitude towards the issue (or that they had changed their mind since their original answer). To illustrate this point, inspired by research on attribution (Jones, 1965), in paper 4 the reasons participants gave when correcting were categorized. One third self-attributed the corrections internally and gave reasons such as "I must have misread the question" or "Oh, I misinterpreted how the scale worked" etc. Another third attributed it externally, and gave reasons such as "It must have been a glitch in the survey application" etc. One third reported spontaneous change such as "I have changed my mind" or "Now when I read the question again I want to change my response". In essence, this might indicate that even corrections are based on inferential self-perception (i.e. essentially being confabulations), just with a different outcome than in the accepted trials. We also found that politically involved participants were slightly more likely to attribute the correction externally. Now, the sample size was rather small, in particular when broken up into subcategories. But the experience of correcting is another potential source of insights when trying to untangle what makes people correct as

well as how the incongruence induced by the choice blindness affects the participants.

It is also possible to step outside the choice blindness domain and into the political sciences. Table 1 shows all scientific papers published the past hundred years which included the words 'political' and either one of 'attitude', 'opinion', 'judgment', 'preference' or 'belief' in the topic.

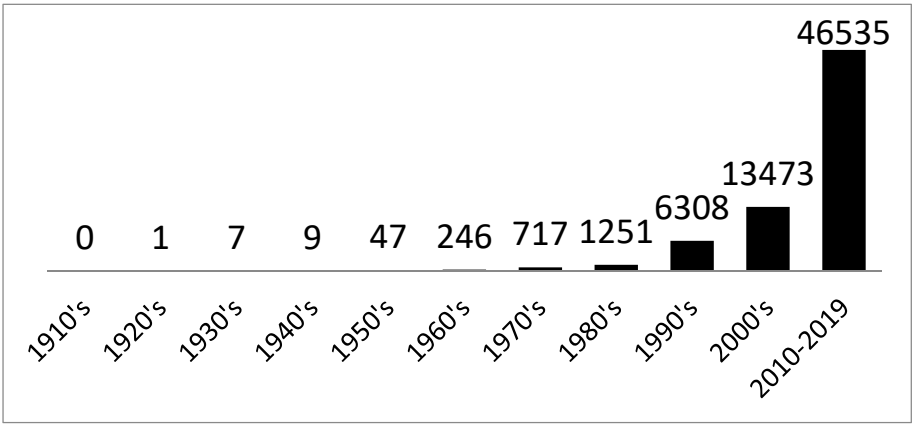

Table 1 – A bibliometric analysis in Web of Science's CORE collection database shows exponential growth in research output covering political attitudes. The x-axis shows the 10 past decades and the y-axis number of publications. In 1920's there was one publication on political attitudes and 2010's over 46,000.

Thus, there is an increased interest in research on political attitudes, and almost 70% of all papers were published the last decade. Given this increased interest in political attitude research, it is inviting to introduce a novel choice paradigm like choice blindness to the political scientists' theoretical and methodological repertoire. And there is a strange dynamic between attitude stability and flexibility which deserves more experimental attention. Political attitudes has exhibited remarkable stability when measured over years and even decades (Gerber, Huber, Doherty, & Dowling, 2011; Hatemi et al., 2009; Hooghe & Wilkenfeld, 2008; Lewis, 2018; Sears & Funk, 1990; Alwin, 1994; Sears, 1983). This stability is particularly prevalent at an aggregate level, and for example election outcomes are often accurately predicted (Sohlberg, Gilljam & Martinsson, 2017; Silver, 2018; but see research showing that asking people what their friends will vote for is a better predictor of election outcome than asking people what they will vote for themselves! Galesic, Bruine de Bruin, Dumas, Kapteyn, Darling & Meijer, 2018;

Bruine de Bruin, Galesic, Bååth, Bresser, Hall, Johansson, Strandberg & van Soest, in press). This general view is also corroborated by our own data. In paper 3, non-manipulated responses and manipulated responses that was instantly corrected, were rated on the same position throughout the entire experiment. Yet, the accepted trials of all the experiments presented in this thesis indicate that what appears to be the same set of attitudes can exhibit flexibility when self-perceptive and argumentative processes interact with situational changes. This shows how attitudes can exhibit longitudinal stability while at the same time being malleable in everyday life. And this result indicates that it does not have to be a conflict between attitude stability and flexibility. Instead, it opens up new and interesting questions regarding the nature of attitudes, and how they are continually regulated by our environments, social interactions, and perceived behavior.

Concluding remarks

I like to end by concluding that despite of (or possibly thanks to) all the hard and cognitively draining work, long days and late hours, feelings of frustration and hopelessness, I have enjoyed [almost] every moment of this endeavor. In particular, I have enjoyed working with such smart, innovative and genuinely curious collaborators from Lund University and McGill University. The melding of our different personalities, habits and mindsets have always rendered interesting research output. I am also particularly appreciative (and proud) that I so often have stepped out of the lab, to a large extent resisted the convenience of the online, and taken the more "anthropological" route by collecting data in the streets. Doing the hard, time-consuming, sometimes inconvenient, nitty-gritty work has perhaps more than anything shaped who I am today.

Bibliography

Aardema, F., Johansson, P., Hall, L., Paradisis, S. M., Zidani, M., & Roberts, S. (2014). Choice blindness, confabulatory introspection, and obsessive-compulsive symptoms: A new area of investigation. *International Journal of Cognitive Therapy, 7*, 83–102.

Alwin, D. (1994). Aging, personality, and social change: The stability of individual differences over the adult life span. In D. L. Featherman, R. M. Lerner, & M. Perlmutter (Eds.), *Life-span development and behavior* (pp. 135–185). Hillsdale, NJ: Erlbaum.

Barden, J., & Tormala, Z. L. (2014). Elaboration and attitude strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass, 8*, 17–29.

Bassili, J. N. (2014). Attitude strength. In W. D. Crano & R. Prislin (Eds.) *Attitudes and attitude change* (pp. 237-286). Taylor & Francis Group, New York: NY.

Bassili, J. N., & Brown, R. D. (2005). Implicit and explicit processes attitudes: Research challenges and theory. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.) *The handbook of attitudes* (pp. 543-574). Mahwah, NJ: Erlbaum.

Bator, R. J., & Cialdini, R. B. (2006). The nature of consistency motivation: Consistency, aconsistency, and anticonsistency in a dissonance paradigm. *Social Influence, 1*, 208–233.

Bem, D. J. (1965). An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology, 1*(3), 199-218.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology, 6*, 1-62.

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

Bem, D. J., and McConnell, K. H. (1970). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology 14*, 23–31.

Berger, I. E. (1992). The nature of attitude accessibility and attitude confidence: A triangulated experiment. *Journal of Consumer Psychology, 1*(2), 103-123.

Berger, I. E., & Mitchel, A. A. (1989). The effect of advertising on attitude accessibility, attitude confidence, and the attitude-behavior relationship. *Journal of Consumer Research, 16*(3), 269-279.

Bowyers, T. (1928). A message from nowhere. *Linking Ring Magazine*.

Brandom, R. (1994). *Making it explicit*. Cambridge, MA: Harvard university press.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology, 52*(3), 384.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352*, 220 –224.

Bonhoeffer, K. (1904). Der Korsakowische Symptomcomplex in seinen beziehungen zu den verschieden krankheitsformen. *Allgemeine Zeitschrift für Psychiatrie, 61*, 744-752.

Bortolotti, L. (2018). Stranger than fiction: Costs and benefits of everyday confabulation. *Review of Philosophy and Psychology, 9*(2), 227–249.

Bortolotti, L., & Sullivan-Bissett, E. (2019). Is choice blindness a case of self-ignorance? *Synthese.*

Bruine de Bruin, W., Galesic, M., Bååth, R., de Bresser, J., Hall, L., Johansson, P., Strandberg, T., & van Soest, A. (in press). Asking about social circles improves election predictions even with many political parties.

Busch, J. T. A., & Lagare, C. H. (2019). Using data to solve problems: Children reason flexibly in response to different kinds of evidence. *Journal of Experimental Child Psychology, 183*, 172-188.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford, UK: Oxford University Press.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*, 121–182.

Chater, N., Johansson, P., & Hall, L. (2011). The non-existence of risk attitude. Frontiers in Psychology, 2, 1–3.

Cheung T., Junghans A. F., Dijksterhuis G. B., Kroese F., Johansson P., Hall

L. & De Ridder D. T. D. (2015). Consumers' choice blindness to ingredient information. *Appetite,* doi:10.1016/j.appet.2015.09.022.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology, 69*(2), 318-328.

Clarkson, J. J., Tormala, Z. L., & Leone, C. (2011). A self-validation perspective on the mere thought effect. *Journal of Experimental Social Psychology, 47*, 449 – 454.

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2018). (Choice) blind justice: Legal implications of the choice blindness phenomenon. *University of California, Irvine Law Review 8*, 85.

Converse, P. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206 –261). New York, NY: The Free Press.

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

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

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

Dennett, D. C. (1991c). Two contrasts: Folk craft versus folk science and belief versus opinion. In J. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135–148). Cambridge: Cambridge University Press

Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 247–282). Lawrence Erlbaum Associates, Inc.

Fazio, R. H. (2007). Attitudes as Object-Evaluation Associations of Varying Strength. *Social Cognition, 25*(5), 603–637

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*, 25– 42.

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

French, L., Garry, M., & Loftus, E. (2009). False memories: A kind of confabulation in non-clinical subjects. In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy* (pp. 33– 66). Oxford, UK: Oxford University Press.

Galesic, M., Bruine de Bruin, W., Dumas, M., Kapteyn, A., Darling, J.E., & Meijer, E. (2018). Asking about social circles improves election predictions. *Nature Human Behaviour, 2*, 187-193.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science, 14*, 265–287.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. Free Press.

Gleckman H. Character vs policy in the 2016 presidential election. *Forbes*. 2016 Nov 1 [cited 2019 Sep 6]

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5). 620-629.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Studies in syntax and semantics III: Speech acts* (pp. 41–58) New York, NY: Academic Press.

Gwinn, R., & Krajbich, I. (2020). Attitudes and attention. *Journal of Experimental Social Psychology*, vol. 86.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLOS ONE, 7*:e45457.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117*, 54 – 61.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814 – 834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* Pantheon Book, New York: NY.

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (pp. 3–24). American Psychological Association.

Hatemi, P. K., Funk, C. L., Medland, S. E., Maes, H. M., Silberg, J. L., Martin, N. G., & Eaves, L. J. (2009). Genetic and environmental transmission of political attitudes over a life time. *The Journal of Politics, 71*, 1141–1156.

Haugtvedt, C. P., & Petty, R. E. (1992). Personality and persuasion: need for cognition moderates the persistence and resistance of attitude change. *Journal of Personality and Social Psychology, 63*(2), 308-319.

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

Hirstein, W. (2009). *Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy*. Oxford, UK: Oxford University Press.

Hooghe, M., & Wilkenfeld, B. (2008). The stability of political attitudes and behaviors across adolescence and early adulthood: A comparison of survey data on

adolescents and young adults in eight countries. *Journal of Youth and Adolescence, 37*, 155–167.

Jack, A., & Roepstorff, A. (2004). Trusting the subject II: The use of introspective evidence in cognitive science. *Imprint Academic*, pp. 5-22

Janis, I. L., & King, B. T. (1954). The influence of role playing on opinion change. *Journal of Abnormal Psychology, 49*, 211–218.

Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia, 51*, 142–155.

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.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15*, 673– 692.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making, 27*, 281–289.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about Telling More Than We Can Know. *Consciousness and Cognition, 15*, 673– 692.

Johansson, P., Hall, L., Gulz, A., Haake, M., & Watanabe, K. (2007). Choice blindness and trust in the virtual world. *Technical Report of Institute of Electronics, Information and Communication Engineers – Human Information Processing, 107*(60), 83-86.

Jones, E. E., & Davis, K. E. (1965) From acts to dispositions: the attribution proces in social psychology, in L. Berkowitz (ed.), *Advances in Experimental social Psychology* (Volume 2, pp. 219-266), New York: Academic Press

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making, 8*, 407– 424.

King, B. T., & Janis, I. L. (1956). Comparison of the effectiveness of improvised versus non-improvised role-playing in producing opinion changes. *Human Relations, 9*, 177–186.

Krosnick, J. A. (1990). Government policy and citizen passion: A study of issue publics in contemporary America. *Political Behavior, 12*, 59-92.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 1– 24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, G. J. (2018). Early-childhood conduct problems predict Economic and political discontent in adulthood: Evidence from two large, longitudinal U. K. Cohorts. *Psychological Science, 29*, 711–722.

Loftus, E. F., Garry, M., & Feldman, J. (1994). Forgetting sexual trauma: What does it mean when 38% forget? *Journal of Consulting and Clinical Psychology, 62*(6), 1177–81.

Loftus E. F., Hoffman H. G. (1989). Misinformation and memory: the creation of new memories. *Journal of Experimental Psychology: General.*

Loftus, E., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society, 5*, 86 – 88.

McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making, 8*(5), 561–572.

Martin, J. R., & Pacherie, E. (2013). Out of nowhere: Thought insertion, ownership and context-integration. *Consciousness and Cognition, 22*, 111–122.

Merckelbach, H., Jelicic, M., Pieters, M. (2011) The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical Experimental Neuropsychology, 33*, 131–139.

Moore, J., & Haggard, P. (2006). Commentary on "How something can be said about telling more than we can know: On choice blindness and introspection." *Consciousness and Cognition, 15*(4), 693–696.

Moran, R. (2001). Authority and estrangement: An essay on self-knowledge. Princeton, NJ: Princeton University Press.

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

O'Regan, J.K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology , 46*, 461–488

Rensink, R.A.,O'Regan, J.K., & Clark,J.J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368–373.

Robinson, W. (1898). *Spirit slate writing and kindred phenomena.* Munn & co.

Pennycook, G., Ross, M. R., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24,* 1774-1784.

Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39-50.

Petty, R. E., Brinol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 318–329). The Guilford Press.

Petty, R. E., Briñol, P., & Tormala, Z. L. (2002). Thought confidence as a determinant of persuasion: The self-validation hypothesis. *Journal of Personality and Social Psychology, 82*(5), 722–741.

Rieznik, A., Moscovich, L., Frieiro, A., Figini, J., Catalano, R., Garrido, J. M., Heduan, F. A., Sigman, M., & Gonzalez, P. A. (2017). A massive choice blindness experiment on choice blindness political decisions: Confidence, confabulation, and unconscious detection of self-deception. *PLOS ONE, 14*;12(2):e0171108.

Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law, 22*, 303–314.

Sagana A., Sauerland M., Merckelbach H. (2013). Witnesses' blindness for their own facial recognition decisions: a field study. *Behavioral Science & The Law, 31*, 624–636.

Sagana A., Sauerland M., Merckelbach H. (2017). Witnesses' failure to detect covert manipulations in their written statements. *Journal of Investigative Psychology and Offender Profiling, 14*(3), 320-331

Sagana A., Sauerland M., Merckelbach H. (2014). This is the person you selected: Eyewitnesses' blindness for their own facial recognition decisions. *Applied Cognitive Psychology, 28*(5), 753-764.

Sauerland M., Sagana A., Otgaar H. (2013). Theoretical and legal issues related to choice blindness for voices. *Legal and Criminological Psychology, 18*(2), 371–381.

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 426-457). Malden, MA: Blackwell.

Sears, D. O. (1983). The persistence of early political predispositions: The roles of attitude object and life stage. *Review of Personality and Social Psychology, 4*, 79 – 116.

Sears, D. O., & Funk, D. L. (1990). Evidence of long-term persistence of adults' political predispositions. *The Journal of Politics, 61*, 1–28.

Shanker-Krishnamn, H., & Smith, R. R. (1998). The relative endurance of attitudes, confidence, and attitude-behavior consistency: The role of information source and delay. *Journal of Consumer Psychology, 7*(3), 273-298.

Skinner B.F. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.

Skinner, B. F. (1953). *Science and human behavior*. Macmillan.

Silver, N. (2018). The Polls Are All Right. *Fivethirtyeight*. 2018 May 30 [cited 2020 May 18]

Statler-Throckmorton A. Personality over policy. *Stanford Politics*. 2016 Mar 3 [cited 2019 Sep 6]

Solhlberg, J., Gilljam, M., & Martinsson, J. (2017). Determinants of polling accuracy: the effect of opt-in Internet surveys. *Journal of Elections Public Opinion and Parties 27*(2),1-15

Somerville, J., & McGowan, F. (2013) Can chocolate cure blindness? Investigating the effect of preference strength and incentives on the incidence of choice blindness. *Journal of Behavioral and Experimental, 61*(C), 1-11.

Taber, M. and Lodge, C. S. (2013). *The rationalizing voter*. New York, NY: Cambridge University Press.

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PLOS ONE, 9*, e108515.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition, 39*(7), 1275-1289.

Visser, P. S., Krosnick, J. A., & Simmons, J. P. (2003). Distinguishing the cognitive and behavioral consequences of attitude importance and certainty: A new approach to testing the common-factor hypothesis. *Journal of Experimental Social Psychology, 39*, 118-141.

Waldman P. Why the 2016 campaign may be the most personality-driven ever. *The Washington Post*. 2016 June 10 [cited 2019 Sep 6]

Wang, Y., Zhao, S., Zhang, Z., & Feng, W. (2018). Sad facial expression increase choice blindness. *Frontiers in Psychology,* doi.org/10.3389/fpsyg.2017.02300.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review, 87*(1), 105-112.

Wicker, A. W. (1971). An examination of the "other variables" explanations of attitude-behavior inconsistency. *Journal of Personality and Social Psychology, 19*, 18-30.

Wood, W., Rhodes, N., & Biek, M. (1995). Working knowledge and attitude strength: The effect of amounts and accuracy of information. In R. E. Petty & J. A. Krosnick (Eds.) *Attitude strength: Antecedents and consequences* (pp. 283-324). Mahwah, NJ: Erlbaum.

Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York, NY: Cambridge University Press.

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.

Zawidzi, T. W. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations, 11*(3), 193-210.

Paper I

How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions

Lars Hall, Thomas Strandberg, Philip Pärnamets, Betty Tärning, Andreas Lind and Petter Johansson

Abstract: Political candidates often believe they must focus their campaign efforts on a small number of swing voters open for ideological change. Based on the wisdom of opinion polls, this might seem like a good idea. But do most voters really hold their political attitudes so firmly that they are unreceptive to persuasion? We tested this premise during the most recent general election in Sweden, in which a left- and a right-wing coalition were locked in a close race. We asked our participants to state their voter intention, and presented them with a political survey of wedge issues between the two coalitions. Using a sleight-of-hand we then altered their replies to place them in the opposite political camp, and invited them to reason about their attitudes on the manipulated issues. Finally, we summarized their survey score, and asked for their voter intention again. The results showed that no more than 22% of the manipulated replies were detected, and that a full 92% of the participants accepted and endorsed our altered political survey score. Furthermore, the final voter intention question indicated that as many as 48% (±9.2%) were willing to consider a left-right coalition shift. This can be contrasted with the established polls tracking the Swedish election, which registered maximally 10% voters open for a swing. Our results indicate that political attitudes and partisan divisions can be far more flexible than what is assumed by the polls, and that people can reason about the factual issues of the campaign with considerable openness to change.

Introduction

With the proliferation of public polls from both media, political organizations, and the parties involved, European and US elections now seems to generates almost as much controversy about the polling as the candidates and issues themselves. In particular, it has become commonplace to question the scientific integrity of the polls, and view them as partisan instruments of persuasion [1]. For example, during the recent 2012 US presidential campaign many political commentators suggested the mainstream polls were based on flawed assumptions, and harbored a systematic bias that needed to be 'unskewed' [2-4]. However, in the aftermath of the election it was concluded that professional polling organizations generally did a good job of predicting the outcome (albeit underestimating the winning margin for president Obama [5]), and that independent aggregators of the polls, such as Votamatic, FiveThirtyEight, Princeton Election Consortium, or the HuffPost Pollster was particularly accurate in their calls.

But success in calling the outcome of a race on the eve of the election is only one aspect of the prediction game. More important in both understanding and running a campaign is the effort to delineate what *could* happen, to pinpoint how many voters are receptive to different messages, and open to ideological change. To use another example from the recent US presidential campaign; seven weeks before the election, a video was released of republican candidate Mitt Romney, secretly filmed during a fundraiser in Florida. In this video Romney declares that it is not his job not to worry about the 47% of Americans that pay no income tax, because they are not receptive to his campaign message. Instead, he asserts that there only are 5-10% of voters that are open to move across the partisan divide, and that those are the target demographic he needs to convince to win the election (for the relevant quotes, see Online Material S1). Independently of whether the message of the leaked tape contributed to the failure of the Romney campaign, one might legitimately ask whether it is a sound strategy to run a presidential race on the premise that maximally 10% of the electorate can be swung across party lines? Are most voters so firmly locked in their views that they are unreceptive to any attempts at persuasion, even from the concentrated effort of a billion dollar campaign machinery [6]?

Looking at the research, this seems to be the case. The most salient contrast across the political landscape in the US and the EU is the left vs. right wing

division. Despite a trend towards diminishing party affiliation among voters, partisanship across the left-right divide still holds a firm grip on the international Western electorate, and has even shown evidence of further polarization in recent years (e.g. see [7-10] for analysis relating to the condition in the US, and [11-13] for the EU perspective, see also [14, 15] for cross cultural comparisons).

We were given an opportunity to test this premise during the final stretch of the 2010 general election in Sweden. Based on our previous research on the phenomenon of choice blindness (CB [16, 17]) our hypothesis was that if we could direct the focus of our participants towards the dividing policy issues of the campaign, and away from the overarching ideological labels of the competing parties, we could use CB to demonstrate far greater flexibility in their political affiliations than what is standardly assumed.

Like in the US, the Swedish electorate is regarded as one of the most securely divided populations in the world (albeit shifted somewhat to the left compared to the US continuum). When we entered into the study, the tracking polls from commercial and government institutes were polling the Swedish electorate at about 10% undecided between the two opposing coalitions social democrats/green vs. conservatives (provided by Statistics Sweden (J. Eklund, unpublished data, 2012)), with the conventional wisdom of political science identifying very few additional voters open for a swing at the final stretch of the campaign [18-20].

Method

Participants

In total, 162 volunteers (98 female) divided in two conditions (manipulated and control) participated in the study. Ages ranged from 18 to 88 years (M=29.7, SD 14.1). We recruited our participants from various locations in the cities of Malmö and Lund in Sweden, and asked them if they wanted to fill in a questionnaire concerning their views on political issues. Participants who did not intend to vote, or who had already voted by mail were not admitted into the study. Two participants were removed due to technical problems with the manipulation process (the glued-on piece of paper did not stick and fell off during the

discussion, see procedure figure 1). All participants gave informed consent. The study was approved by the Lund University Ethics board, D.nr. 2008–2435.

Procedure and materials

We introduced ourselves as researchers from Lund University with an interest in knowing the general nature of political opinions. We emphasized that participation was fully anonymous, that we had no political agenda, and that we would not argue with or judge the participants in any way. After this, we presented the participants with an 'election compass'; a survey with salient issues from the ongoing election campaign where the left- and the right-wing coalition held opposite positions.

At the start of the questionnaire, the participants were asked to indicate how politically engaged they were (on a scale from extremely disengaged, to extremely engaged), and how certain they were in their political views (from extremely uncertain, to extremely certain). Next, they were asked to indicate the direction and certainty of their current voting intention on a 100mm bidirectional scale (from extremely certain social democrat/green, to extremely certain conservatives, with the midpoint of the scale representing undecided).

The main survey consisted of 12 salient political issues taken from the official coalition platforms where the two sides held opposing views. On the survey, the issues were phrased as statements, such as: "*Gasoline taxes should be increased*" or "*Healthcare benefits should be time limited*". We asked the participants to indicate their level of agreement with the statements on a 0-100% scale (where 0% meant absolutely disagree, and 100% absolutely agree, and the midpoint represented uncertainty/indecision). To avoid any obvious patterning of the answers on the form, the statements were formulated both in the positive and the negative (i.e. to introduce or to remove a particular policy) and counterbalanced for the left and right wing coalitions (see table 1).

Table 1 – The political issue statements.

1.	The gas tax should be increased
2.	Healthcare benefits should be time limited
3.	It should be possible to remove disruptive students from a school even against the parents or students wishes
4.	Family leave benefits reserve 2 months out of a total of 13 months for each parent. The earmarked months for each parent should be increased, to ensure more equality
5.	Employee income taxes have been lowered for the past several years through the Earned Incom Tax Credit. Incomtaxes should be lowered further
6.	The law that give the Swedish government the right to monitor email- and telephone traffic, if it suspects an external threat against Sweden, should be abolished
7.	Sweden decided in 1997 that nuclear energy should be shut down. That law should now be repealed
8.	A tax deduction for housekeeping services was established in 2007. It should be abolished
9.	Running major hospitals as private establishments should be permitted
10.	The legal age for criminal responsibility should be lowered
11.	The maximum unemployment insurance benefit is about 11 000 Swedish Kronor per month after taxes. It should be increased
12.	The wealth tax was abolished in 2007. It should be reinstated

In the neutral condition (N=47), after having rated their agreement with the 12 statements, we asked the participants to explain and justify their stance on some of the issues. When they had completed these justifications we then overlaid a color-coded semi-transparent coalition template on their answering profile, with red indicating left-wing and blue right-wing (note, these colors are inverted in US politics). In collaboration with the participants, we then tallied an aggregate 'compass score' for the right and left wing side, indicating which political coalition they favored based on the policy issues presented. We then asked the participants to explain and comment on the summary score, and as the final step of the experiment, to once again indicate the direction and strength of their voting intention for the upcoming election.

However, in the manipulated condition (N=113), while observing the participants filling out the form, we surreptitiously filled out an answer sheet identical to the one given to the participants, but created a pattern of responses supporting the opposite of their stated voting intention. Thus, if their voting intention supported the social democrat/green coalition, we made a summary compass score supporting the conservatives, and vice versa (for those that were unsure in their original voting intentions, we created an answer profile that was the opposite of their compass score). Then, before we asked the participants to discuss and justify their ratings of the individual questions, we performed a

sleight-of-hand to overlay and attach our manipulated profile on top of their original answers (see Figure 1). Consequently, when we asked the participants to discuss their answers, they were faced with an altered position supporting the opposing coalition. For example, if they previously thought the gasoline tax ought to be raised, they were now asked to explain why they had indicated it ought to be lowered.

Figure 1 – A step-by-step demonstration of the manipulation procedure. A. Participants indicate the direction and strength of their voting intention for the upcoming election, and rate to what extent they agree with 12 statements that differentiates between the two political coalitions. Meanwhile, the experimenter monitors the markings of the participants and creates an alternative answering profile favoring the opposite view. B. The experimenter hides his alternative profile under his notebook. C. When the participants have completed the questionnaire, they hand it back to the experimenter. The backside of the profile is prepared with an adhesive, and when the experimenter places the notebook over the questionnaire it attaches and occludes the section containing the original ratings. D. Next, the participants are confronted with the reversed answers, and are asked to justify the manipulated opinions. E. Then the experimenter adds a color-coded semi-transparent coalition template, and sums up which side the participants favor. F. Finally, they are asked to justify their aggregate position, and once again indicate the direction and strength of their current voting intention. See http://www.lucs.lu.se/cbp for a video illustration of the experiment.

The goal of our alterations was to bring the sum of the participants' answers securely to the opposing side. Thus, the number of altered responses we made on the mirrored profile depended on how directionally skewed the original answers were (say 11-1 vs. 7-5). In addition, there was no predetermined rule for the size of the manipulations across the scale. Instead, each manipulation was made with the intent of creating an overall believable pattern of responses on the profile (i.e.

as the level of polarization generally varied between questions, it would invite suspicion to simply move all responses the minimal distance across the midline of the scale). During the discussion, and later during the summation, if the participants realized their answers were not expressing their original opinion, they were given the opportunity to change the rating to what they instead felt appropriate. This way, our efforts at creating a coalition shift could be nullified by the number of corrections made by the participants. As in the neutral condition, after reacting to the summary score, the final step of the experiment was for the participants to once again indicate their voting intentions for the upcoming election.

After the experiment we explained the true purpose of the study to all participants, and demonstrated the procedure of the manipulation. At this point we asked whether they had suspected anything was wrong with their answers (over and above any previously registered corrections). We then interviewed the participants about how they felt about the experiment, and finally, everybody gave written consent to have their results included in the analysis. After the study, the experimenter took notes about the comments and explanations of the participants.

Results

Correction of manipulated answers

Each participant had on average 6.8 (SD=1.9) answers manipulated, with a mean manipulated distance of 35.7mm (SD=18.7) on the 100mm scale. The participants were explicitly asked to state reasons on average 4.0 (SD=1.6) of the manipulated trials, and of those were on average 0.9 (SD=1.0) answers corrected by the participants to better match their original intention (i.e. a trial-based correction rate of 22%). At an individual level, 47% of the participants did not correct any answers, while 53% corrected between 1-4 answers. For all answers classified as corrected, the participants indicated that they had misread the question, or marked the wrong end of the scale. Only a single participant expressed any suspicion that we had manipulated her profile.

The number of corrected answers were not related to gender, age, or political affiliation as defined by prior voting intention (p=n.s.). The distance being manipulated on the scale did not differ between corrected and non-corrected answers (p=n.s.). Finally, there were no differences in self-rated political engagement or in political certainty between participants who corrected no answers and participants who made one or more corrections (p=n.s.).

Endorsement of compass score

As very few manipulated issues were corrected, we were able to create a mismatch between the initial voting intention (or original compass score for the uncertain group) and the manipulated summary score for a full 92% of the participants, all of which acknowledged and endorsed the manipulated score as their own.

Change in voting intention

In order to establish if the mismatch between the initial voting intention and the manipulated compass score also influenced the participants final voting intention, we measured the change in voting intention from pre- to post-test, and classified it as a positive change if it was congruent with the manipulated compass score, and as a negative change otherwise. For example, if the participants had a (manipulated) compass score biased towards the right wing, and their voting intention shifted towards the right-wing coalition, this was classified as a positive change. For the control condition, the change between initial and final voting intention was classified as positive or negative against their unaltered compass score. Using this measure to compare the amount of change in voting intention between the manipulated and the control condition, we find that there is a very large change in the manipulated condition (M=15.9, SD=24.7) while there is virtually no change (M=1.72, SD=9.9) in the control condition (Wilcoxon Rank Sum Test, W=3857.5, $p < .00001$, r=0.35, see figure 2). In the manipulated condition, we also find that the skewness of the compass score correlates with the amount of change in voting intention, e.g. if an initially right-wing participant finds herself with a left wing aggregate score of 10 vs 2, she is likely to change her voting intention more than if the balance was 7 vs 5 (Pearson correlation, r=0.28, $p < 0.005$). As was the case with level of correction, we found no connection between gender, age, level of political engagement, overall political certainty, or

initial political affiliation, in relation to magnitude of change in voting intention (p=n.s.).

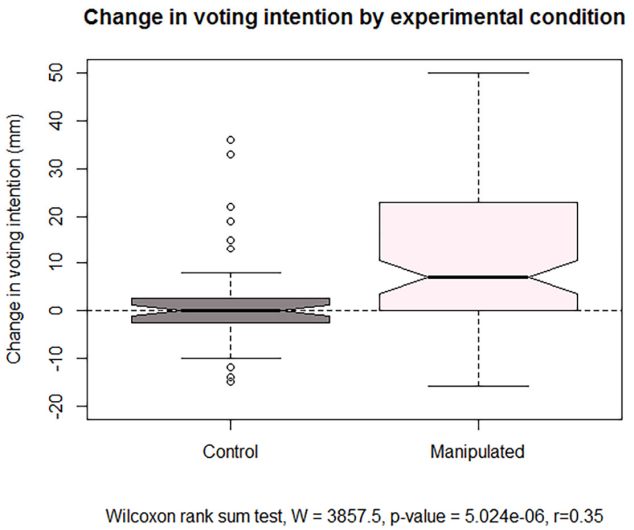

Figure 2 – Change in voting intention in the control and in the manipulated condition.

If we translate the change in voting intention to categorical political affiliation, what we find is that 10% of the participants in the manipulated condition moved across the full ideological span, and switched their voting intention from firmly right wing to firmly left wing, or in the opposite direction (with a mean movement of voting intention across the scale = 71mm, SD=30.2). A further 19% went from expressing certain coalition support (left or right), to becoming entirely undecided (M=27.2, SD=13.2), and 6% went from being undecided to having a clear voting intention (M=12.0, SD=26.9). If we add to this the 12% that were undecided both before and after the experiment, it means that 48% (±9.2%) of the participants were willing to consider a coalition shift. In addition, a further 10% of the participants recorded substantial movement in the manipulated direction, moving 20mm or more on the 100mm scale.

Excluding the initially undecided participants (as they are per definition open to change), the average certainty of the initial voting intentions of the participants was notably high (M= 37.4mm, SD= 13.45, with the 100mm bidirectional scale transformed to a 50mm unidirectional scale). If we compare the participants that

altered their voting intention with those that did not change, we find that the latter group has a higher level of polarization (M=34.0, SD=14.40; M=40.5, SD=11.89, Wilcoxon Rank Sum Test, W = 789.5, p-value < 0.05), indicating that they are somewhat more resistant to change. However, there were no differences in certainty of initial voting intentions between participants who made corrections (M=30.0, SD=18.58) and participants who did not make any corrections (M=31.3, SD=19.36; (Wilcoxon Rank Sum Test, W=1681, p=n.s.), which indicates that greater certainty of voting intentions does not in itself translate to a greater general awareness about one's political attitudes.

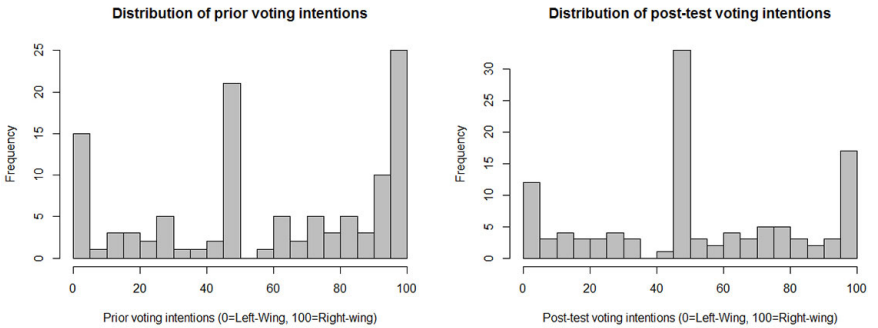

Figure 3 – (A) Distribution of prior voting intentions and (B) distribution of post-test voting intentions. The graph shows how the intentions become less polarized after the experiment.

When looking at the post-experiment notes, one salient pattern we find is that around 50% of the participants who were not influenced by the manipulation referred to their ideological identity or prior voting behavior as a reason for ignoring the incongruent compass score. More generally, for all categories of participants, many also expressed clear surprise and curiosity over the fact that they failed to correct the manipulations, then argued the opposite of their original views, and finally accepted the altered compass score.

Discussion

There are three key steps in the current result.

First, the low correction rate of the manipulated campaign issues. As reported above, the manipulations we made were generally not drastic, but constituted substantial movement on the scale, and each one of them had definitive policy implications by moving the participants across the coalition divide on issues that would be implemented or revoked at the coming term of government (yes, politicians keep most of their promises! [21, 22]). It is unlikely that the low level of corrections resulted from our use of a continuous response profile, as we observed similar results in a previous study of morality with a discrete numerical scale [17]. In fact, the survey concerned highly salient issues like income- and wealth taxation, health- and unemployment insurance, and environmental policies on gasoline and nuclear power. As such, they were both familiar and consequential, and the participants often presented knowledgeable and coherent arguments for the manipulated position (e.g. in contrast to [23, 24], who argue that voters generally lack knowledge about political facts).

Another noteworthy finding here is that we found no relationship between level of corrections and self-rated political engagement or certainty. That is, participants who rated themselves as politically engaged, or certain in their political convictions, were just as likely to fail to notice a manipulation. This complements a similar result from [17], and indicates that general self-reports of moral- or political conviction has a low sensitivity to predict correction rates on CB tasks.

The second main step of the study was the summation of the compass score. Here, an overwhelming majority of the participants accepted and endorsed a manipulated political profile that placed them in the opposite political camp. As we see it, this result is both obvious and remarkable; obvious, in that unless the participants had suspected some form of manipulation on our side, endorsement of the score follows logically from the summation (the adding was fully transparent, so it must be *their* score); and remarkable in that a few individual CB manipulations can add up to seriously challenge something as foundational as left- or right wing identity, a division seen by both academic research and commercial polling as one of the most stable constructs in the political landscape [7, 8].

But one can have many other reasons for giving political support than enthusiasm or disdain for specific policies (issues having to do with ideological commitment, trustworthiness, leadership, etc). So, the third and most critical part of the study concerned whether the participants' endorsement of the 'factual' compass score would translate to a willingness to change their actual voting intentions. Here, it must be remembered that the study was conducted at the final stretch of a real election campaign, and our ratings indicated our participants were highly certain in their voting intentions from the onset. Despite this, what we found was that no less than 48% of them were being open for movement across the great partisan divide (or 'in play', as the pollsters would say). Adding to this the further 10% that moved more than 20mm in the manipulated direction, often from positions at the absolute far ends of the scale, it is clear that our participants demonstrate a great deal of ideological flexibility.

This result can be compared to recent studies that have emphasized how hard it is to influence peoples' voting intentions with 'regular' social psychology tools, like framing and dissonance induction [25, 26] (but see [27]). Still, most likely, our findings underestimate the number of participants open to a coalition shift. As we measured voting intentions both before and after the survey, we set up a clear incentive for the participants to be consistent across measurement (e.g. [28-31]). If we instead had measured voting intention only at the end of the experiment, and used the untampered compass score as a proxy for their political affiliation, they would have had no previous anchor weighing on the final voting question, and the amount of influence would probably have been larger. Similarly, our survey contained the critical wedge issues separating the coalitions, but not any party specific interests, and some participants found they could dismiss the compass score as not representative of their critical concerns (whether this was a post-hoc rationalization or not, we cannot know). However, as our result revealed there was no difference in correction rate between smaller and larger manipulations on the scale, to gain additional force for the summation score, we could have allowed the participants to indicate which issues they cared the most about, and then focused our CB manipulations there.

As argued by Haidt [32, 33], political affiliation can be seen as primarily being about emotional attachment, an almost tribal sense of belonging at the ideological level. The goal of our study was to use CB to circumvent this attachment, and get our participants to exercise their powers of reasoning (post-hoc, or not) on the factual issues of the campaign. Previous research has shown that voters engaging

in ideologically motivated reasoning can be stubbornly resistant to correcting any factual misperceptions, even to the point where contradictory information presented to them only serve to strengthen their convictions [34]. Thus, in no part of the experiment did we provide arguments in support or opposition to the expressed views of the participants, instead they did all the cognitive work themselves when reasoning about the manipulated issues and the summary score. This way, it seems, we were able to peel back the bumper sticker mentality encouraged by coalition attachments, and reveal a much more nuanced stance among our participants. But nevertheless, we get a clue about the pervasive influence of ideology from what the participants reported at the end of the experiment. Particularly interesting are those participants that did not alter their voting intention. In this category, many referred to an overarching sense of coalition identity to motivate why the manipulated compass score did not influence them. Sometimes these participants even expressed a form of ideological relief at the debriefing stage ("pheeew... I'm not a social democrat after all!").

In summary, we have demonstrated considerable levels of voter flexibility at the cusp of a national election, with almost half of our participants willing to consider a jump across the left-right divide. As the recent assessment of the polling organizations and the polling aggregators in the US confirmed, stated voting intentions in the final weeks before an election are generally very reliable [18, 19]. This was precisely the reason we chose to conduct our study at the stretch of a real campaign. But our result provides a dramatic contrast to the established polls tracking the Swedish election, which indicated that maximally 10% of the population would be open to swing their votes, or the 5-10% of uncertain voters that Mitt Romney revealed as the exclusive target of his US presidential campaign (already in May, half a year before election day). In this way, it can be seen how the polls can be spot on about what will likely happen at the vote, yet dead wrong about the true potential for change among the voters. We are happy that only five dollars' worth of paper and glue is required to make this point, rather than a billion dollar campaign industry, but we would advise politicians against taking to the streets with a merry horde of choice blindness pollsters! Our result shows there is a world beyond ideological labels and partisan divisions, where people can approach the political issues of the campaign with considerable openness to change. Unfortunately, the question remains how to enter this world with no sleights-of-hand to pave the way.

Acknowledgement

We thank Anthony Barnhart, Steve Macknik and Max Maven for their advice concerning the historical roots of the magic trick.

References

1. Holtz-Bacha C, Stromback J (2012) Opinion Polls and the Media: Reflecting and Shaping Public Opinion. New York, NY: Palgrave McMillan.
2. Jordan J (2012) Nate Silver's Flawed Model. National Review Online. Available: http://www.nationalreview.com/articles/331192/nate-silver-s-flawed-model-josh-jordan#. Accessed 25 January 2013.
3. McLaughlin D (2012) On Polling Models, Skewed and Unskewed. Red State. Available: http://www.redstate.com/2012/10/31/on-polling-models-skewed-unskewed/. Accessed 25 January 2013.
4. Easley J (2012) GOP Takes Aim at 'Skewed' Polls. The Hill. Available: http://thehill.com/homenews/campaign/251413-gop-takes-aim-at-skewed-polls. Accessed 25 January 2013.
5. Mayer W (2012) The Disappearing - But Still Important - Swing Voter. The Forum. DOI: 10.1515/1540-8884.1520.
6. Ashkenas J, Ericson M, Parlapiano A, Willis D (2012) The 2012 Money Race: Compare the Candidates. The New York Times. Available: http://elections.nytimes.com/2012/campaign-finance. Accessed 25 January 2013.
7. Abramowitz AI, Saunders KL (2008) Is Polarization a Myth? Journal of Politics 70: 542-555.
8. Lewis-Beck MS, Norpoth H, Jacoby WG, Weisberg HF (2008) The American Voter Revisited. Ann Arbor, MI: University of Michigan Press.
9. Bafumi J, Shapiro RY (2009) A New Partisan Voter. Journal of Politics 71: 1–24.
10. Dodson K (2010) The Return of the American Voter? Party Polarization and Voting Behavior, 1988 to 2004. Sociological Perspective 53: 443–449.
11. Clarke HD, Sanders D, Stewart MC, Whiteley P (2009) The American Voter's British Cousin. Electoral Studies 28: 632–641.
12. Kitschelt H (2010) The Comparative Analysis of Electoral and Partisan Politics: A Comment on a Special Issue of West European Politics. West European Politics 33: 659.

13. Enyedi Z, Deegan-Krause K (2010) Introduction: The Structure of Political Competition in Western Europe. West European Politics 33: 415.

14. Dalton RJ (2009) Parties, Partisanship, and Democratic Politics. Perspectives on Politics 7: 628–629.

15. Cwalina W, Falkowski A, Newman B (2010) Towards the Development of a Cross-Cultural Model of Voter Behavior: Comparative Analysis of Poland and the US. European Journal of Marketing 44: 351–368.

16. 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

17. Hall L, Johansson P, Strandberg T (2012) Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. PLoS ONE 7: e45457.

18. Petrocik JR (2009) Measuring Party Support: Leaners are not Independents. Electoral Studies 28: 562–572.

19. Holmberg S, Oscarsson H (2004) Väljare. Svenskt Väljarbeteende Under 50 år. Stockholm: Norstedts Juridik.

20. Oscarsson H (2007) A Matter of Fact? Knowledge Effects on the Vote in Swedish General Elections, 1985–2002. Scandinavian Political Studies 30: 301–322.

21. Sulkin T, Swigger N (2008) Is There Truth in Advertising? Campaign Ad Images as Signals About Legislative Behavior. Journal of Politics 70: 232-244.

22. Sulkin T (2009) Campaign Appeals and Legislative Action. Journal of Politics 71: 1093-1108.

23. Delli Carpini M, Keeter S (1996) What Americans Know About Politics and Why it Matters. New Haven, CT: Yale University Press.

24. Kuklinski JH, Quirk PJ (2000) Reconsidering the Rational Public: Cognition, Heuristics, and Mass Opinion. In Lupia A, McCubbins MD, Popkin SL, editors. Elements of Reason*:* Understanding and Expanding the Limits of Political Rationality. London: Cambridge University Press.

25. Druckman J (2004) Political Preference Formation: Competition, Deliberation, and the (Ir) Relevance of Framing Effects. American Political Science Review 98: 671–686.

26. Elinder M (2009) Correcting Mistakes: Cognitive Dissonance and Political Attitudes in Sweden and the United States. Working Paper Series, Uppsala University, Department of Economics, 2009:12.

27. Carter TJ, Ferguson MJ, Hassin RR (2011) A Single Exposure to the American Flag Shifts Support Toward Republicanism Up to 8 Months Later. Psychological Science 23.

28. Wilson T, Dunn D, Kraft D, Lisle D (1989) Introspection, Attitude Change, and Attitude–Behavior Consistency: The Disruptive Effects of Explaining Why We Feel the Way We Do. Advances in Experimental Social Psychology 19: 123–205.

29. Krosnick JA, Abelson RP (1992) The Case for Measuring Attitude Strength in Surveys. In Tanur JM, editor. Questions About Questions: Inquiries Into the Cognitive Bases of Surveys. New York: Russell Sage Foundation.

30. Ariely D, Norton M (2007) How Actions Create - Not Just Reveal - Preferences. TRENDS in Cognitive Sciences, 12: 13-16.

31. Lee L, Amir O, Ariely D (2009) In Search of Homo Economicus: Cognitive Noise and the Role of Emotion in Preference Consistency. Journal of Consumer Research 36: 173-187.

32. Haidt J (2007) The New Synthesis in Moral Psychology. Science 316: 998-1002.

33. Haidt J (2012) The Righteous Mind: Why Good People are Divided by Politics and Religion. New York, NY: Pantheon.

34. Nyhan B, Reifler J (2010) When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior 32: 303-330.

Paper II

Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback

Thomas Strandberg, Jay Olson, Lars Hall, Andy woods and Petter Johansson

Abstract: American politics is becoming increasingly polarized, which biases decision-making and reduces open-minded debate. In two experiments, we demonstrate that despite this polarization, a simple manipulation can make people express and endorse less polarized views about competing political candidates. In Study 1, we approached 136 participants at the first 2016 presidential debate and on the streets of New York City. Participants completed a survey evaluating Hillary Clinton and Donald Trump on various personality traits; 72% gave responses favoring a single candidate. We then covertly manipulated their surveys so that the majority of their responses became moderate instead. Participants only noticed and corrected a few of these manipulations. When asked to explain their responses, 94% accepted the manipulated responses as their own and rationalized this neutral position accordingly, even though they reported more polarized views moments earlier. In Study 2, we replicated the experiment online with a more politically diverse sample of 498 participants. Both Clinton and Trump supporters showed nearly identical rates of acceptance and rationalization of their manipulated-to-neutral positions. These studies demonstrate how false feedback can powerfully shape the expression of political views. More generally, our findings reveal the potential for open-minded discussion even in a fundamentally divided political climate.

Introduction

The political landscape in the United States is becoming increasingly polarized [1-4]. Studies have shown that this polarization biases political decisions as well as reduces informative and critical thinking. For example, people tend to automatically support policy issues proposed by their own party and reject those coming from the opposition [5]. Even during effortful deliberation, people usually side with their own party's stance on various issues [6]. Furthermore, polarization strongly correlates with confirmation bias: polarized individuals are more inclined to seek and interpret information to confirm their present ideas about the world [7]. Recent studies have also indicated that people are more susceptible to disinformation and less likely to trust sources that do not fit their agenda [8].

Polarization also extends beyond policy issues into personal relations. The levels of animosity directed towards the opposition have dramatically increased over the past decade. In 2008, about 20% of Democrat supporters and 30% of Republican supporters reported feelings of hatred for their counterparts. In 2016, levels of hatred had risen to about 50% for both parties [1]. Most voters now report that people supporting the opposition anger and even scare them [9-10]. In a telling example, Chen and Rohla [11] found that Thanksgiving dinners in 2016 were 30 to 50 minutes shorter for families consisting of both Democrats and Republicans, compared to same-party families. Across the United States, this meant a loss of up to 34 million hours of cross-partisan Thanksgiving discussions that year, likely contributing to further polarization.

Candidates and campaign strategists leverage this powerful affective dimension of polarization to highlight their personality and leadership abilities [12-13]. This strategy was particularly salient during the 2016 American presidential election. Indeed, the contrast in personality and character between the candidates became a near obsession in both the campaigns and the media [14-15], a pattern likely to repeat in the upcoming election cycle. For example, during the final two presidential debates, the majority of questions that the moderator asked concerned the candidates' characters — even including questions such as whether it is okay for a president to be "two-faced". In the aftermath of the election, analysts expressed concerns that this trend of personality over policy would lead to even further polarization and animosity among voters [16-17]. These concerns have also persisted throughout Trump's presidency, culminating in debate about whether his rhetoric might have contributed to the increase in politically

motivated hate crimes [18] and acts of domestic terrorism such as the mail bombs sent to Democratic politicians [19-22]. Given this troublesome situation, attempts have been made to create a civic depolarization movement to promote open-minded attitudes and to make people more accepting of different political views [2, 23-25]. However, to be effective, such a movement would require a firm grasp on the nature of attitude depolarization. Thus, there is a pressing need for research that provides more knowledge about people's propensity to be more open and flexible in their political reasoning.

One way to experimentally make people consider ideas that are ideologically different from their own is through the *choice blindness paradigm*. Choice blindness is a cognitive phenomenon that occurs when people receive false feedback about a choice they had made, leading them to accept the outcome as their own and confabulate reasons for having made that choice in the first place (see [26] for details). Recently, choice blindness has been applied to the study of attitude change, an area of research that struggles to elucidate the dynamics between the stability and flexibility of attitudes. For example, in Hall, Johansson, and Strandberg [27], participants accepted 60% of the manipulations to a survey on moral dilemmas as their own attitudes. Similar findings have been reported during general elections in both Sweden [28] and Argentina [29]. Hall and colleagues [28] also found that participants not only changed their attitudes on political issues, but their actual voting intention was also affected in the direction of the false feedback (which was not found in [29]). Notably, Strandberg and colleagues [30] found that when participants accepted the manipulations of political attitudes, their attitudes shifted congruently with the false feedback and even persisted one week later.

Choice blindness has proven to be an effective tool for creating situations in which people's flexibility and openness to different political perspectives can be studied. However, as far as we know, choice blindness has never been applied during an American election on a topic as polarized, salient, and contentious as the character of presidential candidates. Given the need for reconciliation and open-mindedness in American politics [2, 23-25], we aimed to test whether we could depolarize American voters, making them more open in their judgments of competing candidates. A few weeks before the 2016 election, we asked participants to fill out a survey assessing the character traits of presidential candidates Donald Trump and Hillary Clinton. We then covertly shifted their polarized ratings to become more moderate. We hypothesized that participants

would fail to notice this manipulation and would instead accept and rationalize the altered position as their own. We also wanted to see whether changes in perceived open-mindedness would generalize to judgments of presidential competency.

Experiment 1

Method

Participants

Posing as political researchers, we recruited 136 participants in New York during the week of the first 2016 presidential debate, six weeks before the election. A third of the participants (n = 41) were recruited at the debate itself (around Hofstra University); the rest were recruited during the same week at parks in New York City (Central Park and Washington Square Park). We excluded data from 14 participants: one was too young to vote, one had trouble seeing the survey, one wished to have his data removed after the debriefing, and the rest had errors in the experimental procedure. After exclusions, 122 participants remained in the final sample (87 females; aged 18 to 42, M = 21.7, SD = 4.3). Most of them were students (75%), and the others had a wide range of occupations including journalists, professors, farmers, retailers, lawyers, and film makers. Based on a voting intention question at the end of the experiment, 89% said they planned to vote for Clinton, 3% for Trump, and 8% for a third party. The study was approved by the Lund University Ethics Board, D.nr. 2016-1046. The design and analysis were pre-registered online (see https://osf.io/gzymp); the confirmatory tests are explicitly labelled as such throughout. There was one deviation from the pre-registration: we had initially intended to exclude participants who began with more moderate views, but after analysis we decided to keep them and focus on another set of interesting yet exploratory results. This change did not affect any of the confirmatory hypothesis outcomes.

Materials and procedure

We designed a political survey to assess the leadership traits of two presidential candidates: Hillary Clinton and Donald Trump. The survey items were chosen based on traits that the public usually deems important in a president [31-32]. Participants rated the candidates on 12 adjectives describing leadership traits: analytic, trustworthy, decisive, patriotic, experienced, empathetic, visionary, courageous, diplomatic, passionate, charismatic, and principled. Each trait on the survey was shown on a visual analog scale with pictures of the candidates at either end-point. We asked participants to rate the candidates on each trait; for example, if they thought Clinton was more analytic, they would mark that scale closer to her, or if they thought Trump was, they would mark it closer to him (Figure 1A). To minimize response bias, we randomized which side of the scale Clinton or Trump appeared on for each item. Overall, the responses had good internal consistency (Cronbach's α = .82, 95% CI [.78, .87]).

Participants were randomly assigned to either the control group (n = 53) or the experimental group (n = 69). We randomized participants such that the majority would be in the experimental condition, since we were more interested in this group. In the experimental group, our goal was to make it appear as if participants had more moderate views than they initially reported. To accomplish this, while the participants rated the candidates on the 12 leadership items, we discreetly observed their responses. At the same time, we filled out an identical slip of paper with some of their most polarized responses shifted closer to the midpoint of the scales (Figure 1A). When the participants finished the questionnaire, we briefly took it to ostensibly review the responses. At this point, we covertly pasted our paper slip with the manipulated moderate responses on top of the participants' original responses (Figure 1B), then we handed the questionnaire back to them. It now appeared as if the participants had given primarily moderate responses to the questions. This replacement was inconspicuous and took only a few seconds to complete. In the control group, we performed a similar procedure but without manipulating any of the responses.

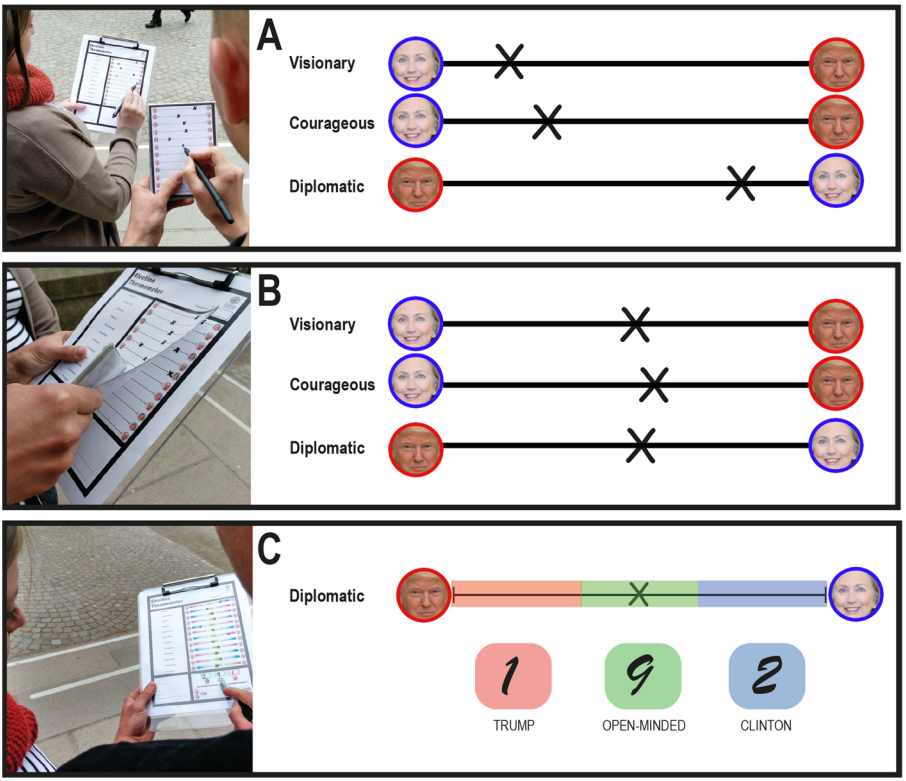

Figure 1 – The paper survey. Participants filled out a paper survey rating Hillary Clinton and Donald Trump on 12 leadership traits, such as *courageous* and *diplomatic*. In the experimental group, while participants rated the candidates, we discreetly looked at their ratings and filled out an identical slip of paper with the majority of their polarized ratings shifted closer to the midpoint **(A)**. When the participants finished the survey, we briefly took it and covertly pasted our paper slip with the manipulated moderate responses on top of the participants' original responses **(B)**. We then asked the participants to explain some of their (manipulated) ratings. Next, we overlaid a transparent sheet that categorized their ratings into: favoring Trump, favoring Clinton, or "open-minded" (i.e., neutral). Together with the participants, we tallied their ratings and asked them to explain their overall score. All participants in the experimental group now had a primarily open-minded score **(C)**. The participants in the control group did not receive any manipulations and instead explained their own original score. (Politician photographs from Wikimedia Commons).

We then asked participants in the control group to explain the reasoning behind approximately three arbitrary non-manipulated responses; in the experimental group, we asked about three manipulated ones. The experimenter would ask, for example, "Why do you think that Trump is more analytic?". If the participants hesitated, or behaved as if something were wrong, the experimenter would inform them that they could change their response (operationalized as *correction*) and

instead explain their reasoning behind that response. We tape-recorded the reasons participants gave to each of these responses.

Next, we told the participants we would calculate a summary score of their responses using a transparent overlay that segmented the scales into three categories: a clear preference for Trump, a clear preference for Clinton, or "open-minded" in the middle 30% of the scale (Figure 1C). Together with the participants, we tallied their 12 responses into the three categories. Using this segmentation rule, participants received summary feedback that their score had a majority of either Trump, Clinton, or open-minded responses. We then showed the participants their overall score and asked them, "Most of your responses were in the open-minded (or Clinton, or Trump) category – do you know why this would be?" We tape-recorded as participants explained their overall score. (Two participants did not want their voices recorded and were thus excluded from this measure.) Two independent judges later assessed whether participants justified the manipulated position. In particular, the judges rated whether participants provided clear justifications (e.g., "My parents raised me to be open-minded"), versus whether they either rejected the score (e.g., "I don't think I'm that open-minded") or did not justify it at all (e.g., "I don't know"). We conservatively defined justification as occurring only when both judges agreed that the participant justified the score; the judges agreed on 75% of their ratings.

Having discussed their aggregate score, we next asked participants to rate the candidates' competency ("How competent are these candidates as leaders?"), to see if the manipulation and confabulation would affect these more general attitudes. Here, each candidate had a visual analog scale ranging from "Extremely incompetent" to "Extremely competent". We then debriefed the participants, asked who they were planning to vote for, and finally asked for consent to use their data.

Results

Correction of the false feedback

In the experimental group, we manipulated an average of 8.53 responses closer to the midpoint of the scale, with 3.55 of these moving from supporting one candidate to being in the open-minded category. We then asked participants to explain approximately 3 (M = 3.1, SD = 0.49) of these manipulated responses,

and they only corrected 12% (95% CI [8%, 17%]) of these. Overall, 28% of the participants corrected one manipulation and only 4% corrected two. None corrected more than two of the discussed responses. The participants who made the corrections said that they had either made an error or changed their mind about the rating. No participants expressed any suspicion that their responses had been manipulated, even when asked after the study if they had noticed anything unusual. Accordingly, the participants accepted the large majority of the manipulated responses as their own. After accepting the manipulated responses, participants often gave elaborate arguments for them. For example, one participant marked his response to the *experienced* item as 94% on the Clinton side of the scale, which we manipulated to a more neutral position closer to the middle of the scale (59%). When asked to explain the latter rating, he said, "I think they're both experienced in their field. Trump is a really successful businessman... And then, Hillary has had a lot of years [of] practice in office. So I ... feel like they both are really experienced." Another participant originally rated *diplomatic* as 73% on the Clinton side, which we changed to more neutral (57%). She stated, "Hillary has been in the political scene for a very long time, but I think also Trump has a diplomatic aspect to him just because he is very passionate ... about the country." Participants thus offered arguments for moderate positions even though they had originally reported more polarized opinions just moments earlier.

Manipulation, acceptance, and justification of the aggregate survey score

Our false feedback made it appear as if participants were overall less polarized. In the experimental group, participants originally had an average of 4.32 (95% CI [3.88, 4.75]) neutral responses out of 12; after the manipulation and correction phase, the participants were given the feedback that they had 7.87 [7.52, 8.20] of them (Figure 2A). Looking only at participants that had an overall polarized score (i.e. a majority of responses favoring a single candidate), they had 3.20 [2.79, 3.59] neutral responses before the manipulation and 7.27 [6.70, 7.77] after it. Originally, 25% [15%, 37%] of participants in the experimental group had a majority of neutral responses, and the false feedback suggested that almost all of them (97%) did. The control group experienced no manipulation, and 30% [19%, 45%] of them had primarily neutral responses. As expected, in the control group, the large majority of participants (90% [77%, 96%]) verbally justified their own original views, whether neutral or polarized (Figure 2B).

Surprisingly, in the manipulation group, a similar number of participants justified their *manipulated* views which they did not hold moments earlier (94% [84%, 98%]). For example, one participant heavily favored Trump; after the false feedback about open-mindedness, he claimed, "I feel like Clinton and Trump are both in the middle and I don't really stand for either of them." Another participant who initially favored Clinton stated, "I guess I fall somewhere in the middle – I'd like to think I'm a little moderate. … I think at this point it's important to be open-minded." Others discussed balancing the strengths and weaknesses of both candidates: "In terms of being decisive, Trump is more exact and confident in his decisions, so that could be viewed as being decisive. But then Hillary has a track record in which she's changed her mind about a lot of issues, but that's kind of like her educating herself and having developed thought. So that's two different ways of looking at it."

Figure 2 – Frequency of "open-minded" responses and justification rates. The feedback made it appear as if participants had provided more open-minded responses (A); they then explained the reasons behind their original views or the manipulated ones (B).

Competency rating

At the end of the experiment, we asked participants to evaluate both candidates' competence as leaders. The average absolute difference in competency ratings between the candidates was 48.37 [39.39, 56.04] in the control group and 53.45

[47.03, 59.82] in the experimental group. Confirmatory tests showed that these differences did not vary by group ($t(120) = 0.95$, $p = .345$), nor did individual ratings for Clinton (Wilcoxon-Mann-Whitney $Z = -.400$, $p = .691$) or Trump ($Z = .599$, $p = .550$). This indicates that while participants in the experimental group often endorsed and rationalized their seeming open-mindedness, the manipulation did not affect their overall candidate judgments.

Summary of Experiment 1

We found that participants rarely detected when their evaluations of the two presidential candidates had been manipulated into a more "open-minded" position. Instead, they accepted the altered responses as their own and offered unequivocal justifications for them. In the end, this made them endorse a substantially more neutral position compared to their original score. This finding builds upon and supports previous studies exploring false feedback and political attitudes [28-30]. However, choice blindness had never been applied to study depolarization of candidate evaluations during an American election.

Experiment 2

One major caveat of Experiment 1 is that, due to the location of the first presidential debate in New York, our sample was heavily skewed towards the Democratic Party. Looking at the overall tally of the responses for all participants, 85% had more responses favoring Clinton and only 11% favored Trump. This was further reflected in the general competency rating: 89% of participants thought Clinton was more competent and planned to vote for her. Typically, we would not be concerned with this limitation, as we have no prior reason to expect that Republican supporters would behave differently from Democrats. Choice blindness studies generally have given few indications that individual differences are key to explaining the effect. However, two factors may make the present situation unique. First, the stakes are considerably higher, as research on political attitudes is often weaponized and wielded in the public debate on polarization. Second, and more important, studies on potential individual differences between liberals and conservatives have become a hotbed of activity, with many contentious results and speculative interpretations. A choice blindness study with

participants from the full political spectrum could provide a valuable contribution to this debate. Thus, we decided to run a second experiment with a larger and more representative sample.

In the ongoing chase for dissimilarities in personality and cognitive processing between liberals and conservatives, there is some evidence that personality might differ between them. In the popular Big Five personality inventory, liberals score higher on openness to experience whereas conservatives score higher on conscientiousness [33-34]. When it comes to universal values, people on the left tend to value universalism and benevolence, whereas people on the right tend to value achievement and tradition [35]. Researchers have also underlined differences in moral reasoning; liberals tend to favor particular foundations (e.g., harm/care, fairness/reciprocity) whereas conservatives put more emphasis on others (e.g., authority/respect [36-37]). Several studies have also found differences in thinking styles: conservatives have been seen as more intuitive and heuristic, whereas liberals have been seen as more analytic and systematic (e.g. [38-39]). In line with this, two studies found indications that "bullshit receptivity" — the propensity to believe statements independent of their truth — was higher for conservatives [40-41].

On the other hand, it is unclear how these findings translate to the realm of polarization, as studies of political cognitive processing seem to indicate that conservatives and liberals are similarly sensitive to various biases. For example, Frimer, Skitka and Motyl [42] found that the opposing camps were equally averse to statements that did not support their political position. Even when participants had a chance to earn money by simply reading counter-ideological statements, about two thirds of both liberals and conservatives declined to do so, indicating that there is a considerable mental "cost" involved in exposing oneself to opposing information and arguments. Furthermore, in a meta-analysis of 43 studies investigating various biases, the researchers found almost identical levels of partisan bias and confirmation bias for both liberals and conservatives [43]. Similarly, the propensity to believe fake news has also been found to rely on factors such as analytic thinking and prior exposure, rather than partisanship [44-45].

It remains unclear whether liberals and conservatives would differ on a novel decision measure like choice blindness, which involves a combination of false feedback and potential confabulation not used in any of the studies previously discussed. Susceptibility to false feedback has not systematically been linked to ideology, and political choice blindness studies conducted in Sweden and

Argentina have yielded mixed results (see [27-30] for details). However, the two-party electoral system in the United States, fueled by higher levels of polarization, is an ideal domain to explore this research question. Thus, in Experiment 2, we aimed to replicate Experiment 1 testing both liberals and conservatives. To accomplish this, we designed an online version of the first experiment in order to reach a larger and more representative population.

Method

Participants

Experiment 2 took place a few days before the general election being held on November 8, 2016. Participants were 498 (60% male) American citizens with an average age of 31.1 years (SD = 10.1). They were recruited through the online survey platform Prolific Academic [46] and asked to participate in a political survey. Participants were randomly assigned to either the experimental condition (n = 405) or the control condition (n = 93). The experiments ran on the software Xperiment version 2 [47]. Participants received $2.50 USD as compensation. The study was approved by the Lund University Ethics Board, D.nr. 2016-1046.

Materials and procedure

Experiment 2 followed the same general design and procedure as Experiment 1. The participants completed a 12-item survey and were given a chance to change their responses. They then received a summary score giving them feedback about their level of open-mindedness. The survey consisted of the same leadership traits as used in Experiment 1 (e.g., analytic, trustworthy). At the start, all items were presented as a randomized list on the same page, with continuous scales ranging between Clinton and Trump (Figure 3). Rather than using a pen and paper as in Experiment 1, the participants used their mouse to draw an 'X' on the scale where it best represented their attitude towards each item. After the participants had answered all 12 items, they received the following cover story and instructions: "Researchers have found that people sometimes are influenced by the order in which the questions are asked. Therefore, we would like you to take a second look at your answers". They were then presented with the items and their responses again, but in a different order, and asked to verify or change their previous

responses. They were informed that they could change any response by clicking 'edit' and drawing a new 'X'. The items were presented one at a time, with the other items blurred.

Figure 3 – The online survey. Participants rated Hillary Clinton and Donald Trump on 12 leadership traits **(A)**. They were then instructed to look over their responses and were told that they could change any response by drawing a new one **(B)**. The items were presented one by one with the rest blurred. Participants in the experimental group received five manipulations that moved each response to a more moderate position **(C)**. Participants were then told that their score would be summarized **(D)**. They received a score showing how many of their ratings were in each of the three categories (Clinton, Trump, and open-minded). They were also told their degree of open-mindedness based on the number of their responses in the green middle segment and were asked to explain this in text **(E)**. They then rated their overall preference for the candidates **(F)**.

All participants in the experimental condition were given false feedback regarding 5 of their 12 responses. The manipulation mechanism was as follows: select the first five responses at the extremes of the scales (i.e. between 0% and 35% or 65% and 100%), and move them to a random position within the middle 30%. Should a participant have fewer than five responses outside of the middle 30%, the items farthest from the midpoint would be moved closer (by a random amount) towards the midpoint. Thus, all participants received five manipulations shifting their original responses closer to a more open-minded position.

As in Experiment 1, participants then received a summary score showing the list of all 12 items as well as their responses and their associated categories (i.e. Trump, open-minded, Clinton). The participants' degree of open-mindedness was also described in text: "judging by your score, you have a…" followed by:

"...somewhat open-minded attitude" (0-2 open-minded responses), "...open-minded attitude" (3-6), "...very open-minded attitude" (7-10), or "...extremely open-minded attitude" (11-12). Participants were then prompted to type an explanation describing their degree of open-mindedness. The last part of the experiment consisted of the question, "How would you compare the two candidates?", with a continuous scale between Clinton and Trump. Finally, the participants were debriefed and asked for their data to be used for research purposes.

Results

Analysis

In Experiment 2, we did not explicitly ask participants who they were going to vote for in the election. Instead, we based their candidate support on their original aggregate survey score and categorized the participants as either Clinton supporters or Trump supporters using a simple majority rule. Participants with a majority of responses favoring Clinton were categorized as Clinton supporters, participants with a majority favoring Trump were Trump supporters, and participants with a majority of "open-minded" responses were categorized as open-minded. Following this rule, the sample consisted of 234 Clinton supporters, 75 Trump supporters, 147 open-minded, and 42 ties in which no category has a majority. To further corroborate this classification, we compared how Clinton and Trump supporters answered the favorability question ("How would you compare the two candidates?"), with a scale ranging from Trump (0) to Clinton (100). As expected, the two groups differed in their ratings (Clinton supporters: M = 86.92 [84.84, 88.94], Trump supporters: M = 18.93 [13.94, 24.38]) indicating that this is a valid categorization of the participants' candidate preference. Similar to Experiment 1, two independent judges categorized participants' explanations based on whether they justified or rejected their ostensible open-mindedness. The judges agreed on 62% of their ratings, which was lower than in Experiment 1. This lower reliability was likely due to the poorer quality of responses; judges were making their decisions based on short phrases or sentences, while in Experiment 1 they had audio recordings lasting several minutes to provide more context.

Correction of the false feedback

We manipulated five responses for each participant to a more neutral position, and the participants were confronted with all manipulations. Of these, 41% of the total 2025 manipulations were corrected. On average, participants corrected 2.06 [1.85, 2.24] manipulations. In total, 154 participants made no corrections, and 71 corrected all of the manipulations. When we compare the correction rates of Clinton and Trump supporters, we find no difference: Trump supporters corrected 2.36 [1.87, 2.88] items on average while Clinton supporters corrected 2.32 [2.02, 2.60] (Wilcoxon-Mann-Whitney $Z = .18$, $p = .861$). Participants who began with a majority of responses in the open-minded category had a lower correction rate (1.47 [1.15, 1.77]) compared to participants favoring a specific candidate (Wilcoxon-Mann-Whitney $Z = 3.66$, $p < .001$). However, this is probably best explained by the fact that the manipulation seemed less extreme since they were already more neutral.

Manipulation, acceptance, and justification of the aggregate survey score

Originally, the participants had on average 4.06 [3.80, 4.33] neutral responses; after being exposed to and correcting the manipulations, they had 6.71 [6.37, 7.04] neutral responses (Figure 4A). Importantly, both Clinton (2.43 [2.21, 2.63]) and Trump supporters (2.33 [1.95, 2.74]) began with the same number of neutral responses. After the manipulation and corrections, this amount had doubled (Clinton supporters: $M = 5.07$ [4.65, 5.47]; Trump supporters: $M = 5.07$ [4.44, 5.71]). As a result of this, when participants received a description at the end about their level of open-mindedness, they were most often told "you have an open-minded attitude" (i.e. between 4 and 7 open-minded responses). They were then given the opportunity to explain their open-mindedness in text and these were analyzed by independent judges. Overall, the confabulation rates in the experimental group were high (71% [64%, 78%] for Clinton supporters and 73% [60%, 83%] for Trump supporters; Figure 4B), meaning that both Clinton and Trump supporters justified their apparent open-mindedness. There was no difference in their degree of justification ($\chi^2(1, N=245) = 0.03$, $p = .872$).

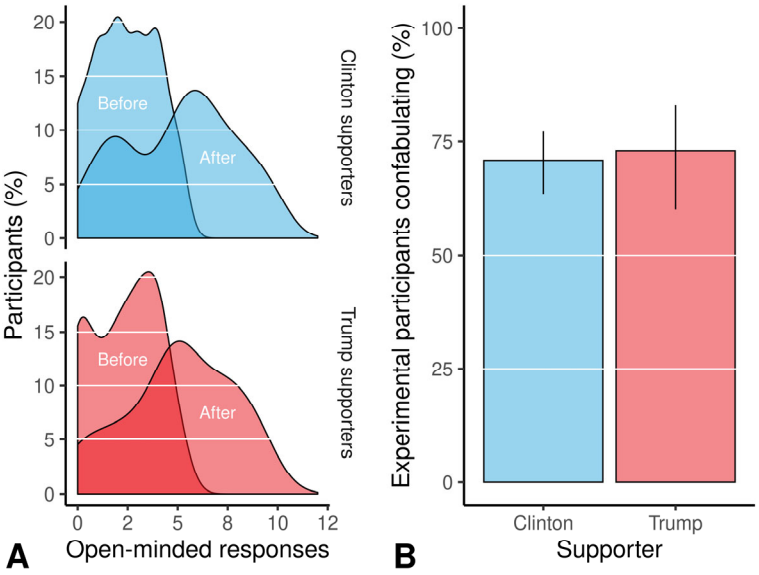

Figure 4 – Frequency of "open-minded" responses and confabulation rates in the experimental group. As in Experiment 1, the manipulation made it appear as if the participants had provided more open-minded responses (A); they then explained the reasons behind their original views or the manipulated ones (B). We saw similar rates for both Clinton and Trump supporters.

Favorability rating

In Experiment 1, the open-mindedness manipulation did not influence participants' overall competency ratings. In Experiment 2, we instead asked participants to rate their favorability: "How would you compare the two candidates?" Again, we saw no differences between the control group (M = 69.40 [63.32, 75.27]) and the experimental group (M = 64.08 [61.06, 67.03]; Wilcoxon-Mann-Whitney Z = 1.43, p = .154), and in both groups Clinton supporters favored Clinton (M = 86.93 [84.77, 88.80]) whereas Trump supporters favored Trump (M = 18.93 [13.93, 24.71]). This shows that even though participants in the experimental group endorsed and justified their apparent open-mindedness, Trump supporters still rated Trump as more favorable, and Clinton supporters rated Clinton as more favorable. As in Experiment 1, changes in individual character evaluations do not necessarily influence overall favorability.

Discussion

There is an ongoing quest to create a less polarized and more open-minded political climate in the United States [2, 23-25]. We believe this to be an important effort for several reasons. Studies show that polarization can bias information processing and decision making in detrimental ways [5-6, 48]. As a result, it often leads to fear, anger, and animosity towards the opposition [1, 9-10]. Polarization is also associated with dogmatic intolerance, which in turn increases the propensity to behave antisocially and to deny free speech [49]. Furthermore, polarization erodes central parts of civic society, such as trust in the government and media [50]. However, for a depolarization movement to be effective, we need to advance our theories on political attitude change and better understand the mechanisms underlying depolarization.

To contribute to this effort, we tested the choice blindness paradigm [26] with American voters just before the 2016 American general election. Our aim was to investigate whether participants could become less polarized in their political views. Study 1 was conducted during the week of the first presidential debate; Study 2 was conducted online with a larger and more representative sample. Participants responded to a survey comparing Hillary Clinton and Donald Trump on various leadership traits. In both studies, the participants in our sample were clearly polarized when entering the study. Participants that favored either of the candidates had on average only 2 to 3 "open-minded" responses out of 12, defined by a response in the middle 30% of the visual analog scales. Participants then received false feedback about their responses: we nearly doubled the number of items that participants had in the open-minded category. Only a few of these manipulations were detected and corrected, which resulted in an overall score that made it appear as if the participants were more open-minded in their views towards the candidates. When asked to explain their score, the great majority of the participants accepted and justified their apparent open-mindedness, even though they had reported more polarized views moments earlier.

Supporters of Clinton and Trump are similarly susceptible to false feedback

In Experiment 2, both Clinton and Trump supporters behaved similarly on the experimental measures: they had similar correction rates to the choice blindness manipulations and justified their open-minded score to similar degrees. This is

the first study we are aware of that demonstrates that liberals and conservatives are equally susceptible to false feedback about their own attitudes. Given previous findings that acceptance and justification of false survey feedback can lead to lasting changes in political attitudes [30], we see the lack of difference between Trump supporters and Clinton supporters as contributing to the ongoing research on the psychology of ideology. So far, this line of research indicates that liberals and conservatives are different in some aspects, such as personality [33], values [35-36], and thinking styles [38-39]. However, they are both similarly susceptible to cognitive biases [42-43]. Our findings show that choice blindness applies equally to conservatives and liberals. More generally, choice blindness offers a useful tool to test how liberals and conservatives reason — or rationalize — when presented with false information.

Choice blindness as a method to study depolarization

The current study was not intended as a practical method to influence voters but rather as a novel investigation of experimental depolarization in the political domain. We find that giving people false feedback can be an effective way to, at least momentarily, make them perceive themselves as more open towards competing candidates. This shows that even deeply held beliefs depend on situational factors and can be flexible under certain circumstances. From a theoretical perspective, we believe that participants interpret their own behavior — in this case their survey responses — and infer the reasons behind these responses [51-54]. Choice blindness could therefore be useful to study the depolarization of extreme views. For example, we could measure how susceptibility to choice blindness and confabulation are affected by the direction of the manipulation, such as going from polarized to moderate, or vice versa. This could help us understand whether being moderate or undecided is a distinct pole of its own. If so, we could explore whether these moderate views are more or less susceptible to false information. Here, the framing of moderate views may play an important role. In our studies, participants received *positive* false feedback about their survey responses. Instead of suggesting to people that they are open-minded, we might have found different results if participants had been told that they were "wishy-washy", "flip-flopping", "uncertain", "centrist", or even "moderate". Future work could examine how participants behave when they are given false *negative* or more neutral feedback as well.

The effectiveness of choice blindness in the political domain distinguishes it from many other forms of persuasion, such as perspective-taking [55-56]. In a recent study, Catapano and colleagues [57] found that such methods are less effective for deep-seated attitudes, such as those relating to politics. In fact, imagining the perspectives of out-group members can even backfire and hinder subsequent attitude change. This could partially be explained by the fact that in those paradigms, participants are fully aware that the perspective they consider is not their own and that the arguments they express are hypothetical. In choice blindness experiments, however, participants often believe that the response they are asked to explain reflects their own true attitude.

Limitations and future studies

In Experiment 1, only 12% of all manipulations were corrected, but in Experiment 2, 41% of them were. The reasons behind this difference are difficult to isolate given the variation in design between the two studies (such as the number of manipulations, the instructions for revisiting their responses, and verbal versus written explanations). One potential explanation is the plausibility of the manipulation. In Experiment 1, the manipulations were performed using a magic trick, which is extremely improbable in the context of a typical political opinion survey. Likely none of the participants had ever filled out a pen-and-paper survey that changed seconds later. Thus, if the participants lack perfect access to their own attitudes (or if political attitudes are not stored for us to access; [58-59]), then the manipulated survey responses ought to function as a prime source of evidence about their own attitudes [51-52]. The (presumably non-conscious) inference may look something like: "I wrote these responses, so either they must be my true attitudes, or else I made several large errors". So, if people see themselves as competent at answering a simple questionnaire, making a series of large errors would seem less plausible. In contrast, in Experiment 2, even though we attempted to replicate the general procedure of the original trick, participants were faced with a far less magical procedure. People are familiar with malfunctioning computer programs and websites, and thus our participants would have had little difficulty in concluding that there may have simply been a software error when saving their responses that needs correcting. Another explanation might be the difference between verbally explaining versus silently revising the manipulations. While participants in Experiment 2 were also

confronted with the manipulations, they did not have to engage in the mental task of having to recall or generate arguments for them. On the face of it, one might expect this additional reasoning process to generate more corrections, presumably by helping participants think more deeply about the issue and discovering that they do not agree with the manipulated position. However, if deliberation serves not as attitudinal fact-checking but as a way for participants to further commit to and defend their own ostensible attitudes, the reasoning process might lead to fewer corrections [53-54]. A third explanation could be simply that Experiment 2 was conducted closer to the election compared to Experiment 1, and that a larger proportion of the participants in Experiment 2 had firmly decided who they would vote for. Finally, it could also have been that the cover story in Experiment 2 — telling participants to check their responses in case they had been affected by presentation order — may have primed participants be more attentive and to search for inconsistencies.

Prior to the current study, choice blindness had only been used to study what might be called "repolarization" — for example by shifting people from agreeing to disagreeing with a statement. Here, for the first time, we show that it is possible to use the same methodology to depolarize people, by making them adopt the idea that they are more "open-minded". In future studies, we could also explore more global attitude shifts. In the two experiments presented here, the manipulations did not influence the candidate competency/favorability ratings. Had this been found, it would have been a unique case of attitude generalization where manipulation on some character judgments would bleed over and affect another more general trait. Perhaps political competency is judged somewhat independently of the specific traits in our survey.

Conclusion

Our findings corroborate a recent large-scale analysis of survey data with answers from 140 000 people across over 60 countries [60]. The researchers found that people across the political spectrum were more similar than they were different on several moral and political attitudes. We share their conclusion that similarities between the attitudes of people and groups tend to be overlooked, suggesting that the "us versus them" dichotomy is a prevalent but perhaps exaggerated narrative. We hope our findings can be used to simulate polarizing societal forces and thus

contribute to the search for an effective remedy sought by the political depolarization movements [2, 23-25]. Our study reveals that American voters at either end of the political spectrum are willing to endorse more open views about both candidates with surprisingly little intervention. Here, suggesting to people that they are more open-minded removed their political blinders and nudged them to consider and argue for more moderate views. These results offer hope in a divided political climate: even polarized people can become — at least momentarily — open to opposing views.

Acknowledgments

We would like to thank Julia Biris, Denis Chmoulevitch, Victoria De Braga, Johnny Nahas, Madalina Prostean, Claire Suisman, and Léah Suissa-Rocheleau for transcribing or judging the interviews, Despina Artenie and Mariève Cyr for helping prepare the stimuli, Kylar D'Aigle and Jason Da Silva Castanheira for transcribing the questionnaires, Dasha Sandra for feedback, and Alain Al Bikaii for formatting the manuscript. We would also like to thank the editor and the reviewers for their valuable comments.

Supplemental materials: https://osf.io/xh2tq/

References

1. Hetherington MJ, Weiler J. Prius or pickup? How the answers to four simple questions explain America's great divide. Boston: Houghton Mifflin Harcourt; 2018. 259 p.
2. Lukianoff G, Haidt J. The coddling of the American mind: how good intentions and bad ideas are setting up a generation for failure. New York: Penguin Press; 2018. 338 p.
3. Hunter JD, Bowman CD. The vanishing center of American democracy [Internet]. Institute for Advanced Studies in Culture; 2016 [cited 2019 Sep 6].
4. McCarty NM, Poole KT, Rosenthal H. Polarized America: The dance of ideology and unequal riches. Second edition. Cambridge, MA: MIT Press; 2016. 255 p. (Walras-Pareto lectures).
5. Bolsen T, Druckman JN, Cook FL. The Influence of partisan motivated reasoning on public opinion. Polit Behav. 2014 Jun;36(2):235–62.
6. Cohen GL. Party over policy: The dominating impact of group influence on political beliefs. Journal of Personality and Social Psychology. 2003;85(5):808–22.
7. Lord CG, Ross L, Lepper MR. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology. 1979;37(11):2098–109.
8. Flynn DJ, Nyhan B, Reifler J. The nature and origins of misperceptions: understanding false and unsupported beliefs about politics: Nature and origins of misperceptions. Advances in Political Psychology. 2017 Feb;38:127–50.
9. Iyengar S, Westwood SJ. Fear and loathing across party lines: New evidence on group polarization. American Journal of Political Science. 2015 Jul;59(3):690–707.
10. Partisanship and Political Animosity in 2016. Pew Research Center; 2016 [cited 2019 Sep 6].
11. Chen MK, Rohla R. The effect of partisanship and political advertising on close family ties. Science. 2018 Jun 1;360(6392):1020–4.
12. Druckman JN, Jacobs LR. Who governs?: Presidents, public opinion, and manipulation [Internet]. University of Chicago Press; 2015 [cited 2019 Sep 6].
13. Statler-Throckmorton A. Personality over policy. Stanford Politics [Internet]. 2016 Mar 3 [cited 2019 Sep 6];US.
14. Gleckman H. Character vs policy in the 2016 presidential election. Forbes [Internet]. 2016 Nov 1 [cited 2019 Sep 6];Business.

15. Waldman P. Why the 2016 campaign may be the most personality-driven ever. The Washington Post [Internet]. 2016 June 10 [cited 2019 Sep 6];Blogs.

16. Gerzon M. The Reunited States of America how we can bridge the partisan divide [Internet]. Oakland, CA: Berrett-Koehler Publishers, Inc.; 2016 [cited 2019 Sep 6].

17. French D. Can America's divide be healed? National review [Internet]. 2017 Jan 20 [cited 2019 Sep 6];Politics & Policy.

18. Buchanan S. Rage against change: White supremacy flourish amid fears of immigration and nation's shifting demographics [Internet]. The Southern Poverty Law Center; 2019 [cited 2019 Sep 6].

19. Morin R. Mail bomb suspect appeared to be fervent Trump supporter. Politico [Internet]. 2018 Oct 26 [Cited 2019 Sep 6]; Foreign Affairs.

20. Ioffe J. How much responsibility does Trump bear for synagogue shooting in Pittsburgh? The Washington Post [Internet]. 2018 Oct 28 [cited 2019 Sep 6];Outlook.

21. Stewart E. Republicans don't want to acknowledge that Trump's rhetoric is fueling political divisions. Vox [Internet]. 2018 Oct 28 [cited 2019 Sep 6];Politics & Policy.

22. Essig T. How Trump's psychology of hate unleashed the MAGAbomber. Forbes [Internet]. 2018 Oct 5 [cited 2019 Sep 6];Leadership.

23. Donaldson G. David S. Brown. Moderates: The vital center of American politics, from the founding to today. The American Historical Review. 2017 Dec 1;122(5):1611–2.

24. Haidt J. Can't we all disagree more constructively?. [Internet]. Knopf Doubleday Publishing Group; 2016 [cited 2019 Sep 6].

25. Wheelan, C. (2017). America needs a centrist party now more than ever – Here's how to make it happen. [Internet].

26. Johansson P, Hall L, Sikström S, Olsson A. Failure to detect mismatches between intention and outcome in a simple decision task. Science. 2005 Oct 7;310(5745):116–9.

27. Hall L, Johansson P, Strandberg T. Lifting the Veil of Morality: Choice blindness and attitude reversals on a self-transforming survey. PLoS ONE. 2012 Sep 19;7(9):e45457.

28. Hall L, Strandberg T, Pärnamets P, Lind A, Tärning B, Johansson P. How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. PLoS ONE. 2013 Apr 10;8(4):e60554.

29. Rieznik A, Moscovich L, Frieiro A, Figini J, Catalano R, Garrido JM, et al.. A massive choice blindness experiment on choice blindness political decisions: Confidence, confabulation, and unconscious detection of self-deception. PLoS ONE. 2017 Feb 14;12(2):e0171108.

30. Strandberg T, Sivén D, Hall L, Johansson P, Pärnamets P. False beliefs and confabulation can lead to lasting changes in political attitudes. Journal of Experimental Psychology: General. 2018 Sep;147(9):1382–99.

31. Barber JD. The Presidential character: Predicting performance in the White House. Englewood Cliffs: Prentice Hall; 1972.

32. Costa Lobo M, Curtice J, editors. Personality politics?: The role of leader evaluations in Democratic elections [Internet]. Oxford University Press; 2014 [cited 2019 Sep 6].

33. Carney DR, Jost JT, Gosling SD, Potter J. The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. Political Psychology. 2008 Oct 23;29(6):807–40.

34. Jost JT. The end of the end of ideology. American Psychologist. 2006;61(7):651–70.

35. Caprara GV, Schwartz S, Capanna C, Vecchione M, Barbaranelli C. Personality and politics: Values, traits, and political choice. Political Psychology. 2006 Feb;27(1):1–28.

36. Graham J, Haidt J, Nosek BA. Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology. 2009;96(5):1029–46.

37. Haidt J, Graham J. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Soc Just Res. 2007 Jun 1;20(1):98–116.

38. Stern C, West TV, Jost JT, Rule NO. The politics of gaydar: Ideological differences in the use of gendered cues in categorizing sexual orientation. Journal of Personality and Social Psychology. 2013;104(3):520–41.

39. Deppe KD, Gonzalez FJ, Neiman JL, Jacobs C, Pahlke J, Smith KB et al. Reflective liberals and intuitive conservatives: A look at the cognitive reflection test and ideology. Judgment and Decision Making. 2015 Jul 1;10(4):314-331.

40. Pfattheicher S, Schindler S. Misperceiving bullshit as profound is associated with favorable views of Cruz, Rubio, Trump and conservatism. Runco MA, editor. PLoS ONE. 2016 Apr 29;11(4):e0153419.

41. Sterling J, Jost J, Pennycook G. Are neoliberals more susceptible to bullshit? Judgment and Decision Making. 2016 Jul 1;11(4):352-360.

42. Frimer JA, Skitka LJ, Motyl M. Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. Journal of Experimental Social Psychology. 2017 Sep;72:1–12.

43. Ditto P, Liu B, Clark CJ, Wojcik S, Chen E, Grady RH, et al. At least bias is bipartisan: A meta-analytic comparison of partisan bias in Liberals and Conservatives. Perspectives on Psychological Science: 2019;14(2):273–291.

44. Pennycook G, Rand DG. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of Personality: 2019;1–16.

45. Pennycook G, Cannon TD, Rand DG. Prior exposure increases perceived accuracy of fake news. Journal of Experimental Psychology: General. 2018 Dec;147(12):1865–80.

46. Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance. 2018 Mar;17:22–7.

47. Woods AT, Velasco C, Levitan CA, Wan X, Spence C. Conducting perception research over the internet: A tutorial review. PeerJ. 2015 Jul 23;3:e1058.

48. Mason L. Ideologues without issues: The polarizing consequences of ideological identities. Public Opinion Quarterly. 2018 Apr 11;82(S1):866–87.

49. van Prooijen J-W, Krouwel APM. Extreme political beliefs predict dogmatic intolerance. Social Psychological and Personality Science. 2017 Apr;8(3):292–300.

50. Hetherington MJ, Rudolph TJ. Why Washington won't work: Polarization, political trust, and the governing crisis [Internet]. University of Chicago Press; 2015 [cited 2019 Sep 6].

51. Bem DJ. Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychological Review. 1967;74(3):183–200.

52. Carruthers P. How we know our own minds: The relationship between mindreading and metacognition. Behav Brain Sci. 2009 Apr;32(2):121–38.

53. Mercier H, Sperber D. Why do humans reason? Arguments for an argumentative theory. Behav Brain Sci. 2011 Apr;34(2):57–74.

54. Mercier H, Landemore H. Reasoning is for arguing: Understanding the successes and failures of deliberation. Political Psychology. 2012 Apr;33(2):243–58.

55. Broockman D, Kalla J. Durably reducing transphobia: A field experiment on door-to-door canvassing. Science. 2016 Apr 8;352(6282):220–4.

56. Janis IL, King BT. The influence of role playing on opinion change. The Journal of Abnormal and Social Psychology. 1954;49(2):211–8.

57. Catapano R, Tormala ZL, Rucker DD. Perspective taking and self-persuasion: Why "putting yourself in their shoes" reduces openness to attitude change. Psychol Sci. 2019 Mar;30(3):424–35.

58. Zaller JR. The nature and origins of mass opinion [Internet]. Cambridge: Cambridge University Press; 1992 [cited 2019 Sep 6].

59. Haidt J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review. 2001;108(4):814–34.

60. Hanel PHP, Maio GR, Manstead ASR. A new way to look at the data: Similarities between groups of people are large and important. Journal of Personality and Social Psychology. 2019 Apr;116(4):541–62.

Paper III

False beliefs and confabulation can lead to lasting changes in political attitudes

Thomas Strandberg, David Sivén, Lars Hall, Petter Johansson and Philip Pärnamets

Abstract: In times of increasing polarization and political acrimony, fueled by distrust of government and media disinformation, it is ever more important to understand the cognitive mechanisms behind political attitude change. In two experiments, we present evidence that false beliefs about one's own prior attitudes and confabulatory reasoning, can lead to lasting changes in political attitudes. In Experiment 1 (N=140), participants stated their opinions about salient political issues, and using the Choice Blindness Paradigm we covertly altered some of their responses to indicate an opposite position. In the first condition, we asked the participants to immediately verify the manipulated responses, and in the second, we also asked them to provide underlying arguments behind their attitudes. Only half of the manipulations were corrected by the participants. To measure lasting attitude change, we asked the participants to rate the same issues again later in the experiment, as well as one week after the first session. Participants in both conditions exhibited lasting shifts in attitudes, but the effect was considerably larger in the group that confabulated supporting arguments. We fully replicated these findings in Experiment 2 (N=232). In addition, we found that participants' analytical skill correlated with their correction of the manipulation, whereas political involvement did not. This study contributes to the understanding of how confabulatory reasoning and self-perceptive processes can interact in lasting attitude change. It also highlights how political expressions can be both stable in the context of everyday life, yet flexible when argumentative processes are engaged.

Introduction

In an increasingly polarized political landscape, as exemplified by the dramatic U.K decision to leave the European Union and the acrimonious 2016 U.S General Election, it is ever more important to understand the sources and dynamics of political attitude change. On the one hand, social psychological experiments have indicated that political attitudes can be flexible and sensitive to contextual influences, and that these attitudes either may be constructed in the moment (Bishop, 2005; Haidt, 2001; Zaller, 1992; Converse, 1975; 1964), or easily altered by the deliberation of the respondents (Hall, Johansson & Strandberg, 2012; Hall et al., 2013). This perspective has long prompted a concern about the power of corporate capital and the political elite to shape the public agenda (Bullock, 2011; Burke, 1774). More recently, it has led to a common recognition of the malicious persuasive potential of fake news spreading through social networks and media outlets (McNair, 2017). On the other hand, longitudinal studies have demonstrated a remarkable stability in political attitudes over the lifespan, and traced their genesis to developmental context and personality traits (Lewis, 2018; Gerber et al. 2011; Hatemi et al. 2009; Hooghe & Wikenfeld, 2007). One large-scale study found that partisan affiliation remained unchanged when measured over the course of almost four decades (Sears and Funk, 1999). They also found that only a minority of the individual attitudes fluctuated, and that these fluctuations occurred in incremental and consistent ways (see also Alvin, 1994; Sears, 1983). Similarly, much work within political science has underlined stability and resistance to change as central characteristics of political attitudes (Bartels, 2002). In light of this, when a recent study of door to door canvassing showed how 10 minutes of induced perspective taking could change participants' attitudes towards transgender persons (Broockman & Kalla, 2016), it was widely seen as a political sensation (Ledford, 2016).

But how can these differing perspectives, one focusing on attitude stability and the other on attitude flexibility, be reconciled? Here we use the *Choice Blindness* paradigm (CBP) to contribute to these questions. In the original CBP study (Johansson, Hall, Sikström & Olsson, 2005), participants decided which face they found most attractive in a pair, but sometimes the opposite alternative was presented as their actual choice. The results showed that participants often failed to notice these manipulations, and instead accepted the false feedback as their preferred choice. In addition, participants readily gave verbal explanations of why

they preferred the manipulated outcome, thus confabulating reasons for a choice they did not make. These results indicated a striking dissociation between the act of making a choice and its later justification and highlight the perils of assuming infallible self-knowledge about preferences, as is common in cognitive and economic models of decision making (Johansson et al., 2005).

The CBP, and its underlying methodology of creating dissociations between action and outcome, has since been widely replicated in a variety of different domains. These include taste preferences in a supermarket setting (Hall et al., 2010), financial decisions (McLaughlin & Somerville, 2013), eye-witness testimony (Sagana, Sauerland & Merkelbach, 2016), haptic feedback (Steenfeldt-Kristensen, & Thornton, 2013), and speech intentions (Lind, Hall, Breidegaard, Balkenius & Johansson, 2014). Recent work has also demonstrated interesting downstream effects of accepting the false feedback in the CBP, both on later memories for past choices (Pärnamets, Hall & Johansson, 2015), and for later preferences themselves (Johansson et al. 2014; Taya, Gupta, Farber, & Mullette-Gillman, 2014; Luo & Yu, 2016). In these latter experiments, not only are the participants' ratings of alternatives influenced, but also their later choices so that they become more likely to choose an alternative they previously received false feedback about choosing.

The format of the decisions in the CBP, which includes both deliberation and explanation, makes it well suited for application to political attitudes, where this type of explicit reasoning often is highlighted as an important ideal (Druckman, 2004; Anand & Krosnick, 2003, Taber & Lodge 2013). In previous work, we have demonstrated that salient moral (Hall, Johansson & Strandberg, 2012) and political attitudes (Hall et al., 2013) are susceptible to false feedback manipulations. In these studies, participants' responses were reversed to indicate the opposite of what they had answered, and more than half of these manipulated responses were accepted by the participants as being their original attitudes. Yet, it is unclear whether CBP can induce lasting attitude change, as the participants in these studies were debriefed about the false feedback soon after the study and were reacquainted with their original answers. Using faces as stimuli, Taya et al. (2014) found preference change resulting from the false feedback in the short-term, but no effect when measured a week later. However, the influence of the false feedback in Hall et al. (2013; 2012) was considerable, and it is likely it might have been sustained if the debrief had been postponed and the participants queried at a later time. Thus, the first aim of the current study is to investigate

whether false feedback about one's own survey responses can result in lasting change to one's political attitudes.

Second, if this is the case, what might the mechanisms be? In a classic study, Janis and King (1954) used role playing as a manipulation and had participants actively arguing for hypothetical future events, such as an estimation of the amount of movie theatres still open in three years' time. They found that participants who expressed verbal arguments in favor of an estimate were more likely to change their attitude to correspond with it, compared to a passive control group that did not verbally engage with the issue. They also found that participants in the experimental condition reported a higher confidence in their attitude. Similar kinds of attitude change have also been reported for groups, for example, when groups' jointly decided attitudes toward specific issues were rated as more extreme compared to the mean original rating of each individual (Kogan & Wallach, 1967). In particular, the attitudes of actively discussing groups changed more compared to groups that only listened to recordings of another group's discussion (Kogan & Wallach, 1967; Isenberg, 1986). In another more recent line of work, Clarkson, Tormala and Leone (2011) found that if participants get to think about an object for up to 300s compared to 60s, their confidence regarding their own attitudes directed at this object was increased and their attitudes became more extreme. In Barden and Tormala (2014), participants' attitude strength was similarly influenced by how they experienced their own arguments: the more arguments the participants expressed in favor of a cause, the stronger their pro-attitude for that cause became. These findings illustrate that the perception and verbalization of one's own reasoning processes can largely impact one's attitudes (Knowles & Linn, 2004; Tormala & Petty, 2002).

Reasoning is a core element in the CBP, since participants are asked to verbally explain their (putative) choice (Johansson, Hall, Sikström, Tärning & Lind, 2006). What is interesting is that we can be certain that these explanations are confabulatory, since the participants give reasons for a choice they in fact did not make (Johansson et al., 2005). The majority of previous research on confabulation has described it as a clinical spectrum disorder (Fotopoulou, Conway & Solms, 2007, Hirstein, 2009). Confabulation has also been implicated in (false) memory formation (Loftus & Zanni, 1975, Berstein, Laney, Morris & Loftus, 2005), and there are indications it might be prevalent in typical peoples' everyday lives (French, Garry & Loftus, 2009). This possibility is strengthened by the lack of

semantic and emotional differences found in CBP contrast analysis between the non-manipulated and manipulated verbal reports (for detailed analyses of such reports, see Johansson et al. 2005; 2006 and Hall, Johansson & Strandberg, 2012). Potentially, the process behind all introspective reports might be confabulatory at its core (Dennett, 1987). However, without a wedge like CBP to get between the decisions of the participants and their reports, it is difficult to question the subjective authority of the participants. Consequently the impact of confabulatory reasoning on attitude change has not been studied at all. Since confabulatory reasoning has been found to strengthen false beliefs, and since depth of reasoning in general can influence attitudes, we hypothesized that the amount of confabulation a participant engages in when justifying a false feedback response, will increase the self-induced attitude change, as well as its persistence over time.

To investigate this as well as the longevity of attitude change following false feedback, we conducted two experiments. In Experiment 1 our participants filled out a political attitude survey on several specific political issues in the areas of health care, education, and environment. They then received false feedback about some of their responses to these issues (see Figure 1). Half of the participants were assigned to the *Acknowledge* condition, and asked to merely acknowledge their responses, whereas the other half was assigned to the *Confabulation* condition and asked to give verbal explanations behind some of their responses. We then asked participants to state their attitudes to the same issues a second time, a few minutes after having been confronted with the false feedback. Participants were also invited to a third attitude survey one week later. In Experiment 2 we sought to replicate the findings of Experiment 1, as well as adding additional measures to investigate some possible moderators of the reported effects.

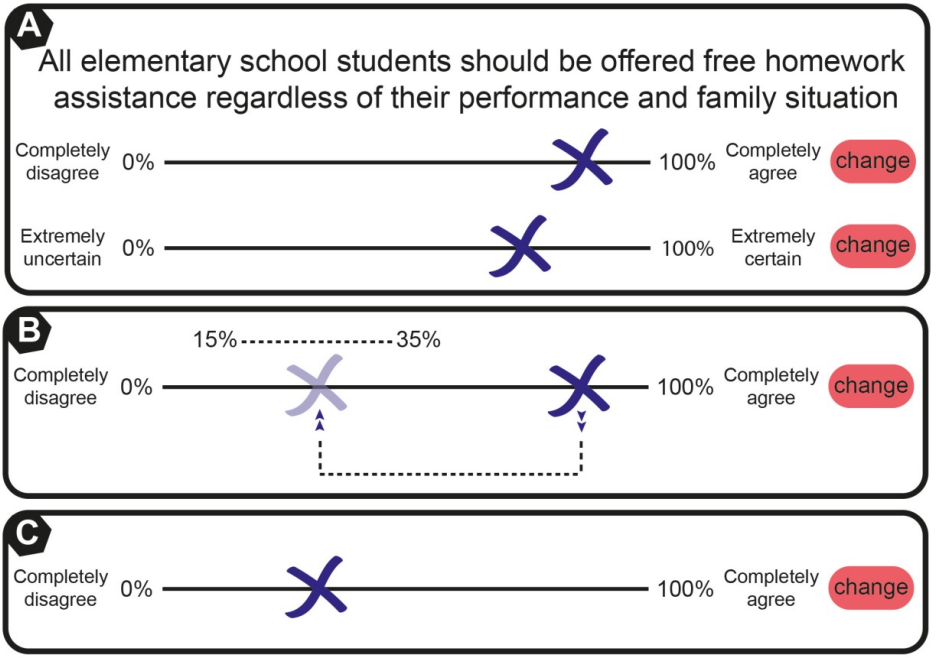

Figure 1 – Manipulation. Participants rate to what extent they agree with a political statement as well as their level of confidence on a visual-analog scale ranging from 0% to 100% **(A)**. After responding to all 12 statements, participants are asked to go over four of the responses together with the experimenter. At this stage, the application has moved two of their responses to the opposite side of the scale. The manipulation moves the responses across the midline and randomly place them between 15% and 35%, or 65% and 85% **(B)**. In the acknowledge condition, participants are asked to just verify their responses. In the confabulation condition, they are also asked to explain the reasons behind each response **(C)**. Participants can always change a response by clicking change **(A-C)**.

Experiment 1

Method

Participants. We recruited a total of 150 participants (91 female), with an average age of 22.7 years (*SD* = 3.0), at Lund University campus. Ten participants were excluded from the final analysis: of these four participants did not show up for the second session, and six experienced a malfunction with the experimental apparatus. One hundred and forty participants were included in the final analysis. Participants received two cinema vouchers in exchange for their participation in two experimental sessions, roughly one week apart (average 6.3 days (*SD* = 1.8)).

At the start of the experiment, we described the general purpose and the outline of the experiment, but without telling the participants that some of their answers would be manipulated. We also informed the participants that they could quit the experiment at any time and request their data to be erased. All participants were fully debriefed after the second follow-up of the experiment, before consenting to their anonymized data to be used by signing a consent form. The participants that did not show up for the second follow-up were debriefed over the telephone. The experiment was approved by the Lund University Ethics board, D.nr. 2008–2435.

Materials and design. Three questionnaires were administered during the experiment. One questionnaire was a tablet application specifically developed for giving participants false feedback about their survey ratings, the Self Transforming Survey. It was developed in the programming language Python with Django framework as a backend on the server side. The front end was coded in HTML, CSS bootstrap and the dynamical functionality in Javascript with the help of JQuery library. The remaining two surveys were regular pen and paper surveys. Further, an audio recorder was used to capture the verbal reports given by the participants.

The political statements were divided into three categories: health care, education, and environment. Six of the statements were used in all three questionnaires. Of these six, four were target statements that were randomly assigned as either manipulated or non-manipulated, taken from the environment and education categories. All statements concerned salient political topics in Sweden at the time of the experiment and were constructed to state a proposed policy and give a brief explanation of that policy. One example of a target statement:

> "The Swedish elementary school should be re-nationalized. Local municipalities would then lose some influence, and the state would become head of the school and assume the responsibility for resource allocation and quality assurance" (see OSF repository for complete list of statements).

Procedure. The experiment consisted of three sessions: initial rating and interaction with the manipulated and non-manipulated responses (T1); a second rating session following their interaction with the experimenter and their initial ratings (T2); and a third rating session around one week later to measure lasting

attitude change (T3). The participants were randomly assigned to one of two conditions: Acknowledge or Confabulation.

The experiment proceeded as follows: the participants were recruited from the common areas of a university building and asked if they would be willing to answer a political questionnaire. If they accepted, participants were brought to a separate room, seated in front of a tablet, and explained the general outline of the procedure, but without mentioning the false feedback. The questionnaire ran on The Self-Transforming Survey (STS), a tablet application specifically developed for giving participants false feedback about their survey ratings. The questionnaire contained 12 political statements, presented one at a time, and the participants' task was to rate to what extent they agreed or disagreed with each statement by drawing a mark on a visual-analogue scale with end-points anchored at "Completely disagree" to "Completely agree". Below each statement they also estimated how confident they felt about their attitude, on a similar scale but with endpoints going from "Extremely uncertain" to "Extremely certain" (Figure 1A). The participants were left to answer the questionnaire at their own pace. The attitude ratings obtained during this initial portion of the experiment are referred to as the T1 ratings and serve as the baseline to which later attitudes are compared.

Afterwards, the experimenter re-entered the room and informed the participants that the application would now randomly display four of the statements, one at a time, together with their ratings (but without the confidence rating). Here, the participants' ratings to two of the four displayed statements had been manipulated by the application (Figure 1B-C). The participants in the both the Acknowledge and Confabulation conditions were instructed to read each displayed statement aloud, tell where on the scale their rating was, if this implicated that they agreed or disagreed with the statement, and to what extent (for example by saying "I agree with that to some extent"). Participants in the Confabulation condition were also instructed to explain their reasoning behind each response. After a participant had stated a position, the experimenter asked: "Why do you [to some extent] agree with that statement?" but avoided interacting with the participants while they were explaining. If a participant, for example, had questions the experimenter just mentioned that it was up to the participant to interpret the statement. Thus, all participants in the Confabulation condition received the same treatment and the experimenter was not involved in the reasoning task.

During a manipulated trial, the participants' rating was always moved across the mid-line of the 0-100% scale, thus shifting the participants stated attitude from agreeing to disagreeing with the statement (or vice versa). The manipulated rating was randomly placed between 15%- 35%, or between 65%-85%, depending on the direction of the manipulation (see Figure 1). Additionally, each scale was coupled with a change button, so while filling out the survey, as well as when going over the ratings with the experimenter, the participants always had the option to change a rating should they feel that it did not reflect their attitude towards a particular issue. If the participants hesitated, or behaved like something was wrong, the experimenter informed them that they could change their response by clicking change and then draw another rating. A manipulation was automatically registered as corrected when the participants clicked the change button and drew a new rating on the scale.

After the tablet survey and the interaction with the four target statements was finished the participants were asked to fill out another questionnaire, this time on paper. These ratings are referred to as T2 ratings. The questionnaire also contained 12 political statements: six from the first questionnaire, including the two manipulated and the two non-manipulated statements, as well as six new statements. The participants were told that it was possible that some of the statements that they had already responded to on the application might reappear, since they were all randomly drawn from the same bank of statements.

The participants were scheduled to return in one week for the second follow-up, which took place on average 6.3 days (*SD* = 1.8) later. These are referred to as T3 ratings. In this follow-up, the participants answered another paper survey containing 12 political statements, including the same six statements from the previous questionnaires (two manipulated, two non-manipulated, and two filler statements) mixed with six new statements. Finally, the participants were debriefed in full, and signed data release statements.

Analysis All ratings were converted to a 0-100mm scale to facilitate comparisons between mediums (i.e. STS (T1) and paper-pen (T2 and T3)). For our analyses we used the ratings in two ways, outlined here.

First, we investigated if attitude strength at T1 predicts correction in the task. To simplify the analysis, we converted the attitude ratings to a 0-50 scale. This was done by centering the scale, so it ranged from -50 to +50 and then used the absolute resulting values. Thus, a rating of 0 (maximum disagree) and a rating of

100 (maximum agree) would both correspond to an attitude strength of 50 (maximum strength). A rating of 50 (no opinion or undecided) would be 0 on the attitude strength scale.

Second, for the main dependent measure, attitude change, we wanted to analyze changes to the participants' stated attitudes over time. To do this, we first needed to realign the attitude ratings, to make them comparable regardless if the participants agreed or disagreed with the statements. This was done at all time-steps of the experiment. We then used the realigned ratings to measure the difference between the original attitude (T1) and later attitudes (at T2 and T3). Both steps are described below.

Participants' ratings on the 0-100mm scale were numerically realigned to facilitate comparison between participants who would otherwise have opposing opinions on an issue. For statements where the participants' T1 ratings were under the midline of the scale (<50), all ratings from that participant to that statement were flipped over the midline. For example, if the participants responded 25 at T1, 60 at T2 and 30 at T3 to some statement, these values were recoded to 75 at T1, 40 at T2 and 70 at T3. For statements where the participants' T1 ratings were over the midline of the scale (>=50), no changes were made. All participants′ ratings at all time-steps of the experiment are shown on the same directional scale

Since our main hypotheses concerned attitude change, the T2 and T3 ratings were analyzed as differences compared to the original T1 rating. A negative difference represents a movement in the attitude towards or beyond the midline, and for manipulated trials, in the direction of the false feedback. Referring back to our earlier example, if the participant's realigned rating at T1 was 75 and the rating at T2 was 40, this represents an attitude change score of -35. We refer to such changes as a weakening of the attitude. Conversely, if the participant's rating at T1 was 75 but the rating at T2 had been 80, this represents an attitude change score of +5, and is described as a strengthening of the attitude.

We analyzed our data using (generalized) linear mixed-effects models using the lme4 package in R. Random-effects were modelled as per participant intercepts and slopes mirroring the full fixed-effects structure, or the maximally permitted structure that would converge (Bates et al., 2015). Significance of fixed-effects was assessed using Wald Chi-square tests as implemented in the car package (Fox & Weisberg, 2011). We report marginal model R2 for the fitted models, describing the proportion variance explained by the fixed-factors, using the piecewiseSEM package (Lefcheck, 2015), which is a variance explained measure specific for

mixed-effects models. For interpretation of effects we report unstandardized beta coefficients from our analyses and their standard errors, which can be interpreted on the 0-100 mm scale.

Results

Correction of manipulated responses. Of the 277 manipulated (M) trials, 134 (48.4%) were corrected by the participants, meaning 51.6% were accepted. Average by participant correction rate was 1.0 trials (*SD* = 0.8). Forty-five (32%) participants made no corrections, 56 (40%) made one correction and 39 (28%) made two corrections. All participants and trials were included in the analyses.

Effects of confidence, attitude strength, and condition on correction. In the Confabulation condition, participants corrected 53.3% of manipulations while participants in the Acknowledge condition corrected 43.6% of manipulations. Average attitude strength was *M* = 23.1, *SD* = 14, on a 0-50 scale where 0 represents the indifference point. Next to each political statement, the participants also rated how confident they felt about their response. Average confidence was high with an average of 63 out of 100 (*SD* = 23). Confidence was higher for Corrected trials (*M* = 70, *SD* = 22) than for Accepted trials (*M* = 56, *SD* = 23; Welch t-test $t(191.77) = 5.88$, $p = 1.78*10^{-8}$). Confidence was highly correlated with attitude strength, $r = .64$, 95% *CI* [.59, .69].

We analyzed the effects of confidence, attitude strength and confabulation condition on the probability of correcting the manipulation. Both confidence and attitude strength were standardized prior to analysis to aid model convergence, while condition was deviation coded (Confabulation = 0.5). We found a significant interaction between confidence and attitude strength ($\chi^2_{(1)} = 8.09$, $p = .0044$), but no other significant effects, with marginal model $R^2 = .246$. The regression coefficients of confidence and attitude strength were all positive, indicating that participants were most likely to correct attitudes which were both extreme and confidently held (see Table 1).

Effect of manipulation and correction on future ratings. We tested the effect of the false feedback during the two follow-up surveys (T2 and T3) in two regressions. In the first, we regressed attitude change on manipulated versus non-

manipulated trials together with an interaction with time. All variables were dummy coded taking T2, non-manipulated trials as reference levels. In the second, we regressed attitude change on accepted versus corrected manipulated trials, disregarding non-manipulated trials, together with an interaction with time. All variables were dummy coded taking corrected trials at T2 as reference levels. We report each regression in turn.

Table 1 – All estimated regression coefficients and their standard error for mixed-model analysis of correction. For all predictors Wald Chi-square and p-values are also reported.

Effect	Estimate	Standard error	Wald χ^2 (df=1)	P-value
Intercept	-0.51	0.26	-	-
Confidence	1.05	0.33	3.48	.062
Attitude strength	0.13	0.27	1.28	.26
Condition	0.38	0.50	2.06	.151
Confidence X Attitude strength	0.65	0.22	8.09	.0044
Confidence X Condition	0.11	0.55	0.034	.86
Attitude strength X Condition	0.10	0.54	0.25	.61
Confidence X Attitude strength X Condition	0.40	0.40	1.02	.31

The first regression tested if attitude change differed on average between manipulated (M) and non-manipulated (NM) trials. *We found* significant main effect of Manipulation ($\chi^2_{(1)} = 39.23$, $p = 3.7*10^{-10}$), as well as a significant interaction between Time and Manipulation ($\chi^2_{(1)} = 31.64$, $p = 1.9*10^{-8}$), but no main effect of Time ($\chi^2_{(1)} = 2.41$, $p = .12$), with model marginal $R^2 = .09$. Interpreting the coefficients, participants were highly accurate in restating their original attitude in T2 during non-manipulated (NM) trials ($b_{intercept}$ = -1.1mm, SE = 0.9) and this changed little from T2 to T3 (b_{T3} = -1.2mm, SE = 1.2). There was a large weakening of attitudes at T2 for manipulated (M) trials (b_M = -12.8mm, SE = 1.6) which was attenuated at T3 (b_{T3*M} = 8.2mm, SE = 1.7).

We additionally examined if initial confidence predicted later attitude shifts, by comparing the model fitted above, with one including an additional standardized confidence term and all interactions with Manipulation and Time. However, including the confidence term did not significantly improve fit ($\chi^2_{(3)} = 6.17$, $p = .09$), and the fitted coefficients of confidence indicated that any effects were negligibly small (b_{Conf} = 0.5mm, SE = 1.0; b_{Conf*M} = 2.0mm, SE = 1.4; $b_{Conf*T3}$ = 0.4mm, SE = 1.3; $b_{Conf*T3*M}$ = -2.5mm, SE = 1.9).

The second regression contrasted accepted (A) and corrected (C) manipulated trials, subsetting the data to only include manipulated trials. We found a significant main effect of Correction ($\chi^2_{(1)} = 98.52$, $p = 2.2*10^{-16}$) and of Time ($\chi^2_{(1)} = 33.09$, $p = 8.79*10^{-9}$), as well as a significant interaction between Time and Correction ($\chi^2_{(1)} = 11.21$, $p = .00082$), with model marginal $R^2 = .24$. Interpreting the coefficients, participants displayed virtually no directional change in attitudes at T2 during corrected trials ($b_{intercept}$ = -2.5mm, SE = 1.3) and this changed little from T2 to T3 (b_{T3} = 2.8mm, SE = 1.8). Consistent with our hypotheses, we found a large weakening of attitudes in T2 for accepted (A) trials (b_A = -21.6mm, SE = 2.2), an effect that was attenuated at T3 (b_{T3*A} = 8.3mm, SE = 2.5). To summarize: we found evidence of directional attitude change following from accepted but not for corrected false feedback trials. The effects where largest at T2 but remained robust at T3.

Qualitative shifts in position. Given the changes in ratings at T2 and T3, we examined the proportion of the trials that crossed the mid-line of the attitude scale, indicating a qualitative shift compared to the original T1 attitude. At T2, 73% of responses represented such a shift for Accepted trials, compared to 10% for Corrected trials and 11% for Non-Manipulated trials. At T3, where the attitudinal effects of the manipulation were attenuated, 41% of responses were still qualitatively shifted for Accepted trials compared to 10% for Corrected trials and 12% for Non-Manipulated trials.

Effect of confabulation on future ratings. We investigated the effect of Confabulation condition (dummy coded with the acknowledge condition as reference level), on subsequent attitude change. We first analyzed all trials, following the same analytical strategy as above, contrasting manipulated and non-manipulated trials including interactions with Time and Confabulation condition. We found no main effect of Confabulation ($\chi^2_{(1)} = 0.0082$, $p = .93$; b_{CONFAB} = -0.1mm, SE = 1.9) nor any interaction with Manipulation ($\chi^2_{(1)} = 1.42$, $p = .23$; $b_{M*CONFAB}$ = -3.0mm, SE = 3.1), Time ($\chi^2_{(1)} = 0.83$, $p = .36$; $b_{T3*CONFAB}$ = -1.7mm, SE = 2.4) or three-way interaction ($\chi^2_{(1)} = 0.01$, $p = .92$; $b_{M*T3*CONFAB}$ = -0.4mm, SE = 3.4; (see also Fig. 2A-B). This shows that participants' attitude stability in general was not affected by the method of restating their attitudes. The remainder of the analysis yielded coefficients consistent with previous results (see Supplementary results).

Previously we showed that attitude change was only present for accepted manipulated trials. Therefore, we again subset the data on manipulated trials and contrasted corrected (C) and accepted (A) trials, including interactions with Time and Confabulation condition. We found that participants displayed no directional attitude change in T2, corrected trials in the acknowledge ($b_{intercept}$ = -2.9mm, SE = 2.2) or confabulation conditions (b_{CONFAB} = -0.2mm, SE = 3.1; see Figure 2C), with similar results for T3 trials (b_{T3} = -1.9mm, SE = 2.8; see Fig. 2D). There was a large directional attitude change for the accepted trials (b_{A} = -16.7mm, SE = 2.8; $\chi^2_{(1)} = 124.14$, $p < 2.2*10^{-16}$). Importantly, in line with this we found main effects of Condition ($\chi^2_{(1)} = 4.81$, $p = .028$), and Time ($\chi^2_{(1)} = 30.74$, $p = 3.0*10^{-8}$), and these were qualified by interactions between Correction and Condition ($\chi^2_{(1)} = 8.33$, $p = .0039$) and between Correction and Time ($\chi^2_{(1)} = 10.71$, $p = .0011$). Taken together, this means that the directional changes of accepted trials were, as hypothesized, enhanced in the Confabulation condition at T2, meaning a further weakening of the original attitude ($b_{A*CONFAB}$ = -9.6mm, SE = 4.0). Attitude changes were attenuated at T3 (b_{A*T3} = 7.7mm, SE = 3.6). The interaction between Condition and Time ($\chi^2_{(1)} = 0.77$, $p = .38$; $b_{CONFAB*T3}$ = -1.5mm, SE = 3.7) and the three-way interaction were not significant ($\chi^2_{(1)} = 0.078$, $p = .78$; $b_{A*CONFAB*T3}$ = -1.4mm, SE = 5.1). Model conditional $R^2 = .26$.

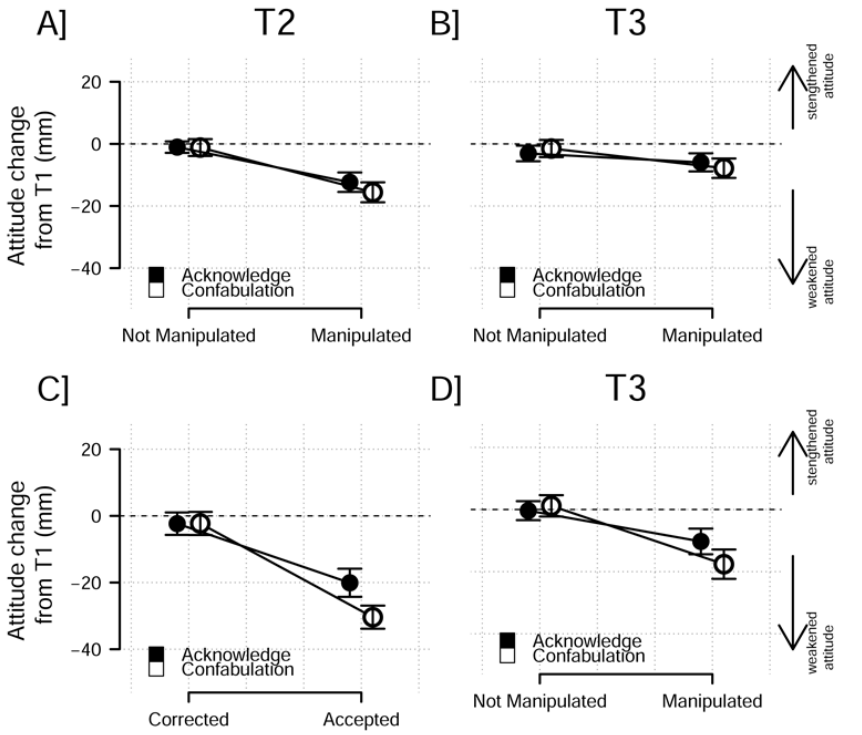

Figure 2 – Attitude change. Average attitude change compared to original (T1) ratings. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction towards the rating indicated by the false feedback. A-B Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. C-D Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Error bars are 95% CI.

Summary of Experiment 1. We investigated if false beliefs about one's own political attitudes, and confabulatory reasoning, could lead to lasting changes in these attitudes. We gave participants false feedback about some of their responses on a political survey, and asked half of them to merely acknowledge their responses, and the other half to also give verbal explanations to their responses. As expected, about half of the manipulations were accepted by the participants as being their own responses. Participants' future attitudes were strongly influenced by the false feedback, both directly following the manipulation and one week later. Additionally, we found that the attitude change was considerably larger if participants were asked to verbalize arguments, compared to only acknowledging its position.

Experiment 2

Experiment 2 was conducted with two aims in mind. The first was to run a high-powered direct replication of the findings in Experiment 1. The second was to investigate some possible factors that could moderate acceptance of the manipulation and the attitude change observed in Experiment 1. These factors are introduced below.

Our main finding in Experiment 1 was that attitude change is greater following confabulatory reasoning during the false feedback as compared to when only acknowledging the manipulated answer. One question that arises from this concerns what relation participants' confabulation stands to their later attitude change. One possibility is that merely engaging in the production of reasons gives an encoding advantage to the new attitude, leading to a greater shift in the participant's attitude. Alternatively, participants' attitude change might reflect a gradual depth of processing, as could be seen in the quantity of arguments given for the false feedback attitude. One simple unobtrusive measure is the amount of time participants spend engaging with the false feedback before answering the next question. If the magnitude of the participants' confabulatory argumentation is helping them cement their new attitude, we should expect the size of attitude change to be positively correlated with the length in time of their confabulatory engagement. To test this, we measured participants' talking time during the false feedback phase of the experiment.

A dominant view in much recent theorizing about information processing and reasoning, particularly in the political domain, has been that it is susceptible to the influence from strong motivational forces (Jost & Amodio, 2012; Taber, Lodge and Glather, 2001; Kunda, 1990; 1987). On this view, implicit motives, such as the need to be right about an issue, or to behave according to one's ideological values, can shape the interpretation of political information and the construction of reasons for having a belief (Jost & Amodio, 2012). This type of inferred justification strategy is supposedly used when there is a discrepancy between a belief and the external evidence contradicting the basis of that belief, and may help explain how people evaluate facts (Ditto & Lopez, 1992) and why some people label news as fake if they come from media houses with a political agenda opposite to their own (Flynn, Nyhan & Reifler, 2017). In our study, participants faced a dilemma of sorts when viewed through a motivational lens. On the one hand, they should be motivated to defend their initial political

attitudes which will, by definition, conflict with the false feedback. On the other, they should be motivated to defend their stated attitude, i.e., whatever is presented to them as being their own attitude. To investigate the impact of global political beliefs on level of acceptance and attitude change, we therefore included a general measure of political involvement and a left- to right-wing ideology scale.

Recently, motivated cognition in politics has also been related to peoples' cognitive style. One common measure is the Cognitive Reflection Test (Frederick, 2005), which is hypothesized to capture individual differences in reflexivity and critical reasoning (Bialek & Pennycook, 2017; Pennycook & Ross, 2016). Kahan (2013) found that high CRT scores associated with greater propensity to engage in politically motivated reasoning. Similarly, higher CRT scores were also found to predict the ability to discern fake news (Pennycook & Rand, 2017). While the false feedback presented to participants in our experiments is not exactly "fake news", it is counter-factual and runs against their prior attitudes. Hence, we can expect that higher CRT scores should correlate with correcting the false feedback.

In sum, we attempted a direct replication of our findings from experiment 1, adding measures of confabulatory reasoning, political attitudes and a CRT task.

Method

Participants. We recruited a total of 264 participants based on prior power calculations indicating that 240 participants would give high power to detect the crucial Correction and Confabulation condition interactions (>95%). Power was calculated based on the regression coefficients for the model estimated in Experiment 1 including Confabulation condition and Correction as factors analyzing attitude change for manipulated trials. We simulated data based on the estimated random and fixed effects, as well as the correction rates observed in Experiment 1 (Gelman & Hill, 2007). Thirty-two participants failed to show up for the T3 measurement or experienced equipment malfunction. The final sample therefore consisted of 232 participants (146 male, 85 female, 1 not identified), with an age range of 18-52 and average age M = 23.6 (SD = 4.6).

Participants received two cinema vouchers in exchange for their participation in two experimental sessions, roughly one week apart (average 6.8 days (SD = 0.9)). Participant information and debriefing followed the procedures described for Experiment 1. The experiment was approved by the Lund University Ethics board, D.nr. 2016-1046.

Materials and design. The choice blindness and attitude change setups were identical to Experiment 1 (a combination of the STS and paper-pen surveys), including the political statements. CRT, political involvement, and left-right ideology were assessed on additional paper surveys. CRT consisted of the following questions, presented on separate pages: (1) *A bat and a ball costs $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?* [answer in cents] (2) *If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?* [answer in minutes] (3) *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?* [answer in days][1]. Political involvement was assessed with the following items: (1) *In your daily life, how engaged in political issues would you say that you are?* (2) *Are you engaged in any of the following: (a) political party, (b) environmental organization (such as Greenpeace) (c) school organization (such as a teacher association)?* [yes/no]. Left-right ideology was assessed using a scale with endpoints going from left to right. Further, participants stated their education level, education subject, age and gender. The political-, educational-, and demographical items were assessed on the final page. Just as in Experiment 1, a tape recorder was used to capture the verbal reports, and a timer used to clock speaking time.

Procedure. Experiment 2 followed exactly the same procedure as Experiment 1, with two extensions. First, questionnaires measuring CRT, political involvement and ideology were administered during the final session (T3), after the participant had completed the political attitude surveys, but before the debriefing. Second, the experimenter timed the participants' argumentation/confabulation using a timer on the computer. This way, the additional measures in Experiment 2 were unobtrusive and did not interfere with the direct replication of Experiment 1.

Analysis. We followed the same analytical strategy as for Experiment 1 with two additions. First, we also estimated random effects (intercept and slopes) grouped by stimulus ID to improve the generalizability of our estimates. Again, random effects were entered as maximal or the maximal that would converge. Second, to provide combined estimates of the effects from both experiment, we conducted

[1] The CRT problems were Swedish translations of the questions used in Frederick (2005).

an analysis of our main findings on the combined dataset using Bayesian estimation techniques of the maximal multi-level model using the *brms* package (Buerkner, 2016). For information about priors, see the Supplemental Material.

Results

Correction of manipulated responses. Of the 464 manipulated (M) trials, 234 (50.4%) were corrected by the participants, meaning 49.6% were accepted. Average by participant correction rate was 1.0 trials (*SD* = 0.8). Sixty-eight (29%) participants made no corrections, 94 (41%) made one correction and 70 (30%) made two corrections. All participants and trials were included in the analyses.

Predictors of correction. Participants corrected 55.6% of manipulations in the Confabulation condition, while participants in the Acknowledge condition corrected 45.7% of manipulations. Average attitude strength was *M* = 26.4, *SD* = 15, on a 0-50 scale where 0 represents the indifference point. Participants also rated how confident they felt about each response. Average confidence was high with an average of 68 out of 100 (*SD* = 25). Confidence was higher for Corrected trials (*M* = 77, *SD* = 20) than for Accepted trials (*M* = 58, *SD* = 24; Welch t-test $t(307.22) = 8.66$, $p = 2.73*10^{-16}$)). Confidence was highly correlated with attitude strength, *r* = .71, 95% *CI* [.68, .74].

We analyzed the effects of nine possible predictors on the probability of correcting the manipulation, three were the same as analyzed in Experiment 1: Confidence, Attitude strength, and, Confabulation condition. Six were added in Experiment 2: participant political involvement, membership in political party, environmental organization or school organization, left-right political attitude and CRT (Cognitive Reflection Test) score. Average political involvement was fairly high, 51 of 100 (*SD* = 21). Membership in organizations was low: 7.8% of participants were members of a political party, 8.7% of an environmental organization and 5.2% of a school organization. Average political attitude on a left-right scale, where 0 is extreme left, 50 is neutral, and 100 is extreme right was *M* = 35, *SD* = 22. For CRT we sampled an even distribution of scores; 32% of participants answered zero questions correct, 28% 1 question, 20% 2 questions and 20% all 3 questions correct. The average score was *M* = 1.3.

All variables were entered in a multi-level regression model together with the interaction between Confidence and Attitude strength. All continuous variables

were standardized, except CRT score which was mean centered. Organization membership variables were also mean centered, with positive values indicating membership. Confabulation condition was coded (-.5 = Acknowledge, .5 = Confabulation). We found four significant predictors of correction. Participants' CRT scores ($\chi^2_{(1)} = 7.76$, $p = .0054$; $b = 0.41$, SE = 0.15), Confidence ($\chi^2_{(1)} = 5.97$, $p = .015$; $b = 0.69$, SE = 0.28), and Attitude strength ($\chi^2_{(1)} = 7.84$, $p = .0051$; $b = 0.79$, SE = 0.28), all positively predicted increasing probabilities of correcting the false feedback. Participants' Left-Right attitudes negatively predicted probability of correcting the false feedback ($\chi^2_{(1)} = 7.22$, $p = .0072$; $b = -0.45$, SE = 0.17), meaning that highly left-leaning participants made more corrections compared to other participants (see Figure S1). The remaining predictors were non-significant (see Table 2), and marginal model $R^2 = .37$.

Effect of manipulation and correction on future ratings. We wanted to see if accepted manipulated ratings would influence future ratings of the same issue. We repeated the analyses reported for Experiment 1 above. For brevity we only report the critical findings here and report the full analysis in the Supplemental Materials. We replicated our findings from Experiment 1 and found once again a large weakening of original attitudes for T2 manipulated (M) trials ($\chi^2_{(1)} = 45.84$, $p = 1.29*10^{-11}$; $b_M = -12.1$mm, $SE = 1.6$), which decreased during T3 ($\chi^2_{(1)} = 12.07$, $p = .00051$; $b_{T3*M} = 4.8$mm, $SE = 1.4$). Similarly, when comparing corrected and Accepted trials only, we found, consistent with our first main hypothesis and our findings in Experiment 1, a large weakening of original T2 attitudes for accepted (A) trials ($\chi^2_{(1)} = 41.45$, $p = 1.2*10^{-10}$; $b_A = -20.9$mm, $SE = 2.8$), which decreased somewhat at T3 ($\chi^2_{(1)} = 14.01$, $p = .00018$; $b_{T3*A} = 7.1$mm, $SE = 1.9$).

Qualitative shifts in position. We examined the proportion of the trials that crossed the mid-line of the attitude spectrum, indicating a qualitative shift compared to the original T1 attitude. In T2, 67% of responses represented such a shift for Accepted trials, compared to 6% for Corrected trials and 13% for Non-Manipulated trials. In T3, where the attitudinal effects of the manipulation were attenuated, 47% of responses were still qualitatively shifted for Accepted trials compared to 8% for Corrected trials and 17% for Non-Manipulated trials. These findings mirrored those of Experiment 1.

Effect of confabulation on future ratings. Next, we analyzed the effect of confabulation condition (acknowledge or confabulation) on attitude change. We first contrasted manipulated and non-manipulated trials (see also Fig. 3A-B). Our findings were largely consistent with those of Experiment 1. We found no main effect of Confabulation ($\chi^2_{(1)} = 1.24$, $p = .27$; $b_{CONFAB} = 0.02$mm, $SE = 2.3$) nor any interaction with Manipulation ($\chi^2_{(1)} = 2.97$, $p = .085$; $b_{M*CONFAB} = -3.4$mm, $SE = 3.7$) , Time ($\chi^2_{(1)} = 0.00$, $p = .99$; $b_{T3*CONFAB} = 1.1$mm, $SE = 2.0$), or three-way interaction ($\chi^2_{(1)} = 0.50$, $p = .48$; $b_{M*T3*CONFAB} = -2.1$mm, $SE = 2.9$). The remaining effects and coefficients were highly similar to those reported for Experiment 1 (see Table S1). Model marginal $R^2 = .09$.

Table 2 – All estimated regression coefficients and their standard error for mixed-model analysis of correction from Experiment 2. For all predictors Wald Chi-square and p-values are also reported.

Effect	Estimate	Standard error	Wald χ^2 (df=1)	P-value
Intercept	0.05	0.26	-	-
Political involvement	-0.05	0.27	0.04	.84
Party member	1.32	0.79	2.81	.094
Environmental org. member	1.55	1.40	1.22	.27
School org. member	0.21	0.80	0.072	.79
Left-Right attitude	-0.47	0.18	7.16	.0075
CRT score	0.41	0.14	7.93	.0049
Confidence	0.70	0.27	6.84	.0089
Attitude strength	0.73	0.29	6.51	.011
Confabulation condition	0.20	0.36	0.32	.57
Confidence X Attitude strength	-0.10	0.23	0.20	.65

Next, we conducted the crucial test if attitude change differed by Confabulation condition and Correction within the manipulated trials. Participants displayed small directional attitude change at T2, corrected trials in the acknowledge condition ($b_{intercept} = -3.6$mm, $SE = 2.1$), and further shifted slightly more in the confabulation condition for T2, Corrected trials ($b_{CONFAB} = -1.4$mm, $SE = 2.8$; see Figure 3C), with similar results for T3 trials ($b_{T3} = 0.6$mm, $SE = 2.2$; see Fig. 3D). For the accepted (A) trials, there was a large directional attitude change ($b_A = -16.3$mm, $SE = 2.4$, $\chi^2_{(1)} = 63.3$, $p < 1.8*10^{-15}$). The main effects of Condition ($\chi^2_{(1)} = 2.16$, $p = .14$), and Time ($\chi^2_{(1)} = 14.62$, $p = .00013$), were, again, qualified by interactions between Correction and Condition ($\chi^2_{(1)} = 4.78$, $p = .029$) and

Correction and Time ($\chi^2_{(1)} = 5.04$, $p = .025$). As expected according to our second main hypothesis, and from Experiment 1, the directional changes of accepted trials were accentuated in the confabulation condition at T2, meaning a further weakening of the original attitude ($b_{A*CONFAB}$ = -10.5mm, SE = 5.6). The attitude change was attenuated in T3 (b_{A*T3} = 8.3mm, SE = 3.9). The interaction between Condition and Time ($\chi^2_{(1)} = 0.00$, $p = .98$; $b_{CONFAB*T3}$ = 0.8mm, SE = 3.0) and the three-way interaction, were not significant ($\chi^2_{(1)} = 0.17$, $p = .68$; $b_{A*CONFAB*T3}$ = -1.7mm, SE = 4.1). Model conditional $R^2 = .24$.

Effect of confabulation length on attitude change. In the Confabulation condition, we additionally measured how long participants took while stating reasons for the presented attitude. Confabulation Length ranged from 36 to 255 seconds, with an average of M = 93s, SD = 39s. To analyze the effects of Length on attitude change we subset the data from the Confabulation condition depending on if the false feedback was corrected or accepted. The reason for doing so is that Length will have slightly different meaning depending on if the false feedback was accepted or not. For each subset we regressed Length, standardized, together Time on Attitude Change.

For accepted trials, Length captures the amount of time participants spend giving confabulatory reasoning for their presented attitude. For these trials, while we found that the estimates were in the expected direction, i.e. longer Length increases attitude change, the magnitude of the estimates was both small and non-significant (b_{LENGTH} = -0.2mm, SE = 2.9; $\chi^2_{(1)} = 0.06$, $p = .81$; $b_{LENGTH*T3}$ = -1.2mm, SE = 2.4; $\chi^2_{(1)} = 0.23$, $p = .63$). For corrected trials, however, Length captures both confabulatory reasoning as well as the time it takes for them to correct the presented attitude and enter a new one onto the tablet.

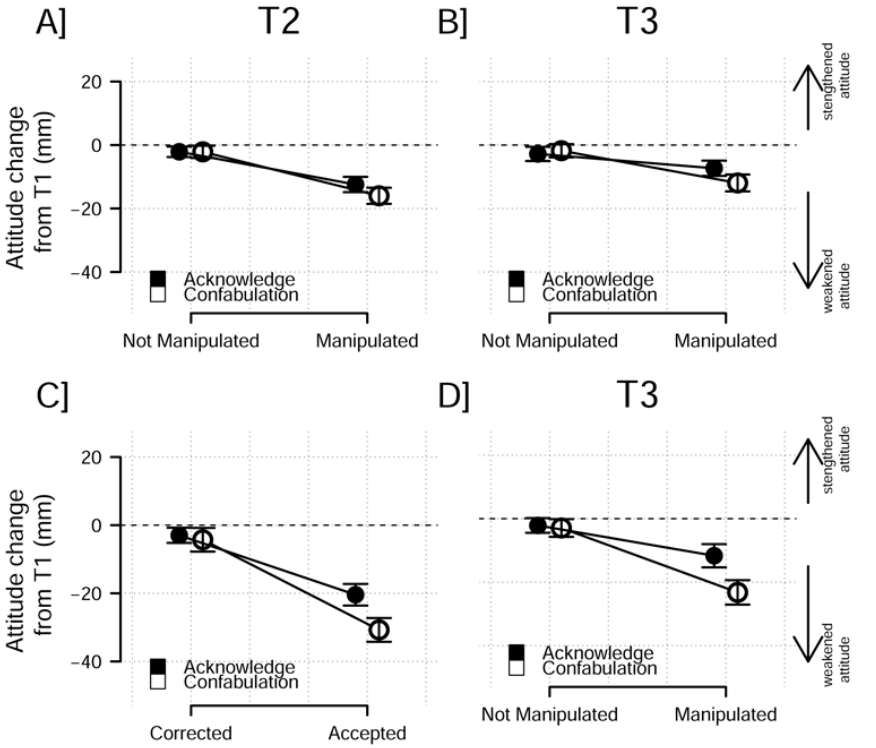

Figure 3 – Attitude change. Average attitude change compared to original (T1) ratings in Experiment 2. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction towards the rating indicated by the false feedback. A-B Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. C-D Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Error bars are 95% CI.

Here we found a main effect of Length (b_{LENGTH} = -4.4mm, SE = 1.6; $\chi^2_{(1)}$ = 9.04, p = .0026), such that participants shifted their attitudes more in the directions of the manipulation the longer time they spent engaging with the false feedback, even if they ultimately corrected the presented attitude. There was no interaction effect of Length and Time ($b_{LENGTH*T3}$ = 0.1mm, SE = 1.5; $\chi^2_{(1)}$ = .006, p = .94), nor any significant effect of time (b_{T3} = 1.6, SE = 1.5; $\chi^2_{(1)}$ = 1.22, p = .27). The intercept, reflecting attitude change at T2 at average Length, was estimated as ($b_{intercept}$ = -5.1mm, SE = 2.5).

Possible moderators of attitude change. We examined three additional potential moderators of the attitude change observed, participants' CRT score, political involvement and left-right attitude. All measures were entered into separate regressions together with Correction, Condition and Time. No effects involving any of the candidate variables reached significance (all *p*s > .066). We report all coefficients and p-values from all three models in Tables S2-4.

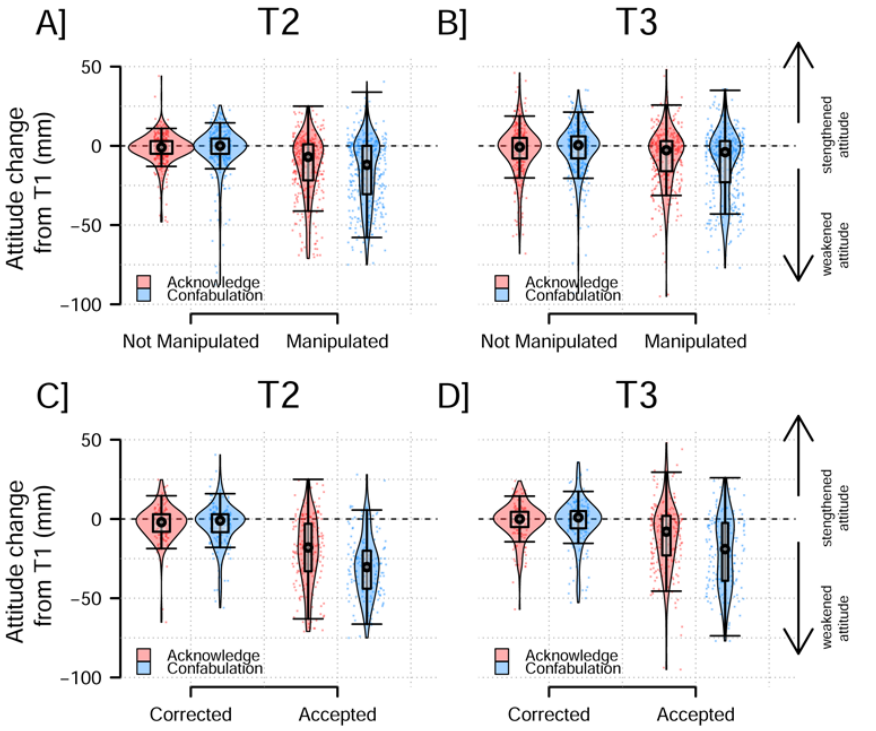

Figure 4 – Data from both experiments. Attitude change compared to original (T1) ratings. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction towards the rating indicated by the false feedback. A-B Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. C-D Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Points represent individual trials. Boxplots depict median (large circle), 25th and 75th quantile (box edges) values, as well as 1.5*interquartile range (hinges).

Bayesian estimation of effects from both Experiments. Finally, we combined the data from Experiment 1 and Experiment 2 and analyzed them using Bayesian multilevel regression estimating attitude change for Corrected and Accepted trials

together with Time and Confabulation condition. This provides our best estimates of the effects of our main findings and of the posterior uncertainty surrounding our estimates. The model was fit using the full random effects structure grouped by both participant and question ID. Figure 4 shows the combined data from both experiments.

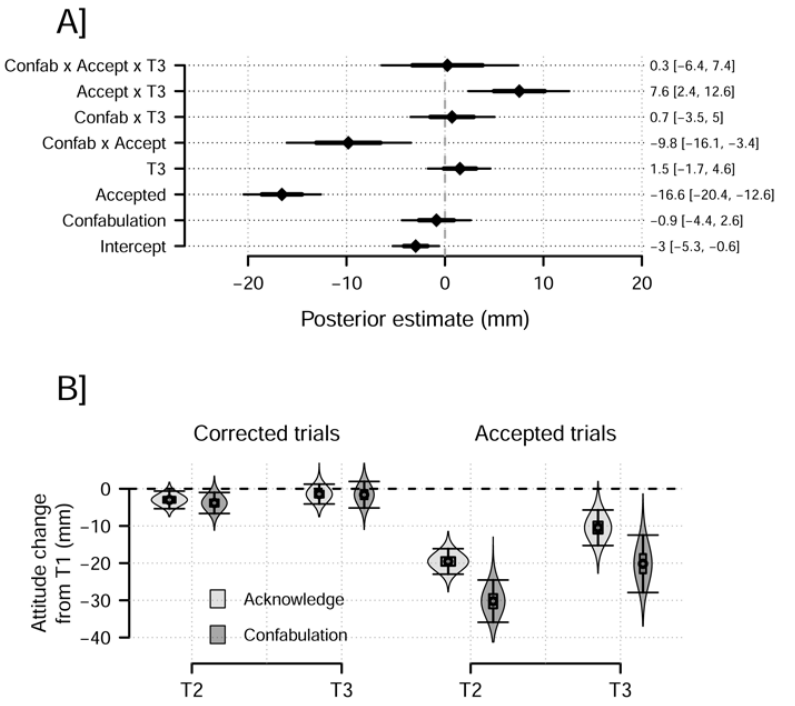

Figure 5 – Results from Bayesian regression. A] Posterior estimates from Bayesian regression combining data from Experiment 1 and Experiment 2 from manipulated trials only. Estimates reflect the coefficients contribution to attitude change measured as a difference from the original (T1) ratings. A negative difference indicates a weakening of the original attitude (in the direction of the false feedback). The reference level captured by the intercept reflects attitude change for Corrected trials in the Acknowledge condition at T2. All regressors were dummy coded. Points represent the mean posterior estimate; thick bars represent the standard deviation of the posterior and thin bars the 95% credible interval. The numerical column displays the mean of the posterior and 95% credible intervals. B] Violin plots depicting distribution of posterior predictions from a Bayesian regression model combining data from Experiment 1 and Experiment 2. Estimates reflect predicted attitude change compared to original (T1). A negative difference indicates a weakening of the original attitude (in the direction of the false feedback). Left panel depicts Corrected trials and right panel depicts Accepted trials. Points represent the mean posterior prediction. Boxes show the inter-quartile range (IQR) and hinges 1.5*IQR.

Figure 5 shows the results from the Bayesian regression, in panel A displaying the regression coefficients mirroring the reporting from the separate analyses provided above. In panel B posterior predictions of the average attitude changes are displayed for Corrected and Accepted trials.

Discussion

In two experiments, we investigated if false feedback concerning specific responses to political statements on a survey would influence later attitudes towards these issues. We found that half of the manipulations were accepted by the participants as being their own responses. Participants' responses were strongly affected by the false feedback, both in a session directly following the manipulation and one week later. In both experiments, we found that attitude change was much larger if participants were asked to reason about why they had stated the attitude falsely presented as their own compared to when only acknowledging its position.

Correction of the false feedback

An important part of any experiment involving the Choice Blindness Paradigm concerns the correction or acceptance of the false feedback. In this study we found that about half of manipulated responses were corrected by the participants, which is in line with our previous results in the moral and political domains (Hall, Johansson & Strandberg, 2012; Hall et al., 2013). Naturally, participants were more likely to correct a manipulated rating if their original response was extreme, and if the confidence rating regarding the attitude was high, however this was not predictive of the size of the ensuing attitude change. To get a better understanding of what increases the likelihood of a manipulation to be accepted or corrected, we added several related individual difference measures. In Experiment 2, participants reported their degree of political involvement, and where they would place themselves on the left-right spectrum. They also completed the Cognitive Reflection Test (CRT; Frederick, 2005), which is a short measure of reflexivity and critical reasoning.

We found no correlation between level of correction and self-rated political involvement. This is noteworthy, given the common assumption that increased political involvement also entails increased political awareness and more stable

attitudes (Zaller, 1992; Converse, 1964), and how the result contrasts with previous findings from our own lab (Hall et al., 2013; Strandberg, Björklund, Pärnamets, Hall & Johansson, 2018). However, political orientation on a left-right political ideology scale predicted correction, such that more left-leaning participants had higher rate of correction. However, this effect is probably best explained by the fact that more participants rated themselves to be strongly left compared than participants being strongly right (see distribution in Figure S1).

It has recently been found that there is a positive correlation between CRT score and ability to differentiate between real and fake news (Pennycook & Rand, 2017), as well as between CRT and measures of politically motivated cognition (Kahan, 2013). Considering this research, and the basic assumption that CRT captures analytic skill, we hypothesized that it would correlate with level of correction. This is also what we found, with participants scoring higher on CRT also having a higher likelihood of correcting the false feedback. Few individual difference predictors of correction have been found in previous research using the CBP (McLaughlin & Somerville, 2013; Sauerland et al., 2016; Strandberg et al., 2018, but see Aardema et al., 2013), making this result of general interest. More research is needed to establish which mechanism is captured by CRT in this context; if it is memory of prior answers, or more elaborate belief structures, or some other factor.

Influence of false feedback on future attitudes

As a backdrop to the false feedback manipulations in our study, and given the debate we outlined in the introduction between stable and flexible attitudes (e.g. Alvin, 1994; Bishop, 2005; Converse, 1975; Gerber et al. 2011; Haidt, 2001; Hall et al., 2013; Hatemi et al. 2009; Hooghe & Wikenfeld, 2007; Sears, 1999; Zaller, 1992), it is important to note that our participants generally displayed stability in their attitudes. For the non-manipulated trials there were no attitude shifts during the first follow-up, and one week later, during the second follow-up, these responses remained at their original positions. Generally, this was the case also for the trials where the participants corrected the false feedback.

In contrast, for the manipulated trials in both experiments, we found that participants' attitudes following the first session, as well as one week later, were shifted in the direction of the false feedback. The observed changes are consistent with previous work demonstrating preference change through choice using

various false feedback procedures (Izuma et al. 2015; Janis & King, 1954; Johansson et al. 2014; Luo & Yu, 2016; Sharot et al. 2012). However, our findings are noteworthy given the prior mixed evidence for more enduring changes in these paradigms (Sharot et al., 2012; Taya et al., 2014). In addition, prior studies have concerned preferential binary choices between pairs of faces and abstract images, or ratings of near equally preferred holiday destination, or hypothetical estimations of future events. To avoid these problems, we employed a more ecological procedure in the form of a political attitude survey focusing on specific, current political issues. This is not only a domain of great general importance, but one where preferences are supposed to be more resilient to change (Bartels, 2002; Gerber et al. 2011; Hatemi et al. 2009; Hooghe & Wikenfeld, 2007; Sears & Funk, 1999), as we also saw with the non-manipulated trials in our experiments. The specificity of the political questions, together with our confrontation procedure which required participants to both read the statement and the presented rating, suggests that the changes observed cannot be explained as being due to any vagueness in the targeted preference statements or a change in abstract values rather than specific attitudes as in some of the past research (e.g. Rokeach, 1971).

In both of our experiments, the average observed changes were large. The differences in ratings between Session 1 compared to Session 2 reached almost a full quarter of the length of the rating scale, and in most of Accepted trials these shifts crossed the mid-line (i.e. clearly defining the position as different from the original attitude). A week following the manipulation, the combined estimates from both experiments indicates that the attitude changes linger between about 10mm and 20mm for the accepted trials (Acknowledge and Confabulation conditions respectively, see Fig. 5). These effect sizes are notable when for example compared to those of around 10 points (of 100) found by Broockman and Kalla (2016) using a considerably longer and more involved intervention. The attitude changes were obtained absent of any reinforcement following the false feedback manipulation; the participants only viewed the manipulation once, and then immersed themselves in their ordinary life for a full week, with their usual sources of information and personal political biases. Even in the confabulation condition, the experimenter only asked the participants to explain the reasons behind their (manipulated) attitudes, and avoided further engagement in the argumentation. Considering this, the findings here present a strong demonstration of the power of even brief false feedback to engender attitude changes.

Confabulating about false feedback influences future responses

To investigate confabulation as a possible vehicle of attitude change, we varied the amount of confabulation participants gave in response to the manipulated ratings. In both experiments, we found that participants who had been asked to explain their responses, compared to those who merely acknowledged their (manipulated) attitude, showed larger attitude changes, both shortly following the manipulation and one week later. The average increase in rating difference was around fifty percent in the confabulation condition compared to the acknowledge condition at T2 and almost twice as large at T3, representing a considerable increase in relative effect size. This shows how the perception and verbalization of one's own reasoning can influence one's attitudes (cf. Barden & Tormala, 2014; Tormala & Petty, 2002), but as far as we know, the effect of confabulatory reasoning in facilitating attitude change is previously unstudied.

In the analysis of the confabulation condition in Experiment 2, we also looked at trial-based speaking time as an estimate of confabulation length. Using this more fine-grained measure, we found no correlation between confabulation length and the magnitude of attitude change in the accepted manipulated trials. This indicates that the exploratory measure of time taken during confabulation is not sufficient to capture what it is about confabulation that engenders attitude change. This is notable given previous research showing that differences in time spent merely thinking about an object can have varying influence on the attitudes towards that object (Clarkson et al. 2011). Testing a greater span of measures, including various forms of content and semantic analysis will be necessary to fully explain the details of the effect confabulation have on attitude change. In the corrected trials, however, we found a correlation between confabulation length and attitude change, such that the longer time the participants spent engaging with the false feedback the more they shifted in the manipulated direction. Our interpretation, based on informal observations, is that these participants often start constructing arguments for the manipulated position before instead backtracking to correct the presented attitude. This indicates that under some circumstances, even small amounts of confabulation can influence a person's beliefs.

While it is important to acknowledge that similar findings have been reported in the literature on self-persuasion using other methods, such as perspective taking (Broockman & Kalla, 2016), imagination (Carroll, 1978; Gregory et al. 1982; Watts, 1967), or counter-attitudinal argumentation (Lord & Lepper, 1984;

Mussweiler et al. 2000; Watts, 1967), these approaches all suffer from different limitations. In the traditional self-persuasion experiments, participants' attitudes are often compared to control groups (Watts, 1967), the original attitude is established several months prior to the experiment (King & Janis, 1956), or they are asked to assess their own attitudinal change (Lord & Lepper, 1984), resulting in uncertainty about what the participants' original attitudes were, and if any change has taken place. Crucially, in those experiments, participants are also fully aware that the attitude they are asked to express is not their own, and that the arguments they produce are hypothetical (e.g. Lord & Lepper, 1984; Janis & King, 1954), whereas in a CBP experiment participants believe the manipulated response to reflect their own true attitude. In the Confabulation condition, the participants produce arguments in favor of that attitude, just like they would have in an everyday interaction. This means that the present study removes the pressing problems of demand effects as an explanation for the observed attitude change, a concern present in most prior studies. Thus, a key contribution of the present study is that it provides clearer and firmer support for the hypothesis that processes of self-perception can be involved in attitude change.

Implications for attitudes and preferences

How do these findings relate to theories of attitudes and preferences more broadly? One lesson to learn from this study, in relation to the overarching tension between views of political attitudes as stable or flexible, is that both perspectives may capture important aspects of how such attitudes function. On the one hand, absent any manipulation, participants gave the same responses throughout the experiment, clearly indicating they had a stable set of political attitudes. On the other hand, the same participants exhibited large lasting attitude shifts after having accepted the false feedback.

We have previously shown that participants often accept false feedback about their political attitudes, thus revealing a previously undiscovered flexibility to reason beyond ideological labels (Hall et al. 2013). However, these attitude shifts were only measured at the moment of the feedback in terms of accepting the manipulation, but no subsequent follow-up attitude measurements were performed. Here we have extended that work, by showing lasting attitude changes measured during two follow-up elicitations, demonstrating that participants' initial attitudinal flexibility extends far beyond that of the immediate

confrontation with the false feedback. The attitude shifts at the latter stages of the study were not as large as those implied by the false feedback and accepted by the participants. This might signal an upper bound on attitude flexibility when translated into future behavior but might also be due to some form of gravitational pull from interlocking opposing attitudes, or counter pushing from everyday influences in the life of the participants (family and friends, selective news circles, etc.), or just simply noise induced by memory decay. If so, reinforcing the shifted attitudes, by for example exposing participants to extra arguments supporting their new position, would likely lead participants to coalesce their position closer to the one implied by the false feedback.

Another way of approaching the stable/flexible dichotomy is through the lens of inferential and constructivist accounts of preference and attitude formation (Ariely & Norton, 2008; Slovic, 1995; Warren, McGraw & Van Boven, 2011). On strong versions of such accounts, the act of choosing has a constitutive role in the genesis of a persons' preference set (Ariely & Norton, 2008; Slovic, 1995), to the point that some choices might reflect purely arbitrary influences on the preference (Ariely, Loewenstein & Prelec, 2003; Chater, Johansson & Hall, 2011). A more balanced view instead holds that preferences and attitudes are calculated to some degree at the time of choice (Warren, McGraw & Van Boven, 2011), recasting the question of stable versus flexible attitudes from a categorical one into a continuum. Instead it becomes key to discover what factors influence the degree of calculation and how that process is supported. In this vein, we have previously argued, based on preference changes for faces induced using the CBP (Johansson et al., 2014), that preference or attitude change in the CBP taps into a specific aspect of preference calculation, namely that preference calculation is supported by a process of self-perception. Inferences about one's own attitudes or preferences go via observations of the outcomes of past behavior. In other words, we often infer our own preferences much like we infer other peoples' preferences, by observing and interpreting our own overt behavior (Bem, 1967; Johansson et al., 2014). Once we believe we have stated some attitude, it follows that we should infer that we also hold that attitude. For example, recent work has demonstrated that once beliefs change, recollections of past beliefs become biased to match the current belief (Wolfe & Williams, 2017).

The proposition that participants rely on their beliefs about their past attitude ratings to inform their new ratings bears structural similarities with "options-as-information" theory, developed to account for some challenges to classical

decision theory arising from observed preference reversals in multi-attribute choice (Sher & McKenzie, 2014; Müller-Trede, Sher & McKenzie, 2015). The theory takes the form of a rational analysis (Oaksford & Chater, 1994), positing that by accounting for participants' prior beliefs going into a decision task, seemingly inconsistent patterns of preferences can be accommodated using a normative framework based on Bayesian updating. The decisions analyzed differ from the conditions of the present study, but nevertheless the question arises to what extent a framework such as "options-as-information", or broadly, a conception of decision makers as performing updating of their attitudes according to Bayesian normative theory, can be useful in explaining the observed attitude changes reported here.

One way of understanding participants' behavior at T2, in the accepted manipulated trials, is that they must reconcile two conflicting representations of their past attitudes. One being the trace of their original attitude, the second being the one presented during the false feedback confrontation. Depending on the weighting between these representations the participants' new attitudes should fall within that interval. If the weighting is equal the average attitude change should be half the average manipulation length, which is consistent with the data presented here, at least for T2. This suggests at least a tentative compatibility of the predictions of a theory like "options-as-information" and our findings, though more formal analysis and experiments specifically designed to test this would be required. Regardless, some rationalization of participants' behavior should be forthcoming. It is important for us to stress that while findings of choice blindness are counterintuitive by folk psychological reasoning, and perhaps the ensuing attitude changes reported here even more so, we do not take the findings presented here to demonstrate some fundamental irrationality on part of the participants. Rather it highlights the continuous and dynamic evolution of attitudes with respect to new information about oneself and one's beliefs.

That beliefs play a role aligns with a growing consensus across the decision sciences regarding the importance of memory processes for understanding value-based choice, where much recent work has focused on the influence of past episodes for the calculation of preferences (Bornstein, Khaw, Shohamy & Daw, 2017; Murty, FeldmanHall, Hunter, Phelps & Davachi, 2016; Shadlen & Shohamy, 2016). Using the CBP, we have previously shown that false feedback about choices leads to systematic distortions of participants' source memory, thus demonstrating that beliefs are formed resulting from acceptance of the false

feedback (Pärnamets, Hall & Johansson, 2015). This is consistent with other work showing source memory distortions when reasoning about past choices (Maher, Shafir & Johnson, 2009). Understood in the light of the present study, observations of our own past political survey responses lead to the inference that we hold those attitudes, this belief then influences later attitude construction when queried in the future.

Strengths, limitations and future studies

Future work should address questions arising both from the findings reported here and from limitations in the study design. We have demonstrated lasting attitude change following a simple false feedback manipulation. One route towards deepening our understanding of this finding is to investigate how far attitudes can be shifted. This would include follow-up sessions over longer periods of time as well as adopting a procedure where participants' false beliefs about their past attitudes were reinforced, perhaps by supplanting participants with additional arguments to buttress their new-found positions. Together this would allow us to better understand the interplay between original and implanted attitudes, and perhaps better model attitude shifts arising from malicious information sources in the world outside the lab. We have also argued that our attitude shifts are dependent on participants gaining false beliefs about their past attitudes. Hence, a key area to look at in future studies would be how false beliefs about past attitudes are integrated into participants' broader belief structure and how resulting changes in participants' memories about their own attitudes are maintained.

There is also the possibility to use CBP to explore other domains than politics, such as personal values, personality traits or character attributes. The case of values is particularly relevant to the present study as values are thought to underpin many political attitudes (Schwartz et al., 2012). While previous work applying CBP to moral questions (Hall, Johansson & Strandberg, 2012), including moral principles, indicates that also values should be susceptible to false feedback manipulations little is known how these effects translate back into attitudes or behavior. Studies have shown that values and value-relevant behavior can be susceptible to influence – for example by priming reasons or making the reasons more salient (Maio, Hahn, Frost & Cheung, 2009), and it is possible that accepting false feedback about values might recruit similar processes on

downstream behavior. Nevertheless, other value changes appear to occur on longer time-scales in relation to significant life events (Bardi, Buchanan, Goodwin, Slabu, & Robinson, 2014) or not at all (Manfredo, Bruskotter, Teel, Fulton, Schwartz, Arlinghaus & Sullivan, 2016). This leaves an important avenue for exploring if people can become, for example, more altruistic, fair, or patriotic, by making them adopt and argue for false beliefs about their values.

To increase the generalizability of our study, replicating it on a sample representative of the general population would be desirable. In a similar vein, assessing if the findings are limited to a WEIRD population is of importance (Henrich, Heine & Norenzayan, 2010). In this study we targeted political attitudes from two salient domains, education and environmental issues. Of course, this does not exhaust the spectrum of political topics, and it is important to assess if political attitudes behave the same across varying topics and questions, with various levels of polarization and acrimony. Nevertheless, unpublished data from studies conducted during the 2016 U.S election indicate that at least some of these effects are transferable to domains involving political leaders and generalize to a broader U.S population (Strandberg, Olson, Hall, Raz & Johansson, 2018).

As we see it, one of the clearest theoretical contributions of the current study is that we create a self-perception situation where the participants truly believe the manipulated attitudes to be their own, thus creating much stronger grounds for consequential self-inferences. As we detail below, this ought not to be interpreted as an irrational, or worse, even pathological, process, but instead as a reasonable inferential response to a peculiar array of evidence. However, more speculatively, some self-perception theories have suggested there might be a special relationship between attitudes and first-person authority, such that attitudes we endorse (either by acknowledgment or confabulation in the current study), also creates a special sense of agency or ownership of that attitude (see Carruthers, 2011; Moran, 2001; Martin & Pacherie, 2013). This phenomenological emotional component might then feed into or enhance the self-inferences seen in the CBP compared to previous paradigms. Unfortunately, there is nothing in the current design that allow us to disentangle these possibilities, so this remains as an exciting avenue for future research.

As detailed above, our preferable way of framing the self-inferential process would be in terms of Bayesian updating of beliefs. From this standpoint, the difference between the Acknowledge- and the Confabulation condition is one of

degree, where confabulation simply adds another layer of evidence to the self-inferences. Similarly, other theoretical frameworks of attitude change, such as the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986; Petty, Haugtvedt & Smith, 1995) could potentially help to explain the differences in change found between the Acknowledge and the Confabulation conditions. According to ELM, in the Confabulation condition, participants can be expected to make thoughtful and deliberate considerations of the arguments they generate. This would allow them to engage in deeper information processing compared to participants that simply acknowledge the stated attitude as their own, and this difference in information processing could be used to explain the different T2 and T3 effects between conditions.

Potentially, the matrix of evidence in CPB might also include our beliefs and expectations about *other* people, and their reactions to our opinions - that is, part of the difference between the two conditions might reside in the confabulations functioning as a public *commitment* (as has been explored in the literature on conversational implicature (Brandom, 1994; Grice, 1975). In future studies, this would be an interesting dimension to explore, by creating contexts with potentially more or less social commitment, for example by comparing the role of a politician to an entertainer, or a teacher to a student.

Conclusions

In summary, the results presented here demonstrate attitude flexibility in the face of accepted false feedback about previously held positions and how confabulatory reasoning facilitates shifts away from the original position. These results were obtained studying political attitudes; a domain of central importance to public life. On the face of it, this might seem like a troubling result, showcasing the shallowness of our political attitudes (Converse, 1975; 1964; Zaller, 1992), and potentially exposing us to manipulation by malicious opponents. Even though our study was not an attempt at a practical canvassing effort, like Broockman and Kalla (2016), this possibility should not be downplayed. While scientific methods can sometimes be misused by unscrupulous individuals, we take issue with the interpretation that the current findings reveal inherent flaws in our attitudes. Indeed, why should it be considered an ideal to have attitudes so firmly chiseled and bounded that one would consistently notice all CB manipulations? This

position is only intelligible against a backdrop of a society where particularly firm opinions are held in reverie, and where undecideds and moderates are derided as "wishy-washers" and "flip-floppers". But this might be a harmful standard (cf. Hall et al. 2012; 2013). As we see it, the current run of hyper-polarization in politics is not only simple aggregation of individual attitudes but also a result of our larger views of what it is to hold an attitude. In times of information bubbles, fake news, political acrimony, and gridlock, we find it encouraging that a brief CBP intervention can nudge people to find support for positions other than those originally held. This opens up new perspectives for understanding across the political divide and serves as a reminder that people can demonstrate flexibility when they are induced to reason about complex political issues.

Context of research

The research reported in this article originated in our earlier work observing choice blindness for political attitudes as well as effects of choice blindness on later choices and memories for simpler preferential decisions. We were interested in testing if political attitudes could be changed by giving false feedback to participants about their own prior responses. Additionally, this allowed us to visit an underexplored aspect of the choice blindness paradigm: the role of the confabulatory statements participants make in support of the false feedback response. We hypothesised that if participants have formed a false belief about their past attitude, then confabulating reasons for that attitude should increase the change observed in their later responses. Key ideas for future work will be to compare similarities and differences in argument content and paralinguisic markers when defending manipulated versus non-manipulated responses. We will also investigate how the memory of past attitudes is influenced when false beliefs about one's attitudes are adopted. By implementing a self-inferential, constructivist approach to the study of political attitudes, we believe that this research can contribute to the understanding of mass opinion.

Supplemental materials: http://dx.doi.org/10.1037/xge0000489.supp

References

Anand, S., & Krosnick, J. A. (2003). The impact of attitudes towards foreign policy goals on public prefrences among presidential candidates: A study of issue publics and the attentive public in the 2000 U.S presidential election. *Presidential Studies Quarterly*, *33*, 31-71.

Aardema, F., & Johansson, P. (2014). Choice blindness, confabulatory introspection, and Obsessive-Compulsive Symptoms: A new area of investigation. *International Journal of Cognitive Therapy, 7*(1), 83–102.

Alwin, D. (1994). Aging, personality, and social change: The stability of individual differences over the adult life span. In D. L. Featherman, R. M. Lerner, & M. Perlmutter (Eds.), *Life-Span Development and Behavior* (pp. 135-185). Hillsdale, NJ: Erlbaum.

Ariely, D., & Norton, M. I. (2008). How actions create – not just reveal – preferences. *Trends in Cognitive Sciences*, *12*(1), 13–16.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73-106.

Barden, J., & Tormala, Z. L. (2014). Elaboration and attitude strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass*, *8*(1), 17-29.

Bardi, A., Buchanan, K. E., Goodwin, R., Slabu, L., & Robinson, M. (2014). Value stability and change during self-chosen life transitions: Self-selection versus socialization effects. Journal of Personality and Social Psychology, 106, 131-147.

Bartels, L. M. (2002). Beyond the running tally: Partisan bias in political perceptions. *Political Behavior*, *24*(2), 117-150.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1-48.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200.

Bernstein, D. M., Laney, C., Morris, E.K., & Loftus, E. (2005). False memories about food can lead to food avoidance. *Social Cognition*, *23*, 11-34.

Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods*, 1-7.

Bishop, G. F. (2005). The illusion of public opinion: Fact and artifact in american public opinion polls. Lanham, MD: Rowman and Littlefield publishers inc.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352*, 220–224.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices biases decisions for reward in humans. *Nature Communications, 27*(8): 15958.

Brandom, R. (1994). Making it explicit. Cambridge, MA: Harvard university press.

Buerkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bullock, J. G. (2011). Elite influence on public opinion in an informed electorate. *American Political Science Review, 105*, 496-515.

Burke, E. (1774*). On American Taxation.* Indianapolis, IN: Liberty fund.

Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology, 14,* 88-96.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford, UK: Oxford university press.

Chater, N., Johansson, P., & Hall, L. (2011). The non-existence of risk attitude. *Frontiers in psychology*, *2*, 303.

Clarkson, J. J., Tormala, Z. L., & Leone, C. (2011). A self-validation perspective on the mere thought effect. *Journal of Experimental Social Psychology*, *47*(2), 449–454.

Converse, P. (1975). Public opinion and voting behavior. In: F. Greenstein & N Polsby (Eds.), *Handbook of Political Science 4*. Reading, UK: Addison Wesley.

Converse, P. (1964). The nature of belief systems in mass publics. In: D. E. Apter (Ed.), *Ideology and Discontent*. New York, NY: The Free Press.

Druckman, J.N. (2004). Political preference formation: Competition, deliberation, and the (ir)relevance of framing effects. *American Political Science Review*, *98*(4), 671-686.

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

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*(4), 568-584.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, 38*, 127-150.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25-42.

French, L., Garry, M., & Loftus, E. (2009). False memories: A kind of confabulation in non-clinical subjects. In: W. Hirstein (Ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy*. Oxford, UK, Oxford University Press.

Fotopoulou, A., Conway, M. A., & Solms, M. (2007). Confabulation: Motivated reality monitoring. *Neuropsychologia*, *45*(10), 2180–2190.

Fox, J. & Weisberg, S. (2011). *An {R} Companion to Applied Regression*, Second Edition. Thousand Oaks CA: Sage.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science*, *14*, 265-287. doi: 10.1146/annurev-polisci-051010-111659.

Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology, 43*, 89-99

Grice, P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds) *Studies in Syntax and Semantics III: Speech Acts*. New York, NY: Academic Press.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS One, 7*(9), e45457.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, *117*(1), 54-61.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS One*, *8*(4), e60554.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814-834.

Hatemi, P. K., Funk, C. L., Medland, S. E., Maes, H. M., Silberg, J. L., Martin, N. G., & Eaves, L. J. (2009). Genetic and environmental transmission of political attitudes over a life time. *Journal of Politics*, *71*(3), 1141-1156. doi:10.1017/S0022381609090938.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences, 33*, 61-135.

Hirstein, W. (2009). Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy. Oxford, UK: Oxford University Press.

Hooghe, M., & Wilkenfeld, B. (2007). The stability of political attitudes and behaviors across adolescence and early adulthood: A comparison of survey data on adolescents and young adults in eight countries. *Journal of Youth and Adolescence, 37,* 155–167.

Isenberg, I. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50*(6), 1141-1151.

Izuma, K., Akula, S., Murayama, K., Wu, X. D., Iacoboni, X. M., & Adolphs, R. (2015). A Causal role for posterior medial frontal cortex in choice-induced preference change. *Journal of Neuroscience*, *35*(8), 3598–3606.

Janis, I. L., & King, B. T. (1954). The influence of role-playing on opinion change. *Journal of Abnormal and Social Psychology, 49*, 211–218.

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.

Johansson, P., Hall, L., Sikström, S., Tärning, B. & Lind, A. (2006). How something can be said about Telling More Than We Can Know. *Consciousness and Cognition*, *15*, 673-692.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making*, *27*(3), 281–289.

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion, 36*, 55-64.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407-424.

Kogan, N., & Wallach, M. A. (1967). Group risk taking as a function of members' anxiety and defensiveness levels. *Journal of Personality, 35*(1), 50-63.

King, B. T., & Janis, I. L. (1956). Comparison of the effectiveness of improvised versus non-improvised role-playing in producing opinion changes. *Human Relations, 9*, 177-186

Knowles, E. S., & Linn, J. A. (2004). The importance of resistance to persuasion. In E.S. Knowles and J.A. Linn (Eds.), *Resistance and Persuasion* (pp. 3-9). Mahwah, NJ: Lawrence Erlbaum Associates.

Kunda, Z . (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 408-498.

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology, 53*, 636-647.

Ledford, H. (2016, April 7). Door-to-door canvassing reduces transphobia. *Nature News.* doi:10.1038/nature.2016.19713

Lefcheck, J. S. (2015) piecewiseSEM: Piecewise structural equation modeling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*. *7*(5): 573-579. doi: 10.1111/2041-210X.12512

Lewis, G. J. (2018). Early-childhood conduct problems predict Economic and political discontent in adulthood: Evidence from two large, longitudinal UK Cohorts. *Psychological Science, 29*(5), 711-722.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*, *25*(6), 1198-1205.

Loftus, E. & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a questions. *Bulletin of the Psychonomic Society*, *5*, 86-88.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231-1243.

Luo, J., & Yu, R. (2016). The Spreading of alternatives: Is it the perceived choice or actual choice that changes our preference?. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.1967

McNair, B. (2017). *Fake news: falsehood, fabrication and fantasy in journalism.* London, UK: Routledge.

Maio, G. R., Hahn, U., Frost, J. M., & Cheung, W. Y. (2009). Applying the value of equality unequally: Effects of value instantiations that vary in typicality. *Journal of Personality and Social Psychology*, *97*(4), 598.

Manfredo, M. J., Bruskotter, J. T., Teel, T. L., Fulton, D., Schwartz, S. H., Arlinghaus, R., ... & Sullivan, L. (2017). Why social values cannot be changed for the sake of conservation. *Conservation Biology*, *31*(4), 772-780.

Martin, J. R., & Pacherie, E. (2013). Out of nowhere: thought insertion, ownership and context integration. *Consciousness and Cognition, 22* (1): 111–122. doi: 10.1016/j.concog.2012.11.012

Maher, M., Shafir, E., & Johnson, M. K. (2000). Misrememberance of options past: Source monitoring and choice. *Psychological Science, 11,* 132-138.McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making*, *8*(5), 577.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton, NJ: Princeton University Press.

Müller-Trede, J., Sher, S., & McKenzie, C. R. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision*, *2*(4), 280-305.

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General, 145*(5), 548-558.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*, 1142-1150.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608-631.

Pennycook, G. & Rand, D., G. (2017). *Who Falls for Fake News? The Roles of Analytic Thinking, Motivated Reasoning, Political Ideology, and Bullshit Receptivity* (September 12, 2017). Available at SSRN: https://ssrn.com/abstract=3023545

Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00009t

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205.

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength. *Attitude strength, 4*, 93– 130.

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C.D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1823-1828). Austin, TX : Cognitive Science Society.

Rokeach, M. (1971). Long-range experimental modification of values, attitudes, and behavior. *American Psychologist*, *26*(5), 453-459.

Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law*, *22*(4), 303-314.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., ... & Dirilen-Gumus, O. (2012). Refining the theory of basic individual values. *Journal of personality and social psychology*, *103*(4), 663.

Sears, D. O., & Funk, D. L. (1999). Evidence of long-term persistence of adults' political predispositions. *The Journal of Politics*, *61*(1), 1-28.

Sears, D. O. (1983). The persistence of early political predispositions: The roles of attitude object and life stage. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology* (Vol. 4). Beverly Hills, CA: Sage Publications.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron, 90,* 927-939. doi: http://dx.doi.org/10.1016/j.neuron.2016.04.036

Sharot, T., Fleming, SM., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting? *Psychological Science 23*(10), 1123–1129.

Sher, S., & McKenzie, C. R. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General*, *143*(3), 1127-1143.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*(5), 364-371.

Steenfeldt-Kristensen, C., & Thornton, I. M. (2013). Haptic choice blindness. *i-Perception*, *4*(3), 207-210.

Strandberg, T., Olson, J. A., Hall, L., Raz, A., & Johansson, P. (2018). *Opening American minds: False beliefs can induce open-minded evaluations of presidential candidates*. Manuscript submitted for publication.

Strandberg, T., Björklund, F., Pärnamets, P., Hall, L., & Johansson, P. (2018). *The self-transforming survey*. Manuscript in preparation.

Taber, M. and Lodge, C. S. (2013). *The rationalizing voter.* New York, NY: Cambridge University Press.

Taber, M., Lodge, C.S., and Glather, J. (2001). The motivated construction of political judgments. In J. Kuklinski (Ed.), *Citizens and Politics: Perspectives from political psychology* (pp. 198-226). New York, NY: Cambridge University Press.

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O.A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PLoS One 9*(9), e108515.

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me only makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, *83*(6), 1298-1313.

Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(2), 193-205.

Watts, W. A. (1967). Relative persistence of opinion change induced by active compared to passive participation. *Journal of Personality and Social Psychology, 5,* 4-15.

Wolfe, M. B., & Williams, T. J. (2017). Effects of text content and beliefs on informal argument evaluation. *Discourse Processes, 54*, 446-462.

Zaller, J. R. (1992). *The nature and origins of mass opinion.* New York, NY: Cambridge University Press.

Paper IV

Correction of manipulated responses in the choice blindness paradigm: What are the predictors?

Thomas Strandberg, Lars Hall, Petter Johansson, Fredrik Björklund and Philip Pärnamets

Abstract: Choice blindness is a cognitive phenomenon describing that when people receive false feedback about a choice they just made, they often accept the outcome as their own. Little is known about what predisposes people to correct manipulations they are subjected to in choice blindness studies. In this study, 118 participants answered a political attitude survey and were then asked to explain some of their responses out of which three had been manipulated to indicate an opposite position. Just over half (58.4%) of the manipulations were corrected. We measured extremity, centrality and commitment for each attitude, and one week prior to the experiment we assessed participants' preference for consistency, need for cognition and political awareness. Only extremity was able to predict correction. The results highlight the elusiveness of choice blindness and speak against dissonance and lack of motivation to engage in cognitively demanding tasks as explanations why the effect occurs.

Introduction

Choice blindness (CB) is a cognitive phenomenon indicating a dissociation between making a choice and its later justification. It highlights the limitations of our introspective capacity when reasoning about past choices. CB occurs when people receive false feedback about a choice they just made accepting the outcome as their own and reporting seemingly introspective (albeit confabulated) reasons for having made that choice (see Johansson et al., 2005 for details). CB has been reported for many domains and modalities, ranging from taste and smell preferences (Hall, Johansson, Tärning, Sikström & Deutgen, 2010) to eye-witness testimony (Cochran, Greenspan, Bogart & Loftus, 2018), and has been shown to affect both later memories and preferences (e.g. Strandberg, Sivén, Hall, Johansson & Pärnamets, 2018; Pärnamets, Hall & Johansson, 2015; Johansson, Hall, Tärning, Sikström & Chater, 2014). CB has also been applied to the study of attitudes and attitude change, an area of research where deliberation and introspection are often seen as important ingredients. In Hall, Johansson and Strandberg (2012) about 60% of manipulations to a survey on moral dilemmas were accepted by the participants' as being their own attitudes. Hall et al., (2013) reported similar findings for salient political issues in the run up for a Swedish general election. In that study participants not only changed their attitudes on political issues, but their actual voting intention was also affected in the direction of the false feedback. Notably, Strandberg and colleagues (2018) found that when participants accepted the manipulations to political attitudes, these shifted congruently with the false feedback when re-elicited one week later.

Although CB is ubiquitous, and undeniably relevant for the study of attitudes and decisions, little is known about what factors that predisposes people to correct the manipulated responses. So far, only a few studies have attempted to establish CB mediators, and thereby link the effect to other psychological constructs (e.g. Strandberg et al., 2018). However, no studies have focused purely on why people correct the false feedback. In this study, we aim to explore several factors that we have identified as meaningful for understanding why correction in the CB paradigm occurs, particularly in the domain of attitudes.

Subjective experience of attitude strength

One possible key to CB susceptibility could be in the relationship between the individual and the attitude itself. This is supported by the literature describing strong attitudes as "resistant to change, persuasion, and contextual influence" and weak attitudes as "unpredictable, malleable, and created in the moment" (Krosnick & Petty, 1995). Given this definition, it seems reasonable that correction of manipulations to attitudes should correlate with attitude strength. Here we tested three self-report measures adopted from Bassili's (1996) seminal work on attitude strength: extremity, centrality and commitment.

Extremity directly estimates how strongly a person agrees with an issue on a bipolar scale. Extremity, which is basically just the response to the survey item, is what Bassili calls an operative measure based on first order cognitive processing. Extremity is operative because, for example, the experienced valence of the extremity could be directly retrieved from memory and not the product of inference.

Centrality and commitment, on the other hand, are so called meta-attitudes. These are second order impressions of attitudes that rely on people to report on psychological properties not necessarily represented in long-term memory. As such, meta-attitudes are often inferred from sources more or less relevant to the strength of which the attitude is held. Centrality is described as tapping into the importance of an attitude and how it relates to personal values. Studies show that central attitudes are often more memorable and resistant to persuasion and contextual influence compared to peripheral attitudes (Holland, 2003; Pomerantz et al., 1995). Commitment is described as tapping into the confidence in an attitude: the conviction that the attitude is correct and valid. Commitment has been shown to moderate self-perception and contextual influence in attitudes (Holland, 2003; Pomerantz et al., 1995). Since these measures are meant to capture attitude strength – with strong attitudes being defined by their "resistance to change, persuasion, and contextual influence" – they should also correlate with correction of CB manipulations.

Variation in cognitive style

Another possibility is that aspects of the CB task might be experienced as rather cognitively demanding, such that some individuals may be more susceptible to CB than others due to being less motivated to perform them. Previous studies

have shown that individuals with a larger set of general analytic skill are more prone to correct the manipulations (Strandberg et al., 2018). Hence, measures capturing peoples' motivation to engage in cognitively demanding task, such as the *Need for Cognition* (NC; Cacioppo, Petty & Kao, 1984; Cacioppo, Petty, Feinstein & Jarvis, 1996) might also correlate with correction. NC is commonly used in attitude change research, where studies have shown that people with high NC tend to form attitudes that are more resistant to persuasion compared to people with low NC (Haugtvedt & Petty, 1992).

CB could also be affected by a consistency motive, which is the case for dissonance phenomena such as cognitive dissonance, cognitive balance, foot-in-the-door etc. These phenomena show that people often change either their behavior or their attitudes to appear consistent (cf. Festinger, 1957). One measure for estimating peoples' need to have consistent cognitions is *Preference for Consistency* (PFC; Cialdini, Trost & Newsom, 1995). Further, PFC has also been shown to predict if people change their attitudes due to social pressure or external demand (Bator & Cialdini, 2006). Thus, if CB share properties with cognitive dissonance phenomena; or if participants accept manipulations due to demand from the experimental situation, correction may correlate with the PFC score.

Variation in political awareness

We would also like to consider variation in political awareness, since much research in political science highlights political awareness as one of the most important factors when forming strong and resilient political attitudes (Zaller, 1992). Interestingly, recent CB studies involving political attitudes have yielded mixed results. In Hall et al. (2012) politically involved participants were more likely to correct the manipulations, and this was not found in Strandberg et al. (2018). However, since political awareness is supposed to determine how people select, interpret and internalize political information (Sidanius, 1988; Lusk & Judd, 1988) we continue to explore the relationship between various measures of political awareness and participants' behavior in a CB study involving political issues.

Thus, we set out to test if susceptibility to correct manipulated responses in CB could be predicted by any of the attitude strength measures, variation in cognitive style, or political awareness described above.

Method

Participants

A total of 128 (70 female) participants, with ages ranging from 18 to 64 years (M = 23.5, SD = 16.8), were recruited to answer a political survey. Sample size was predetermined based on previous CB studies (e.g. Johansson et al. 2005). Ten participants were excluded due to malfunctions with the experimental equipment. Thus, 118 participants remained for the final analysis. The participants were recruited through posters and flyers distributed at the university campuses of Lund and Malmö and compensated with a cinema voucher. At the start of the experiment, we described the general purpose of the study, but without telling the participants that some of their answers would be manipulated. Participants were informed that they could quit the experiment at any time, request their data to be erased, and still receive the cinema voucher. Participants were fully debriefed at the end of the experiment, before consenting to their anonymized data to be used by signing a consent form. All but six participants allowed their interviews to be recorded (leaving a total of 112 verbal recordings to be analyzed). The study was approved by the Lund University Ethics board, D.nr. 2008–2435.

Materials and design

Pre-test. One week before the main experiment, participants completed an online questionnaire assessing their demographics, political awareness, PFC and NC. PFC was assessed using the abbreviated 9-item version (Cialdini et al., 1995) with scales ranging from 1 (low consistency) to 9 (high consistency). The PFC questionnaire assessed the participants' internal and external consistency and included items such as: "It is important to me that my actions are consistent with my beliefs". For NC, we used the 18-item version (Cacioppo, Petty & Kao, 1984) with scales ranging from 1 to 9 where a nine gave four points and a one subtracted four points (five gave zero points, and so on). The NC questionnaire assessed the participants' attitudes towards effortful thinking, and contained items such as: "I usually end up deliberating about issues even when they do not affect me personally". Further, political awareness was established by assessing the

participants' political interest with a scale ranging from extremely uninterested (1) to extremely interested (9), and whether they were involved in any political party or organization (yes/no). Visit *https://osf.io/zsy47/* for a list of all measures and items.

Main experiment. After the pre-test, participants scheduled to partake in the main experiment being held one week later. It consisted of a questionnaire running on a tablet with a touch-based interface that the participants interacted with using a tablet pen. The experiment consisted of two parts: (1) responding to political issues, (2) explaining the responses, and ended with a full debriefing.

Procedure

Part 1 – responding to political issues. During the first part, participants responded to 12 sets of political issues with each set containing a political statement and corresponding six meta-attitudes; three centrality, such as "how important is this issue to you?", and three commitment, such as "how confident are you about your attitude towards this issue?" (visit *https://osf.io/zsy47/* to see all centrality and commitment items). The political issues were selected together with leading political scientists, and represented 12 of the most salient and important issues in Sweden at the time of the study (Table 1). As such, we believe that the vast majority of our participants were familiar with them. This was also confirmed by the verbal reports: most participants were able to intelligibly and knowingly discuss the various issues. Below each item were visual analog scales with endpoints at 0 and 100 (completely disagree to completely agree for the political statements and for example extremely unimportant to extremely important for the centrality item "importance"). The participants were instructed respond to each item by drawing a mark using the pen. They could change their responses as many times as they wanted by clicking a change icon located to the left of each scale, as well as toggle freely between the 12 sets of issues. The participants were left to complete the questionnaire at own, and told to inform the experimenter when finished.

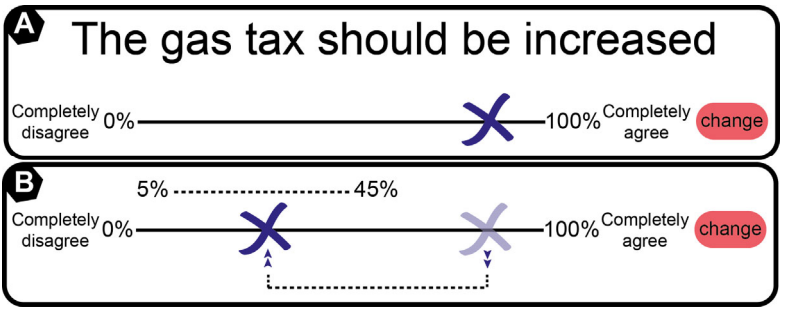

Figure 1 – To respond, participants drew an X on a scale going from completely disagree to completely agree (A). On manipulated trials, participant's X was surreptitiously moved from one side of the scale and then randomly placed on the other side (B). Participants could change their X as many times as they wanted by clicking 'change' (A-B).

Table 1 – The political issue statements.

1.	The gas tax should be increased
2.	A wealth tax should be reinstated
3.	The labor taxes should be lowered
4.	The monarchy should be abolished
5.	The government should run all elementary schools
6.	The punishment for violent crimes should be stricter
7.	The subsidized service for homework assistance should be abolished
8.	High schools should offer more applied and fewer theoretical courses
9.	Women should be recruited to company boards through affirmative action
10.	Private health care companies should be allowed to make profits in the welfare sector
11.	Copyright protected material from internet should be free to download for personal use
12.	The government should be allowed to monitor telephone conversations and internet traffic

False feedback and correction. When going over and explaining the responses, participants had received false feedback on three of the six trials. Trials 2, 4 and 6 had been manipulated by the tablet application to indicate a position opposite to the original (Figure 1). Trials 1, 3 and 5 were non-manipulated controls. The manipulation had two rules: move the participants' rating across the midline of the scale (with a minimum of 5 mm from the middle, i.e. ratings 45 or 55), and then randomly positioned on the opposite axis. If participants in any way indicated that their responses did not correspond with their views, or indicated that something was wrong, the experimenter would tell them that they could change their response if they wanted to, after which they could base their explanation on that response instead. Correction was operationalized when change was clicked and a new response drawn.

Analysis

Consistent with Bassili (1996) extremity was calculated by taking the absolute value of the deviation between a rating on the 100 point scale and the midpoint. All other variables are reported using their averages. Since attitude extremity, and the difference between the original rating and the manipulated rating, labeled 'manipulation length', are core features in CB studies using rating scales; we first tested how well these would predict correction. In our dataset, extremity and manipulation length were highly correlated, $r = .73$, $t_{(333)} = 19.6$, $p = 2.2*10^{-16}$. To address this we performed our analyses using decorrelated variables by transforming manipulation length to be the distance on the scale the manipulated attitude was moved *beyond* the midpoint. The resulting variables were independent, $r = -.028$, $t_{(333)} = -0.52$, $p = .61$. We then used these two variables to fit a baseline for the other predictor variables (i.e. meta-attitudes and cognitive style). We analyzed our data using mixed regression models including by participant varying intercepts and slopes. Models were estimated in a Bayesian framework using the brms package in R (Bürkner, 2016). Weakly regularizing priors were used for all parameters.

Results

On average participants were moderately interested in politics ($M = 6.0$, $SD = 2.1$) and about one fifth identified as politically involved ($M = 22.9$, $SD = 42.2$). As we can see in Table 2, extremity, centrality and commitment was rated fairly strong, averaging between 60 to 65 points of 100. The PFC score in our sample was similar to the 48.9 ($SD = 10.7$) that Cialdini et al. (1995) reported, and the NC score was similar to that reported in a recent meta-analysis of the NC scale ($M = 33.2$, $SD = 10.2$ (de Holanda & Wolf, 2018)).

Table 2 – Means and SD for the main predictor variables.

Predictor	Mean	SD
Extremity	29.2	13.9
Centrality	63.9	18.2
Commitment	65.0	20.1
NC	29.1	17.8
PFC	44.8	12.6

False feedback correction

Participants corrected 58.4% of the total 347 manipulations. Each participant was exposed to three manipulations and the average correction rate was 1.66 (*SD* = 0.98), with 15 participants accepting all manipulations and 27 participants correcting all. After correcting a manipulation participants were instructed to replace it with a new response. This corrected rating was on average placed within 9.43 points (*SD* = 11.7) of their original rating; or -4.45 points (*SD* = 14.4) when taking the direction of the corrected rating into account (defining a weakened new rating as a negative quantity and a strengthened new rating as a positive quantity). As in previous CB studies, correction did not vary as a function of sex, gender, age, or political party.

Predictors of correction

To test for predictors of correction we conducted mixed-effects logistic regression analyses using standardized variables. We first fit a baseline model consisting of extremity and manipulation length. This model (LOO = 402.77, *SE* = 14.86) indicated a large effect of extremity on correction (β = 1.77, *SD* = 0.32, *95% CI* = [1.17, 2.43], BF_{10} > 1.0*10^5), but only a smaller, uncertain effect of manipulation length (β = 0.53, *SD* = 0.28, *95% CI* = [-0.0043, 1.09], BF_{10} = 1.67), with the intercept estimated as β = 0.46 (*SD* = 0.19, *95% CI* = [0.10, 0.84]). See Figure 2 for the marginal posterior predictions of the attitude extremity and manipulation length.

We next fit a full model with all our candidate predictors: extremity, centrality, commitment, preference for consistency (PFC), need for cognition (NC), political involvement, political interest and manipulation length (LOO = 401.65, SE = 17.03). The estimated coefficients, their credible intervals and associated Bayes Factors can be found in Table 3. Marginal posterior predictions are depicted in Figure 3. Notably, when comparing the baseline and full model using LOO we found that the baseline model and the full model did not differ, with a difference of 1.12 (SE = 6.23), this is also mirrored in the estimates where there is little evidence that any of the added predictors are particularly successful at estimating correction.

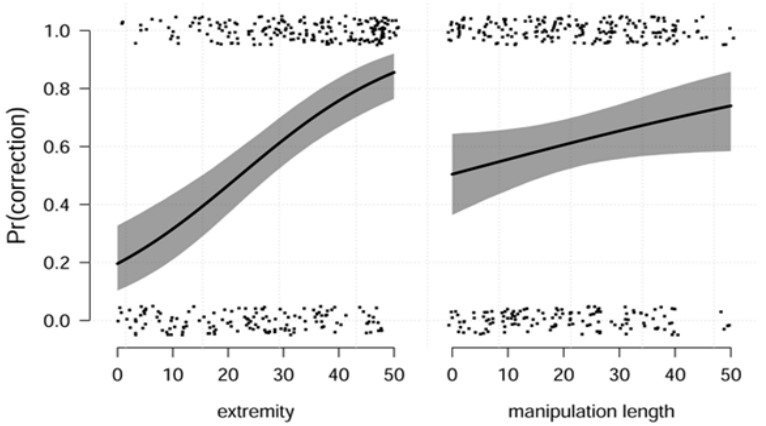

Figure 2 – Marginal posterior predictions from the baseline model. Predictions assume other variable held at its average value (0 for standardized predictors). X-axes renormalized to increase interpretability. Shaded regions indicate 95% posterior intervals.

Table 3 – Estimates and Bayes Factors from the full model.

Predictor	Est (β)	SD	95% CI	BF_{10}
(Intercept)	0.51	0.21	[0.12, 0,94]	-
Extremity	1.34	0.38	[0.06, 2.10]	333.69
Centrality	0.44	0.39	[-0.31, 1.23]	0.71
Commitment	0.69	0.39	[-0.06, 1.46]	1.78
NC	-0.07	0.40	[-0.86, 0.72]	0.40
PFC	-0.12	0.37	[-0.85, 0.59]	0.38
Pol.Involvement	0.71	0.47	[-0.21, 1.62]	1.45
Pol.Interest	-0.22	0.43	[-1.06, 0.65]	0.49
Manip.Length	0.59	0.31	[-0.0013, 1.20]	1.94

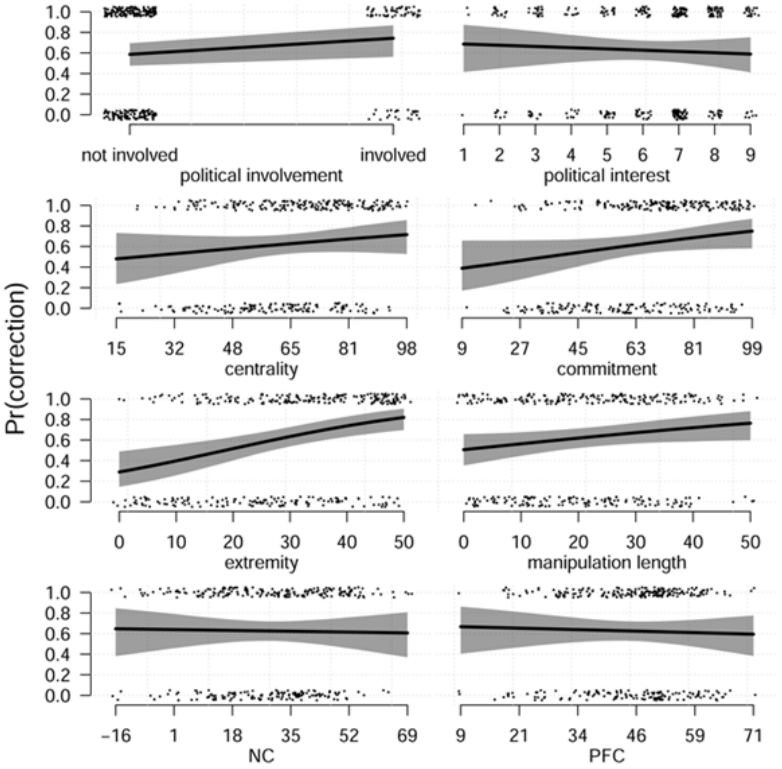

Figure 3 – Marginal posterior predictions from the full model presented with the same properties as Figure 2.

Predictors of correction types

On an exploratory note, we tried to better capture participants' subjective experience of correcting a manipulation. We conducted a simple classification of the reasons participants reported for wanting to correct. One independent rater listened to all the 112 recorded interviews and coded the different reasons participants gave when correcting a manipulation.

We identified three distinct types distributed evenly among the corrections: *internal attribution* (36.8%), when participants claimed to have misinterpreted the question, the scale, or something in the task; *external attribution* (33.9%), when participants blamed the experimental equipment; and *change* (29.2%), when participants felt they had spontaneously changed their minds about the

issue. Only for a few trials did participants report suspicion that their responses had been manipulated; these were categorized as external attribution. A second rater then classified a subset of 40 interviews; the raters agreed on 90% of the classifications. To test for determinants of the correction types we conducted a hierarchical multinomial logistic regression analysis using correction type as dependent variable and the predictors used in previous analyses. Since we were mainly interested in whether people attributed the wish to correct internally or externally, the change category was used as the reference level in the analysis. Consistent with previous findings, most variables were unable to predict whether participants would attribute correction internally (e.g. feeling that they had made a mistake) or externally (e.g. blaming the experimental equipment). However, we did find that the larger absolute difference between the original response and the manipulated response the more likely participants were to attribute correction internally ($\beta = 2.40$, *SD* = 0.57, *95% CI* = [1.03, 3.57], $BF_{10} = 1017.52$). We also found a small negative effect of political involvement, meaning that participants that were uninvolved politically were more likely to attribute correction externally ($\beta = -1.15$, *SD* = 0.77, *95% CI* = [-2.63, 0.37], $BF_{10} = 2.36$). However, the effect size of this latter finding was very small, but could potentially be a subject for future research.

Discussion

To summarize, we first assessed participants' preference for consistency, need for cognition, and political awareness; and one week later measured attitude extremity, centrality and commitment on a questionnaire containing 12 political issues. Participants were then asked to explain their responses to six of these issues out of which three had been manipulated to indicate the opposite position using the Choice Blindness Paradigm. Just over half of the manipulations were corrected by the participants, meaning that the remaining was accepted by the participants as being their own attitudes. This is similar to previous CB studies on political attitudes (Strandberg et al. 2018; Hall et al. 2012).

Attitude strength

In this study we were particularly interested in testing potential underlying factors that predisposes participants to correct the manipulations. We found that correction was mainly predicted by attitude extremity; meaning that the stronger participants agreed with an issue on the bipolar scale, the more likely they were to correct it. That attitude extremity correlates with correction is also in line with previous CB research (Strandberg et al. 2018; Hall et al. 2012; 2013) and corresponds with for example Bassili's (1996) findings on the relationship between extremity and attitude stability.

However, surprisingly, the two meta-attitudes centrality and commitment did not contribute to the correction prediction. One possible explanation to this could be that operative measures of attitude strength, such as extremity, are more relevant to the task compared to second order impressions such as centrality and commitment. Bassili (1996) suggested that extremity is closely associated with the cognitive processing involved in attitude formation and retrieval which is two main components in a CB task. Centrality and commitment on the other hand rather tap into more abstract concepts of the attitude structure (Holland, 2003) not necessarily relevant for scrutinizing one's own survey responses. It could also be that higher extremity is the product of deeper and more involved elaboration, making those responses more salient and memorable (Petty & Cacioppo, 1986).

These results highlight the difficulties in assuming an attitude's strength and stability based on seemingly relevant self-report measures.

Individual difference and cognitive style

The two measures of cognitive style, preference for consistency and need for cognition, were also not able to predict correction.

Preference for consistency. In the case of PFC (Cialdini, Trost & Newsom, 1995), we interpret this as an indicator that the correction of CB manipulations is not based on consistency motives or social influence. Further, PFC is mainly about people self-monitoring and being aware about their own consistency; whereas CB corrections tend to occur outside of the participants' awareness. This could be seen in the reasons people reported when wanting to correct: they were almost exclusively about having made a mistake, detected a glitch in the survey application, or having spontaneously changed their minds. Importantly this result

also distinguishes CB from cognitive dissonance (Festinger, 1957) and other consistency phenomena that are typically highly correlated with PFC. This is useful when discussing CB and its consequences in a larger theoretical context.

Need for cognition. NC (Cacioppo, Petty & Kao, 1984) is often used in social psychology research for its supposed implications to people's attitudes, judgments and decisions. In this literature, NC is described as associated to peoples' tendency to process information and form elaborated and coherent attitudes. Because of this, attitudes of individuals high in NC should be more resilient to change, persuasion, and context effects (e.g. Haugtvedt & Petty, 1992). This is not what we found in this study. However, while individuals high in NC tend to be more resistant to various biases, previous research argue that even these individuals can be influenced if the bias is very subtle (Cacioppo, Petty, Feinstein & Jarvis, 1996). The subtlety factor might help explain why NC and CB correction did not correlate. Further, people with low NC can perform at a comparable level to those with high NC given enough external motivators. One such motivator could be the perception of what participants believe to be their own survey response.

Political awareness. The two political awareness measures (political interest and involvement) also did not correlate with correction. While there is nothing uniquely special to *political* awareness per se, the awareness part addresses a domain specific aspect that could determine the participants' understanding, knowledge, and vested interest about the current CB theme (Zaller, 1992). For example, one previous CB study did find that political involvement correlated with correction (Hall, Johansson & Strandberg, 2012), and in this study we found a tendency (albeit small) that politically involved participants were more likely to attribute the correction externally (e.g. believing that there was some error with the equipment). This tendency at least indicates that politically involved participants experienced the false feedback differently from the uninvolved. It could simply be that politically involved individuals have stronger convictions in the politically attitudes; so when they notice a discrepancy between their original and present response, their main explanation is that software application malfunctioned.

Limitations and future studies

The main limitation of this study was the small number of participants. While we only found a relationship between correction and attitude extremity, the lack of relationship between the other variables might at least be partially explained by the small sample size. Thus, one interesting avenue of future research would be to more systematically, and with more participants, test how a variety of attitude, personality, and performance measures affect correction rates and correction types. This would also allow us to examine subgroups within our sample; for example: what is it that makes some participants correct all the manipulations and some accept all? Importantly, while we found no relationship between correction and any of the two motivated cognition measures (NC and PFC), other more performance based variables might be relevant to CB and worth exploring. For example, in Strandberg et al. (2018) the Cognitive Reflection Test (CRT) correlated with correction, with participants having higher CRT score also being more likely to correct the manipulations. CRT is a performance based cognitive processing measure that captures peoples' ability to use reflective and deliberative thinking instead of gut feelings (Frederick, 2005). Thus, future research could try to link CB to performance based measures that taps into working memory, attention, or perhaps factual knowledge.

Another potential shortcoming of this study was that the majority of the participants were students. Although we have no reason to believe, given previous studies, that a phenomenon such as CB would drastically differ between different demographics, it is always important to establish whether the experimental findings generalize across the public. However, similar levels of correction have been found in experiments with a more diverse and representative sample (Strandberg, Olsson, Hall, Woods & Johansson, in preparation).

Conclusion

Choice blindness is a cognitive phenomenon powerful enough to influence peoples' opinions and reasoning in important political issues. Still, it is difficult to pinpoint what disposes people to accept or correct the manipulations. It seems that the CB manipulation is so surreptitious that it sometimes flies under the radar even for people with strong convictions and motivations to engage in political

reasoning. This study contributes to the understanding of CB, serving as both a backdrop for future research, and an important piece of a broader theoretical puzzle.

Acknowledgments

We would like to thank Anders Lindén at AndVision for implementing the experimental design to a tablet interface, Henrik Ekengren Oscarsson at the University of Gothenburg for helping to select the political issues, and Uno Otterstedts Fund (RFh2014-0390) for financing the movie tickets.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bator, R. J., & Cialdini, R. B. (2006). The nature of consistency motivation: Consistency, aconsistency, and anticonsistency in a dissonance paradigm. *Social Influence, 1*, 208–233.

Bassili, J. N. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of Personality and Social Psychology, 71*(4), 637-653.

Bürkner, P. C. (2016). brms: Bayesian regression models using Stan. *R package version 0.10. 0*.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197-253.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology, 69*(2), 318-328.

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2018). (Choice) blind justice: Legal implications of the choice blindness phenomenon. *University of California, Irvine Law Review 8*, 85.

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010) Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117*, 54–61.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS One, 7*(9), e45457.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS One, 8*(4), e60554.

Haugtvedt, C. P., & Petty, R. E. (1992). Personality and persuasion: need for cognition moderates the persistence and resistance of attitude change. *Journal of Personality and Social Psychology, 63*(2), 308-319.

de Holanda, G. L., P., P. H., & Wolf, L. J. (2018). The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment*.

Holland, R. W. (2003). *On the structure and consequences of attitude strength.* Dissertation Thesis, Radboud University, Nijmegen.

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.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making, 27*(3), 281–289.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4*. Attitude strength: Antecedents and consequences (pp. 1–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lusk, C. M., & Judd, C. M. (1988). Political expertise and the structural mediators of candidate evaluations. *Journal of Experimental Social Psychology, 24*(2), 105-126.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205.

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69*(3), 408–419.

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Sidanius, J. (1988). Political sophistication and political deviance: A structural equation examination of context theory. *Journal of Personality and Social Psychology, 55*(1), 37-51.

Strandberg, T., Sivén, D., Hall., L., Johansson, P., & Pärnamets, P. (2018) False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General, 147*(9), 1382-1399.

Strandberg, T., Olson, J.A., Hall, L., Woods, A.T., Johansson, P. (manuscript in preparation).

Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York: Cambridge University Press.

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

The malleability of political attitudes

This thesis is an empirical and theoretical investigation of choice blindness, in particular in the domain of political attitudes. Choice blindness is a cognitive phenomenon in which people do not notice dramatic mismatches between what they choose and what they get while still offering seemingly introspective arguments to explain their (putative) choice. In four papers, it is demonstrated that the effect also applies to salient political attitudes and evaluations of political candidates. All studies took place in close connection to real elections, and new tools building of the underlying choice blindness methodology has been developed to collect the data. Further, the potential downstream effects are explored, such as influence on voting intentions, and lasting attitude changes. Some mechanisms behind the effect are also investigated and confabulatory reasoning stands out as an important part in facilitating the observed attitude changes.

LUND UNIVERSITY

Faculties of Humanities and Theology
Department of Philosophy
Cognitive Science

Lund University Cognitive Studies 179
ISBN 978-91-89213-06-7
ISSN 1101-8453