Bayesian analysis of the role of metacognition in cognitive control

Bence Palfi

Thesis submitted for the degree of Doctor of Philosophy

School of Psychology

University of Sussex

November 2019

TABLE OF CONTENTS

UNIVERSITY OF SUSSEX

BENCE PALFI, PHD IN PSYCHOLOGY

BAYESIAN ANALYSIS OF THE ROLE OF METACOGNITION IN COGNITIVE CONTROL

SUMMARY

Metacognitive theories of hypnotic responding, such as the cold-control theory, assert that people engage in strategies to create the experience requested by a suggestion (e.g., percept of something that is not present). This act is accompanied with the feeling of involuntariness due to a disrupted metacognition that makes suggestible people (highs) unaware of their intention to be engaged in the strategy. The presented research investigated predictions drawn from the cold-control theory focusing on the claim that hypnotic suggestions should not provide highs with any special abilities, since hypnotic responses are implemented by traditional cognitive control processes. In several experiments, the word-blindness suggestion (suggestion to see meaningless words in the Stroop task) was applied. The suggestion halves the extent of the Stroop interference when it is given to highs, posing a challenge to the metacognitive account of hypnosis and so providing a unique opportunity to test its assumptions. In addition, as the experimental usage of hypnosis requires offline screening to identify potential participants, I investigated whether it is possible to conduct the screening online to ease the costs of recruitment. Throughout this thesis, the Bayes factor was applied for hypothesis-testing. In the last chapter, a case with a 2x2 design (a design that was frequently used in this thesis) is presented to demonstrate how Bayes factors with cut-offs of good enough evidence relate to the old inferential mistake of the neglect of the test of interaction.

This thesis presents empirical evidence that responses to the word-blindness suggestion are not produced by the disruption of reading. First, the extent of the effect is influenced by the proportion of incongruent Stroop trials of the experimental blocks implying that the suggestion alleviates response competition whereas it does not de-automatise reading. Second, when highs are asked to use simple visual strategies, such as blurring and looking-away, to reduce the Stroop interference, the pattern of results produced by the strategies are not in harmony with those of the word-blindness suggestion deeming it unlikely that highs use these strategies when they respond to the suggestion. This thesis also examines whether a voluntary act to imagine meaningless words is sufficient to reduce the Stroop interference. Consistent with the core idea of cold control theory, we found a positive correlation between the extent to which highs reduced interference in the volitional request and in the suggestion conditions. Nonetheless, the current strength of evidence is not good enough to conclude whether or not voluntary and hypnotic responses can reduce the interference to the same extent. The experiment comparing offline and online hypnotic screening demonstrated evidence supporting the notion that hypnotic suggestibility (measured via the SWASH) can be assessed online. Finally, different scenarios of a case study were presented to help researchers develop the right intuition on the issue of why Bayesian evidence for H1 in one group and Bayesian evidence for H0 in another group does not mean Bayesian evidence for the difference between the two by itself.

Declarations

The thesis conforms to an 'article format' in which the middle chapters consist of discrete articles written in a style that is appropriate for publication in peer-reviewed journals in the field. The first and final chapters present overviews and discussions of the field and the research undertaken.

Chapter 2 is written in the style of an article appropriate for *Cortex*.

B. Palfi, B. A. Parris, A. Seth and Z. Dienes contributed to the study concept and design. B. Palfi performed data collection; B. Palfi performed the data analysis and interpretation under the supervision of Z. Dienes. B. Palfi drafted the manuscript, and all authors provided critical revisions. The preregistration and the materials of the study can be accessed at https://osf.io/7ma4t/.

Chapter 3 is written in the style of an article (Report format) appropriate for *Journal of Experimental Psychology: Human Perception and Performance*.

B. Palfi, B. A. Parris, and Z. Dienes contributed to the study concept and design. B. Palfi performed data collection with the help of A. Collins (who used the data as part of his Msc thesis). B. Palfi performed the data analysis and interpretation under the supervision of Z. Dienes. B. Palfi drafted the manuscript, and all authors provided critical revisions. The preregistration and the materials of the study can be accessed at https://osf.io/bkweg/ and the preprint of the Manuscript submitted to *JEP:HPP* at https://psyarxiv.com/ej7w8/.

Parts of Chapter 4 were accepted as a Stage 1 Registered Report at *Cortex*:

Palfi, B., Parris, B. A., McLatchie, N., Kekecs, Z., & Dienes, Z. (2018). Can unconscious intentions be more effective than conscious intentions? Test of the role of metacognition in hypnotic response. *Cortex*, (Stage 1 Registered Report).

All the authors contributed to the study concept and design. The data of the Pilot study were collected by the students of Z. Dienes as part of student projects in 2013 and 2014. B. Palfi performed data collection of the Pre-registered Experiment with the help of N.

Mclatchie. B. Palfi performed the data analysis (both of the Pilot and Pre-registered Experiments) and interpretation under the supervision of Z. Dienes. B. Palfi drafted the manuscript, and all authors provided critical revisions. The accepted Stage 1 version of the manuscript and the materials can be accessed at https://osf.io/h6znt and at https://osf.io/d67u8/, respectively.

Chapter 5 is published in *Psychological Research* as:

Palfi, B., Moga, G., Lush, P., Scott, R. B., & Dienes, Z. (2019). Can hypnotic suggestibility be measured online?. *Psychological Research*, 1-12.

All the authors contributed to the study concept and design. B. Palfi, and G. Moga performed the testing and data collection and B. Palfi performed the data analysis and interpretation under the supervision of Z. Dienes. B. Palfi drafted the manuscript, and Z. Dienes provided critical revisions. All authors approved the final version before submission.

Chapter 6 was submitted as a revised version of an earlier manuscript to *Advances in Methods and Practices in Psychological Science*.

B. Palfi performed the data analysis and wrote the script of the Shiny app with the supervision of Z. Dienes. B. Palfi drafted the manuscript and Z. Dienes provided critical revisions. The materials and the preprint of the manuscript submitted to *AMPPS* can be accessed at https://osf.io/xrctq/ and at https://psyarxiv.com/qjrg4/, respectively.

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature: .....

Acknowledgments

First, I would like to thank my family for their support. Even if living all around Europe makes it difficult to see each other, I felt very much supported and loved by you. Thanks to Pete for helping me get my head around the who is who of hypnosis research and for sharing practical advice on how to minimise the nuisance involved with administrative duties and the recruitment of participants. Thanks to Lina for all the fun and useful conversations, for listening to my rants on the credibility crisis in psychology, and for correcting me when I misunderstood a theory. I`m very greatful to Balazs Aczel for encouraging me to pursue an academic carreer and for supporting me all the way during my PhD. I would like to express my thanks to Anil Seth for his advice and for the opportunity to join the Sackler Centre, a scientifc community, which is full of critical yet friendly people. Finally but most importantly, I would like to thank Zoltan Dienes for being an amazing supervisor, for guiding me in all ascpects of academia from how to ask the right questions to how to handle Reviewer 2. I will truly miss the Wednesday morning meetings.

Chapter I: Introduction

General Introduction

> "In the summer of 1974, a 26-year-old Mayan villager lay drunk in a town square in the Guatemalan highlands. Suddenly he heard a voice that was to change the course of his life and that of his home town, Almolonga. "I was lying there and I saw Jesus saying, 'I love you and I want you to serve me'," says the man, Mariano Riscajche. He dusted himself down, sobered up and soon started preaching, establishing a small Protestant congregation in a room not far from the town's ancient Catholic church." (The Stand, 2017).

This story comes across as an unusual example of an auditory hallucination as it appears that the experience furthered the goals of the person who heard voices. In fact, this is not an isolated case. Hallucination like experiences that are socially appropriate and beneficial for the people sensing them have always been part of our lives. Throughout the history of humanity, these experiences appeared in various rituals. Some of these rituals are religiously motivated, such as the ones inducing spirit possession (Lewis, 2003), whereas others focus more on individual thriving such as mesmerism or as it is called today, hypnosis (Khilstrom, 2008). The fact that these experiences are goal-directed and appropriate to the social context in which they were sensed suggests that they are created by controlled rather than genuinely automatic cognitive processes. However, if these experiences are created by intentional cognitive control processes then why are they felt as if they just happen by themselves? How come the actors have no sense of voluntariness over their own responses?

This thesis focuses on the cold control theory of hypnosis (Dienes & Perner, 2007), which provides a simple mechanistic model explaining how an intentionally formed action can feel involuntary. The theory draws from the seminal consciousness theory of higher-order thoughts (Rosenthal, 1986; Lau & Rosenthal, 2011a) and applies it to intentional actions. In a nutshell, the cold control theory postulates that people have the ability to relinquish the higher-order thought of their intentions, which would create the conscious experience of volition, without affecting their first-order intentions, the intentions that are needed to perform a response. Therefore, according to the theory, hypnosis is about self-deception, and it is purely a metacognitive phenomenon. Although the theory was inspired by hypnosis, its reach goes far beyond that and its explanation can be applied to the phenomena mentioned earlier.

The primary aim of the thesis is to investigate a prediction of cold control theory that follows from the assumption about hypnosis being a metacognitive phenomenon. Cold control asserts that hypnotic responses cannot be more efficient in terms of implementing goal-directed behaviour than non-hypnotic responses, as the mere difference between the two is whether they are accompanied by the feeling of involuntariness or voluntariness. This assumption coincides with the idea that our conscious experiences (at least the conscious experience of voluntariness) and metacognitive processes creating them have limited functions. The thesis tests this hypothesis by investigating a specific hypnotic suggestion to which highly suggestible people can create responses that seemingly disconfirm the prediction of cold control (i.e., hypnosis seems to provide them with skills they do not possess outside of hypnosis). Understanding this suggestion is crucial to explore the boundaries of cold control theory and theories of cognitive control. The secondary aim of the thesis is to contribute to methodological and statistical advancements that can potentially facilitate hypnosis research and increase its credibility. An essential part of any hypnosis research is the screening of suitable participants, which is a time consuming procedure. This thesis advocates the usage of online hypnosis screening over offline screening to foster the conduction of large-scale hypnosis studies. In addition, as this thesis used 2x2 designs in several chapters and in many cases Bayesian evidence for H1 in one condition and Bayesian evidence for H0 in the other condition was expected, the last chapter focuses on explaining why the test of the interaction is essential in these cases to draw any conclusion.

Cognitive control

Contemporary (neuro)cognitive research uses cognitive control as an umbrella term for cognitive functions that allow us to implement and maintain goal-directed behavior even in the presence of distractors (Gratton, Cooper, Fabiani, Carter, & Karayanidis, 2018). According to many models of cognitive control, consciousness is intimately related to control functions and intentional behaviour (e.g., Norman & Shallice, 1986; Jacoby, 1991). This relation is well demonstrated in the influential and simple model of cognitive control of Norman and Shallice (1986; cf. Baddeley & Della Sala, 1996; Cohen, Dunbar, & McClelland, 1990), which postulates a central control system rather than independent control functions. In their model, actions are directly activated by action schematas, which

are regulated by a hierarchical system of cognitive structures. The default and lower-level system called contention scheduling. Contention scheduling handles and implements action schematas once they are activated by external and internal cues. Action schematas can be compatible and incompatible with one another: in the former case, they facilitate each other, whereas in the latter case they inhibit one another. Contention scheduling is essentially responsible for automatic, habitual responses and it is mostly driven externally. The higher-level part of the system, the supervisory attentional system (SAS), monitors and controls the operation of contention scheduling and the action schematas. For instance, the SAS is capable of biasing the activation of certain schematas by either facilitating or inhibiting them in order to achieve a goal-directed response. The SAS operates via conscious control, and its functioning demands the creation of intentions. The SAS is needed to create flexible responses, for instance, to override habitual responses when they cease to be useful, and to plan and form novel actions. Finally, SAS requires attentional resources and it is not capacity-free like contention scheduling.

Later, the link between intentional acts and consciousness was redefined as being putative that needs to be tested and backed by empirical evidence (Jack & Shallice, 2001). The vast majority of research investigating this putative link focused on the demonstration that unconscious information can activate the formation of intentional control processes (e.g., Dehaene et al., 1998; Dehaene, Lau, & Kouider, 2017; Kunde, Reuss & Kiesel, 2012; Lau & Passingham, 2007; Marcel, 1983; van Gaal, De Lange, & Cohen, 2012; contrast Crick & Koch, 1998; Van Gulick, 1994). However, theories of cognitive control (e.g., Norman & Shallice, 1986) motivate research beyond the relationship of unconscious information and control such as questions aiming to unravel the relation of consciousness and intention, which lies at the core of volition and voluntary acts. This thesis presents a framework in which consciousness can be systematically decoupled from intentions without directly influencing the operation of the latter. First, however, I shall discuss the operationalisation of cognitive control.

Measuring cognitive control: the Stroop task

Cognitive control can be probed in interference paradigms where participants are exposed to task-relevant and task-irrelevant stimuli (or stimuli features) at the same time. The Stroop task (Stroop, 1935; for a review see MacLeod, 1991) is one of many interference tasks, and is considered as the gold standard of attentional measures (MacLeod, 1992). In

the Stroop task, the participants are asked to name the color of the presented word while they disregard the meaning of it. On congruent trials, the meaning and the color of the words are in accordance (e.g., RED displayed in red), whereas on incongruent trials, the meaning and the color of the words mismatch (e.g., RED displayed in *blue*). Generally, neutral trials in which the meaning of the word is not related to colors (e.g., SHIP) or in which the presented stimuli are meaningless characters (e.g., xxxx) are used as well in the task. Performance on the task is measured by the accuracy and RTs of the responses. Due to the conflict between the meaning and the color of the incongruent words, people produce the slowest RTs on incongruent trials. The classic Stroop effect is indicated by the RT difference between the incongruent and congruent trials, whereas the Stroop interference effect is the RT difference between incongruent and neutral trials.

The Stroop task is considered to be an index of the conflict between automatic and intentional processes. The Stroop interference effect stems from the fact that the participants read the words and process the meaning of them despite the fact that reading is not part of the task requirements (cf., Tzelgov, 1997, Posner & Snyder, 1975). Word reading qualifies as an automatic process as it is independent of attentional resources and it cannot be brought under control (Brown, Gore, & Carr, 2002; Neeley & Kahan, 2001; Moors, De Houwer, 2006). Nonetheless, Stroop interference is a multicausal phenomenon (e.g., Stirling, 1979): the effect is manufactured partly by the conflict between reading and color processing (e.g., Hock & Egeth, 1970), and partly by the conflict emerging between different response tendencies (e.g., Morton, 1969). In other words, the conflict at the stage of information processing (more precisely semantic conflict) cannot account for the full effect of Stroop interference. While semantic conflict meets the strict criteria of automaticity, the response conflict component of Stroop interference is not completely resistant to controlled processes (Augustinova & Ferrand, 2014). For instance, the presence of humans (Augustinova & Ferrand, 2012b), and substantial training (MacLeod, 1991) decrease the extent of response conflict. These findings challenge the traditional and dichotomous view of automaticity regarding the response conflict component of the Stroop effect. A computational model of the Stroop task focusing on the response conflict component addressed this issue by re-defining automaticity and proposing that automaticity can be graded and that its features depend on the strength of processing pathways (Cohen, Dunbar, & McClelland, 1990; cf. Logan, 1985; MacLeod & Dunbar, 1988). This model presumes that the Stroop interference effect emerges due to the conflict

between automatic and controlled processes while the size of the effect relies upon the involvement of intentional control.

Understanding consciousness as a metacognitive experience

Metacognition was first defined as cognition about cognition (Flavel, 1979). It is-a mental process with which one can monitor (and control) one's own mental processes (e.g., Koriat, Ma`yan, & Nussinson, 2006; Nelson & Narens, 1990). Some theories of metacognition deem metacognitive processes as inherently conscious (e.g., Koriat, Ma`yan, & Nussinson, 2006), whereas other theories assert that many metacognitive processes are unconscious (e.g., Timmermans, Schilbach, Pasquali & Cleeremans, 2012). A popular group of theories that defines conscious experience as being an inherently metacognitive process is the higher-order thought theories of consciousness (Rosenthal, 1986; Lau & Rosenthal, 2011a). There are several variants of the theory but they share the core idea that conscious experience can be altered by a simple change in higher-order representations even if first-order representations remain untouched. A higher-order thought (HOT) is a mental representation that refers to a first-order mental state. For instance, imagine having a first-order state about the world (e.g., "there is a tree"). According to the theory, one has no conscious experience of the world until they possess a HOT asserting one has the first order state (e.g., "I see that there is a tree"). According to HOT theories, it is the HOT that determines the conscious experience of the world and so the conscious experience can be altered by influencing HOTs. Importantly, conscious experience created by a HOT is not the same as self-consciousness or introspection (Rosenthal, 2000; Lau & Rosenthal, 2011b). To be introspective, one needs a third-order state (e.g., "I am aware that I see a tree") making one aware of the conscious first-order state (i.e., of the second order state, which renders the first order state conscious). A second-order state without a third-order state referring to it is unconscious, because one is not aware of that second-order state. `

HOTs are not restricted to perception, they can refer to any mental state, including those with control functions (i.e., intentions). Hence, HOT theories assert that intentional control processes can exist in an unconscious form as well (cf., Jack & Shallice, 2002; contrast Norman & Shallice, 1986). A special case of the HOT theories, the cross-order integration theory (COI; Kriegel, 2007) assumes that first-order states and HOTs are not independent representations. They can bind together to a unified conscious representation

and by that casually influence one another. However, the mainstream theories of HOT assume that first-order states and HOTs are independent representations and that HOTs have limited or even zero function regarding affecting first-order states (Lau & Rosenthal, 2011; Rosenthal, 2008). This assumption is particularly interesting for intentions as it implies that if a HOT of an intention can be replaced without altering the first-order intenion, then control functions can be just as efficient when they are experienced as voluntary and involuntary.

Hypnosis

Richard Feynman underwent hypnosis once during his time at Princeton. He entered the situation with a sceptical mind, suspicious that he will not experience anything special. However, he was labelled as a good subject based on a pre-screening procedure and was selected to participate in the live hypnosis demonstration, where he responded to all kinds of hypnotic suggestions. Feynman summarised his experiences as follows: “So I found hypnosis to be a very interesting experience. All the time you’re saying to yourself, “I could do that, but I won’t”—which is just another way of saying that you can’t.” (Feynman & Leighton, 1992, p. 58). This summary highlights the two distinctive features of hypnosis: 1) responses to hypnotic suggestions feel involuntary as if they happen by themselves, 2) hypnotic responses are goal-directed and the control lies with the hypnotic subject. These features might strike us as being incompatible, but many theories of hypnosis managed to explain both of them within a single model.

Before turning towards the theories of hypnosis let us review the empirical evidence supporting the notion that hypnotic responses are not simply faked by the subjects and that the subjects genuinely experience involuntariness while being in control of their responses. The feeling of involuntariness, which accompanies hypnotic suggestions, is also called the classic suggestion effect due to being intimately tied to hypnosis and hypnotic responses (Weitzenhoffer, 1974, 1980). Our knowledge of the subjects` influenced behaviour and altered sense of reality comes from observations of overt behaviour and administration of subjective reports of experiences. However, subjective reports and overt behaviour are simple to fake and one should assume that it is done, for instance, by participants who are motivated to fulfil the wishes of the experimenter. To test whether participants responses are fabricated to satisfy demand characteristics, Orne (1959) created the real-simulator design in which low suggestible

people are asked to go along with the suggestions as if they would be highly suggestible. Interestingly, highs respond to suggestions even when they believe that they are not observed anymore whereas simulators stop responding to suggestions once they assume that they are not observed (Kirsch, Silva, Carone, Johnston, & Simon, 1989; Orne, Sheehan, & Evans, 1968; Perugini et al., 1998). Furthermore, simulators and hypnotised people produce different electrodermal skin conductance responses when they report their subjective experiences of suggestions. Specifically, by using a criterion of truthfulness on the skin conductance responses, a study found that 89% of the hypnotised participants` and 35% of the simulators subjective reports can be identified as truthful indicating that hypnotised people do not simply lie about their experiences as simulators (Kinnunen, Zamansky, & Block, 1994). The credibility of subjective reports is also corroborated by the finding that there is a correlation between the "pain matrix" (Derbyshire, 2000) and the subjective reports of fibromyalgia[1] patients on the successfulness of a hypnotic suggestion in reducing pain (Derbyshire, Whalley, & Oakley, 2009).

The second defining characteristic of hypnotic responding is that it is goal-directed (Coe, 1978; Spanos, 1986; Spanos & Barber, 1974). In other words, hypnotic subjects cannot be suggested to act against their own will or interest, and their responses are always appropriate to the context, more precisely, to the subjects' understanding of the context. For instance, a hypnotic suggestion per se is not sufficient to induce anti-social behaviour unless the subjects believe that it is appropriate in the current experimental context (Weitzenhoffer, 1949). In addition, being motivated to participate in a hypnotic session and having positive attitudes about it correlate with the successfulness of hypnotic responses and with the strength of experiences (Spanos & Barber, 1974).

Theories of hypnosis can be assigned into three categories: 1) theories not involving cognitive control processes, 2) special process theories of hypnosis[2], 3) metacognitive theories of hypnosis. A prominent theory of the first category is the response expectancy theory (Kirsch, 1985), which provides a unifying explanation for hypnotic responses and responses to placebos. The theory presumes a simple mechanism:

[1] A long-term neuropsychological condition inducing pain all over the body

[2] This term usually reflects that a theory asserts a special hypnotic state, which facilitates responses to suggestions, however, here it is used in reference to theories assuming cognitive processes that are unique to hypnosis and hypnotic responding.

responses to suggestions are produced by expectations and beliefs. According to the response expectancy theory, hypnotic procedures are useful tools to influence the expectations of the participants, which in turn create the experiences of an altered reality. These experiences are sensed to be involuntary as they are not produced by intentional cognitive control processes, they are directly created by the expectations themselves, hence, they are truly unintentional. Nonetheless, measured expectations cannot fully account for the variance in hypnotic responding (Benham, Woody, Wilson, & Nash, 2006; Shor, Pistole, Easton, & Kihlstrom, 1984). These results may be due to measure unreliability but they also pave the way to alternative, cognitive ability based accounts of theories of hypnotic responding. Cognitive ability based theories are not mutually exclusive with the response expectancy theory, there might be multiple factors at play producing responses to suggestions.

Special process theories of hypnosis presume an innate mechanism that is specifically related to hypnosis and hypnotic responding (though innate, its operation in any given person may have environmental influences; Hilgard, 1970). The first modern theory of its kind is the neodissociation theory (Hilgard, 1977) that revitalized the idea of dissociation and was part of a more general theory of cognitive control. Hilgard`s model of cognitive control strongly resembles the one proposed by Norman and Shallice (1986) in that they both assume that automatic action tendencies compete with one another to take control over behavior, while a central, intention driven structure (the 'executive ego') can override these actions. According to neodissociation theory, hypnosis targets the executive ego and splits it into a conscious and (from the perspective of the first part) unconscious part by creating an amnesic barrier between the two. Behaviour can be initiated by any part, but monitoring is operated by the unconscious one, hence, responses to suggestions are intentionally initiated and goal-directed but experienced as involuntary. The dissociated control theory (Bowers, 1992; Woody & Bowers, 1994) asserts that due to hypnosis one can dissociate controlled and automatic processes allowing external cues (i.e., the suggestions) to directly trigger automatic behaviours. In terms of the model of Norman and Shallice (1986), hypnosis targets the SAS and weakens its control over CS, hence, the suggestion can take control over habitual responses, which will be sensed as involuntary due to being genuinely unintentional. However, the idea that hypnosis dampens controlled processes and externally triggers automatic processes is challenged by the ample evidence supporting that hypnotic responding is created top-down (Terhune,

Cleeremans, Raz & Lynn, 2017). For instance, hypnosis does not deteriorate performance on the Stroop task as a model of general cognitive dampening would suggest (Jamieson & Woody, 2007). In addition, the phenomenon of self-hypnosis cannot be explained by a model assuming that hypnotic responding is solely driven by external cues (Kirsch & Lynn, 1998). The integrative model of dissociation theories (Woody & Sadler, 2008) asserts that dissociation due to hypnosis can happen between various cognitive processes: between cognitive control and subsystems of control (dissociated control), and between cognitive control and monitoring (dissociated experience). This latter mechanism assumes that hypnotic responding is controlled but it feels involuntary as monitoring is decoupled from the controlled processes. This mechanism resembles the one assumed by metacognitive theories of hypnosis, however, the metacognitive theories typically regard hypnosis as a non-special process.

Theories from the sociocognitive tradition explain hypnotic phenomena as a metacognitive process in which hypnotic responses are delivered by general cognitive control processes. They stress the role of demand characteristics in forming the subjective experiences involved with hypnotic responding and describe hypnosis as "role playing" that is strategic and appropriate to the specific context (Comey & Kirsch, 1999; Sarbin & Coe, 1972; Spanos, 1986). According to the sociocognitive theories, motivation is fundamental in creating a response. Hence, subjects need to be convinced that behaving in line with the suggestion is in accordance with their goals. These theories assert that subjects volitionally engage in strategies to create the requested experiences (e.g., they divert their attention to reduce pain) while they manage to deceive themselves so that their actions feel involuntary. Nonetheless, these theories do not provide a clear cognitive model elaborating how one can make themselves believe that their action is not initiated internally. The neodissociation theory (Hilgard, 1977) offers a cognitive explanation for this by introducing the amnesic barrier. However, this model presumes the existence of a special, hypnosis-related mechanism as well as the division of the control structure, which results in simultaneous cognitive processing (Kihlstrom, 1997). These might be overcomplicated assumptions.

A simple integration of the different theories is proposed by the cold control theory (Barnier, Dienes & Mitchell, 2008; Dienes, 2012; Dienes & Perner, 2007) as a synthesis of the sociocognitive and neodissociation theories. This theory draws from the higher-order thought theories of consciousness (Lau & Rosenthal, 2011; Rosenthal, 2005)

and offers a purely metacognitive approach to hypnotic responding that is compatible with the cognitive control model of Norman and Shallice (1986). Cold control theory posits that hypnotic responses are implemented by intentional cognitive control processes (e.g., by the SAS in the model of Norman and Shallice), the subjects produce first-order intentions that in turn create the behavior and experience requested by the suggestion. For instance, to experience the hallucination of a noise (e.g., the buzz of a mosquito), one has to have an intention to imagine that particular noise (e.g., an intention with the content "imagine a buzzing mosquito"). This means that subjects engage in intentional strategies just as it was proposed by both the neodissocaition theory (in that the SAS of the split off part can implement executive control) and especially the sociocognitive theories. This part of the theory explains how subjects play an active role in creating the response and why the hypnotic response is goal-directed. Outside of hypnosis, this response would simply feel as imagination as the first-order intention to imagine the buzzing mosquito would be accompanied by the HOT that one is intending to imagine it. Cold control theory posits that the mechanism by which this response becomes hypnotic (i.e., sensed as involuntary) is a process that allows people to replace the accurate HOTs about their intentions with inaccurate ones (Dienes, 2012). Sticking to the example about the hallucinated noise of a mosquito, one needs to form an inaccurate HOT about the first-order intention (e.g. with the content, "I'm not intending to imagine a buzzing mosquito") to experience the noise of the mosquito as perceived rather than as imagined. Consequently, cold control theory claims that the sense of involuntariness is an illusion. Hypnotic responses are only experienced as involuntary but they are in fact products of intentional cognitive processes. The reach of cold control theory may go beyond the hypnotic phenomenon. There are various incidents in which people behave in a goal-directed manner while they experience involuntariness regarding their own behaviour. For instance, cold control may be the underlying mechanism of spirit possession or channelling, automatic writing or glossolalia (Dienes & Perner, 2007).

A central assumption of cold control theory is that one needs to be able to misattribute their intention to experience an act as hypnotic (i.e., involuntary). Hence, the theory predicts a negative relationship between accurate metacognitive monitoring (specifically the monitoring of intentions) and successful hypnotic responding. Indeed, individual differences in the tendency to generate inaccurate HOTs of intending are moderately associated with hypnotisability (Dienes, 2012). In addition, the temporary

disruption of brain areas associated with metacognitive functions enhances hypnotic responding (Coltheart et al, 2018; Dienes & Hutton, 2013; Semmens-Wheeler, Dienes & Duka, 2013). Other core assumptions of the theory are that hypnotic responses are created by intentional cognitive control processes and the sole difference between hypnotic and non-hypnotic responding is the nature of the accompanying HOT. An important prediction follows from these assumptions, namely, that subjects can create any response non-hypnotically that they can produce hypnotically.[3] In other words, hypnosis cannot provide subjects with abilities that they do not possess outside of the hypnotic context. This is a controversial claim, many theorists and models of hypnosis argue otherwise (e.g., Brown & Oakley, 2004; Woody & Sadler, 2008). The empirical evidence in line with this prediction is overwhelming (e.g., Barber, 1966; Erdelyi, 1994; Levitt & Brady, 1964; Nogrady, McConkey, & Perry, 1985; Milling, Kirsch, Meunier & Levine, 2002; Spanos, 1986), however, there are some phenomena that might call into question this idea. Here, I only present one of these phenomena, the *word blindness suggestion* effect. This effect poses the greatest challenge for the investigated prediction of the cold control theory. Other challenging phenomena and the corresponding counter arguments are discussed in Chapter IV.

The word blindness suggestion instructs subjects to see the words during the Stroop task as a meaningless foreign script (Raz, Shapiro, Fan, & Posner, 2002). When this suggestion is given to highly (or sometimes moderately) suggestible people, they can halve the Stroop interference effect compared to a control condition in which they are instructed to not try to make the words meaningless (Parris, Dienes, & Hodgson, 2013). This word blindness effect has been replicated by several independent labs (Augustinova & Ferrand, 2012; Casiglia et al., 2010; Parris, Dienes, & Hodgson, 2012; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006) and even with slightly different wording of the suggestion (Zahedi et al., 2017). The effect is particularly interesting as it is created by the reduction of incongruent RTs rather than the increase of neutral RTs making it unlikely that it would be the result of a simple hold back effect (MacLeod, 2011; Spanos, 1986; Zamansky, Scharf & Brightbill, 1964). Initially, it was proposed that this remarkable reduction is achieved by the de-automatisation of reading (Raz et al., 2002, 2006; Raz & Campbell, 2011). This model posits that suggestion allows highs to take control over otherwise

[3] Note that this prediction coincides with the idea that HOTs have limited or even zero capacity to directly influence first-order states (Lau & Rosenthal, 2011; Rosenthal, 2008).

automatic processes, and by influencing the neural processing of input words (i.e., hindering reading), the conflict can be evaded at an early stage of the process. As a competing model, the response competition account was put forward (Augustinova & Ferrand, 2012; Parris, Dienes, & Hodgson, 2013). This model posits that reading is untouched by the suggestion and so semantic as well as response conflict emerges. The reduction in Stroop interference happens due to an enhanced conflict resolution that alleviates response conflict. The debate whether the word blindness suggestion helps subjects evading the conflict or resolving it is not settled yet as there is behavioral and neural evidence supporting both accounts. The relevant studies will be reviewed in more detail in the next Chapter that aims to investigate this matter in a behavioral experiment. The existence of the word blindness effect poses a challenge for the cold control theory and provides a great tool to understand the boundaries of the theory while raising several interesting questions. For instance, is there a voluntary way to reduce Stroop interference while mimicking the effect of the word blindness suggestion? More generally, does metacognition play a direct role in how well one can implement a first-order intention? Are there cases when an unconscious intention is more efficient than a conscious one? These are the main questions that the next three Chapters are striving to answer.

Measuring individual differences in hypnotic suggestibility

A well-established phenomenon within the field of hypnosis research is that there are substantial and reliable individual differences in the responsiveness to (hypnotic) suggestions (Hilgard, 1965; Laurence, Beaulieu-Prévost, & du Chéné, 2008; Piccione, Hilgard, & Zimbardo, 1989). Basically, all contemporary theories of hypnotic responding are consistent with this observation, although they provide different explanations for it. For instance, cold control theory presumes that individual differences in responsiveness to suggestions are mostly shaped by the ability to monitor one's own intentions (Dienes & Perner, 2007); dissociation theories link responsiveness to dissociative mechanisms through which one can weaken executive control or create temporal unawareness of control processes (e.g., Brown & Oakley, 2004); while according to the response expectancy theory, responsiveness is defined by the beliefs and expectations of the individuals rather than by their cognitive skills (Kirsch, 1985).[4]

[4] Note that all theories acknowledge that individual differences in responsiveness stems from multiple factors. For instance, motivation or cognitive skills such as imagination are needed to create the experience

The magnitude of individual differences in responsiveness is so remarkable that some people barely respond to any of the suggestions (lows), whereas others can easily create responses to the most difficult suggestions as well (highs) such as the auditory or visual percepts of things that are not present. Due to this, undertaking experimental hypnosis research requires the pre-screening of people in order to identify highs who can potentially create the suggested experience and lows who can be deemed as controls. The development of scales particularly focused on the improvement of three factors: 1) how to reduce the cost of the administration or how to make the procedure easier, 2) how to align measures with contemporary theories (e.g., Terhune & Cardena, 2016; Woody & Barnier, 2008), 3) how to improve the measurement properties of the scales. One of the first modern standardised scales is the Stanford Hypnotic Susceptibility Scale: Form A (SHSS:A; Weitzenhoffer & Hilgard, 1959). Since the publication of SHSS:A, more than 30 standardised scales have been developed and published in peer-reviewed journals (Acunzo & Terhune, 2019). Some of these are completely unique but the majority of the scales are updated versions of earlier ones. For instance, the SHSS:A quickly went through several modifications and a group version, the Harvard Group Scale of Hypnotic Susceptibility Form A (HGSHS:A; Shor & Orne, 1963), was created a few years after its publication to speed up the identification of highs and lows. The most recently published scale is the Sussex-Waterloo Scale of Hypnotizability (SWASH; Lush, Moga, McLatchie, & Dienes, 2018), which is a modified version of the Waterloo-Stanford Group Scale of Hypnotic Susceptibility (WSGC; Bowers, 1993). The alterations were motivated by multiple reasons. For instance, the shortening of the induction procedure was justified by theoretical as well as practical reasons (i.e., quicker administration). The administration of the SWASH is very simple, it does not even require the experimenter to interact with the participants during the screening, as the script containing the induction and the suggestions are audio-recorded and played to the participants (Lush, Scott, Moga, & Dienes, 2019). Nonetheless, there is still room for improvement. For example, online measurement of hypnotic suggestibility would further reduce recruitment costs and it would facilitate the application of more heterogeneous samples (see Chapter V); or scales

---

of a percept that is not present. However, these theories diverge in how they explain the emergence of the feeling of involuntariness, which is a crucial constitute of a hypnotic responding. Hence, they hold different assumption on what specific cognitive ability or mental state plays the most important role in determining responsiveness.

that measure responsiveness to suggestions with multiple-trials rather than with a single trial would increase the precision of measurement (Acunzo & Terhune, 2019).

Two reasons to use the Bayes factor instead of frequentists statistics to draw conclusions from the results

Null-Hypothesis Significance Testing (NHST) became the orthodox way of how experimental psychologists test their hypotheses and make inferences based on their results (Fisher, 1925; Neyman & Pearson, 1933). NHST operates through the calculation of the famous (or infamous) p-value, which is the probability that we observe a specific effect size or an effect size that is more extreme while assuming that the null effect is true. This tool provides us with a simple rule on when should we reject the null hypothesis (i.e., psychologists usually apply the threshold of .05), and allows us to control the long-term error rates of our decisions. Nonetheless, NHST bore heavy criticism based on multiple grounds, it was challenged from a philosophical point of view (e.g., Dienes, 2011; Wagenmakers, 2007) as well as due to the misuse and misinterpretation of the tool by researchers (e.g., Gigerenzer, 2004; Greenland et al., 2016; John, Loewenstein, & Prelec, 2012). These issues, among several others, played a critical role in the emergence of the current credibility crisis of psychological science (for a detailed discussion of other factors and the potential solutions, see Chambers, 2017). For instance, NHST is an asymmetrical procedure, when one fails to reject H0, no conclusion follows as a non-significant p-value can either indicate evidence for H0 or insensitive evidence. Due to this, many non-significant findings remained unpublished (i.e., publication bias; Sterling, Rosenbaum, & Weinkam, 1995) creating the (false) impression that the vast majority of psychological phenomena are robust (c.f., Open Science Collaboration, 2015). This asymmetric nature of NHST can lead to other problems, such as the misinterpretation of non-significant tests as clear evidence for the null (Aczel et al., 2018; Hoekstra, Finch, Kiers, & Johnson, 2006) or the invalidation of conclusions due to optional stopping (Armitage, McPherson, & Rowe, 1969).

An alternative tool for hypothesis-testing, the Bayes factor (Jeffreys, 1961), has been proposed as a remedy for many of the issues regarding NHST (Dienes, 2011, Wagenmakers et al., 2011). Importantly, using the Bayes factor requires the researchers to ask fundamentally different questions than they would do when using the NHST. In a nutshell, NHST allows us to control the long-term error rates of our decisions (i.e.,

rejecting the null hypothesis), whereas the Bayes factor offers a more intuitive approach to hypothesis testing. It allows us to calculate and continuously update the probability of a theory in light of new evidence by comparing the predictive ability of two competing models (for more information on the anatomy and calculation of the Bayes factor, see Chapter VI). Nonetheless, here, I am arguing from a pragmatic point of view for the application of the Bayes factor and present two reasons why it was more beneficial for the purpose of the current thesis to use it instead of NHST. First, one can discern evidence for the null from insensitive evidence with the Bayes factor. Second, the Bayes factor retains its meaning regardless of the applied stopping rule, allowing researchers to conduct sequential analyses without invalidating their conclusions.

Frankly, inferring the null is possible within the frequentist framework as well, for instance, by applying equivalence-testing (Lakens, McLatchie, Isager, Scheel, & Dienes, 2018; Schuirmann, 1987). This tool requires the researcher to specify the lower equivalence and the upper equivalence that define a null region containing effect sizes that are not meaningful for our alternative hypothesis. The upper boundary of this region is the smallest positive effect size of interest. To conduct the equivalence test, we need to run two one-sided significance tests assessing whether our data are consistent with the smallest effect sizes or not. Assuming that we observed a smaller effect size than the upper bound and our test is significant as well, we can reject the alternative hypothesis predicting positive effect sizes. If we can reject the alternative hypothesis predicting negative effect sizes as well, then we can assert equivalence, or in other words, we are left with the null (region) hypothesis. One might wonder if the assertion of null is possible within the realm of significance testing then why do we need a fundamentally different approach to draw conclusions about the null? A rather mundane reason is that the predictions of current psychological theories are vague and one can hardly ever pinpoint the smallest effect size of interest. It is difficult to imagine a psychological theory that is consistent with a standardized effect size of 0.20 but it is not consistent with 0.10. The Bayes factor evades this problem as it allows us to draw a conclusion about a point null hypothesis (Dienes, 2014).[5] All we need for this is a model of H0 (which can be a point-null or a null region model) and a model of H1 that represents the effect sizes that are in

[5] This remains true as long as the Standard Error (SE) overlaps with a range of negligible effect sizes. For instance, extremely large sample sizes can create so narrow SEs that result in Bayesian evidence for H1 against H0 even if the observed effect size is meaningless.

accordance with our pet theory. The latter should be less challenging than identifying the smallest effect size of interest. For instance, findings of earlier studies or contextual knowledge (features of a measurement tool such as its boundaries) can be applied to define the values that are consistent with a model (for a range of heuristics, see Dienes, 2019). This feature of the Bayes factor was utilized on many occasions in Chapters II-IV of this thesis. For instance, in Chapter IV, the prediction of cold-control theory could be corroborated by demonstrating evidence for the null when response times of two experimental conditions are compared. The possibility to accept the null is also crucial when one is aiming to equalise a confounding factor (such as expectations) in the two conditions, and it is generally important for studies of unconscious mental states as these studies usually aim to demonstrate that the participants were unaware of the presented stimuli or the goal of the experiment (Dienes, 2015).

Another useful feature of the Bayes factor, which stems from the fact that it is assessing evidence for two competing theories, is that its meaning is independent of the stopping rule used during data collection (Dienes, 2016, Rouder, 2014). That is, optional stopping is not a problem for Bayesian statistics - we can do sequential analysis and check the results after every new participant without invalidating the meaning of the Bayes factor. As long as all relevant evidence is taken into account, the Bayes factor remains a valid measure of relative evidence regarding the compared theories. Optional stopping has many pragmatic advantages over fixed designs from which this thesis benefitted to a great extent. Using sequential analysis until we reach good enough evidence one way or another allows us to minimise the number of participants we need to recruit in order to draw conclusions. The recruitment of highly suggestible people comes with a high cost as to find a single potential participant, one needs to screen ten people on average. Provided that we find good enough evidence for H1 or H0 by using optional stopping earlier than by using fixed design (i.e., collecting data until we recruited as many participants as an a-priori sample size estimation suggested), we can spare a substantial amount of time (e.g., we do not need to screen several groups to find a few more highs). In addition, using NHST combined with fixed designs many times requires us to suspend judgment as, for instance, the obtained p-value was just above the pre-set threshold of .05. In these cases, we would need to run a completely new experiment to settle the matter. However, this is not an issue for Bayesians - if we use the Bayes factor, we can continue

data collection until we deplete our subject pool and even then we can ask collaborators to continue the process for us so that we can get good enough evidence for H1 or H0.

Aims of the thesis

The primary aim of this thesis is to investigate the role of metacognitive processes over intentional control functions by testing predictions derived from the cold control theory of hypnosis. The thesis applies hypnosis as an experimental tool that provides an opportunity to examine altered sense of reality and feeling of involuntariness regarding intentional acts in a controlled yet rich context. The most crucial question tested here is whether an unconscious intention can be more efficient than a conscious one. Cold control clearly predicts that hypnosis (i.e., reduced metacognition over control functions) cannot directly influence intentional behaviour, hence, a finding in which unconscious intentions are shown to be more efficient than conscious ones would require the revision of the cold control theory. The secondary aim of the thesis is to contribute to methodological advancements that have the potential to increase the credibility of hypnosis research. Specifically, this thesis advocates the usage of online hypnosis screening, the appropriate way of Bayesian hypothesis testing in 2x2 designs.

Chapter II presents two experiments providing a test of whether there is a negative relationship between the efficiency of the word blindness suggestion (i.e., extent of Stroop interference) and the proportion of incongruent trials of the experimental blocks. Supporting evidence for this phenomenon would imply two things. First, it would disconfirm the notion that word blindness suggestion deteriorates visual input during the Stroop task (de-automatisation of reading account) and it would support the response conflict resolution account of the word blindness suggestion. Second, assuming that hypnotic responding can be deemed as unconscious control (i.e., control without the feeling of voluntariness and effort), the results would corroborate the idea that cognitive conflict, indexed by the proportion of incongruent trials, plays a crucial role in the activation of control processes regardless of the form of the accompanying HOT over the intention. As a secondary interest, the conscious experience of word meaninglessness was measured via subjective reports. It was predicted that highs experience more meaningless words after the suggestion than in the control condition. As an exploratory question, we tested whether the conscious experience of meaninglessness is related to the reduction in the Stroop interference due to the suggestion.

Chapter III raises the question of whether the effect of the word blindness suggestion can be recreated via simple vision, attention or goal-maintenance related strategies. Cold-control theory asserts that changes in first-order states in hypnosis are the product of engagement in strategies. Hence, highs who reduce Stroop interference when they respond to the word blindness suggestion are expected to be engaged in a strategy or in multiple strategies. Three criteria were defined that a strategy must meet to be qualified as a plausible underlying mechanism of the suggestion: 1) reduce Stroop interference, 2) speed up RTs of incongruent trials, 3) affect solely the response conflict component and not the semantic conflict component of the Stroop interference effect. This question was tested in two experiments. In addition, the first experiment tested whether the extent to which a strategy reduces Stroop interference is related to hypnotic suggestibility measured by the SWASH. Cold-control theory assumes that hypnosis is a purely metacognitive phenomenon, hence, it predicts the lack of a relationship.

Chapter IV presents a test of the idea that hypnosis is a purely metacognitive phenomenon. This notion follows from the cold control theory, which states that hypnosis solely affects the HOT of intending and therefore it should not provide abilities to subjects that they do not possess outside of the hypnotic context. A Registered Report, which was informed by a pilot study, compared the performance of highs on the Stroop task when they responded to the word blindness suggestion and when they responded to a volitional request to imagine that the word are meaningless throughout the task. Cold control predicted that highs should reduce Stroop interference to the same extent in suggestion and the volition conditions (after controlling for the effect of expectations), while they experience involuntariness in the first condition and voluntariness in the second one.

Chapter V investigated whether the online assessment of hypnotic suggestibility is feasible and whether the offline procedure could be replaced by an online one so that the amount of time required to identify low and high suggestible people could be substantially reduced. Students, who completed an offline SWASH (Lush et al., 2018) screening as part of one of their modules, were invited to undertake the screening for a second time and they were randomly assigned to either an offline or an online group. To assess whether the online measurement is viable, the raw SWASH scores and psychometric properties of the offline and online measurement were evaluated and compared. It was required from the online measurement that it provides appropriate psychometric properties, that the scores of the online tool are consistent with the scores

of the offline one (i.e., a person scoring high offline should score high online), and that the online measurement does not create floor or ceiling effects. A secondary aim of the study was to estimate the extent to which the delay between the first and the second screening sessions can affect the raw scores or the psychometric properties of the measurement.

Chapter VI aims to demonstrate that the adage that the difference between significant and non-significant is not necessarily significant by itself applies to Bayesian statistics just as it applies to NHST. This thesis used 2x2 designs in all chapters with hypothesis-testing (Chapters II-IV) and in many cases Bayesian evidence for H1 in one condition and Bayesian evidence for H0 in the other condition was expected. However, to properly test the predictions of the investigated theories, the test of the interaction must be conducted as well in these cases. The chapter serves as a tutorial and presents a case study with two scenarios to highlight the importance of the test of interaction in 2x2 designs and to show that Bayesian evidence for H1 in one condition and Bayesian evidence for H0 in another condition does not necessarily indicate Bayesian evidence for the difference of the conditions. As a sideline to the primary goal, the chapter presents a simple rule with which researchers can calculate the “power” of their designs to find good enough Bayesian evidence for H1 or H0.

Chapter II: Does unconscious control depend on conflict?

Introduction

Cognitive control is the hallmark of human cognition as it allows people to achieve their goals, even in constantly changing environments, through processes of error monitoring, conflict resolution and response control. How people know when to intervene and activate control over their behaviour is a central question for cognitive science (Botvinick, Braver, Barch, Carter, & Cohen, 2001). The *conflict-monitoring theory* (Botvinick et al., 2001; Botvinick, Cohen, & Carter, 2004) provides a behavioural and neuropsychological account, according to which cognitive control can be triggered by the conflict of opposing responses, with this monitoring and evaluation process localised to the dorsal anterior cingulate cortex (dACC). In other words, humans are capable of activating and maintaining goal-directed behavior by constantly assessing task-demands in the form of the evaluation of response conflict. The theory has gained support from behavioural phenomena such as the *Congruency Sequence Effect* (CSE) and the *Proportion Congruency Effect* (PCE). These phenomena can be observed in various interference tasks, such as the Stroop task (Stroop, 1935; see MacLeod, 1991, for a review), and their cornerstone is that people tend to adapt to incongruent tasks (trials in which two conflicting responses can be activated by the presented stimulus) owing to preceding conflicts (Botvinick et al., 2001). For example, the congruency effect (incongruent-congruent trial performance) is shown to be decreased on trial *n* when trial *n*-1 is an incongruent trial compared to when it is a congruent trial (CSE; Duthoo, Abrahamse, Braem, Boehler, & Notebaert, 2014; Egner, 2007; Gratton, Coles, & Donchin, 1992). Moreover, the congruency effect is decreased in contexts where the proportion of incongruent trials is higher compared to contexts with lower proportions of incongruent trials (PCE; Bugg, 2012; Gratton, Coles, & Donchin, 1992; Logan & Zbrodoff, 1979; Tzelgov, Henik & Berger, 1992). In both instances, incongruent trials are considered to be the index of conflict that eventually triggers cognitive control processes demonstrated by the enhancement of performance.

Many models of cognitive control deem control processes to be inherently related to consciousness, particularly to the conscious awareness of one's intentions (Jacoby, 1991; Norman & Shallice, 1986). Empirical research called into question this putative link between control and consciousness by showing that control processes can be

activated by unconscious information about conflicting stimuli (Dehaene, Lau, & Kouider, 2017; Kunde, Reuss & Kiesel, 2012; Lau, 2011; van Gaal, De Lange, & Cohen, 2012; van Gaal, Lamme & Ridderinkhof, 2010). Nonetheless, these studies did not challenge the premise that cognitive control is accompanied by the conscious feeling of effort and volition, and argued merely that the processing of conflict does not reach consciousness. Therefore, it is still unexplored whether the association between conflict and control holds when people experience involuntariness, the lack of intentions, regarding their own behavior.

Building on the *higher-order thought* (HOT) theories of consciousness (Lau & Rosenthal, 2011; Rosenthal, 1986, 2002), control processes can be labelled as unconscious, if the person who is acting cannot report on her intentions concerning the behaviour in question. HOT theories differentiate between mental states that reflect the external world (first-order mental states) and mental states about other mental states (second- or even higher HOTs). The theory postulates that one has to have a second order thought about the first order mental state to be aware of the first order state; therefore, having an intention, but lacking a HOT about it makes the intention unconscious. Possessing an inaccurate HOT about the intention can also indicate unconscious control, but it is not a necessary requirement. Here, we refer to involuntary, and so unconscious, control as a behavior derived from a first-order intent to act (e.g., with content "arm rise!") that is either accompanied by a HOT about the lack of that intention (e.g., with content "I do not intend to raise my arm") or is not accompanied by any HOT regarding the intention.

Hypnotic responding can be understood as an instance of unconscious control, as its defining feature is the feeling of involuntariness that accompanies a usually mundane act such as raising one`s arm or hearing the tune of Happy Birthday (Weitzenhoffer, 1974, 1980; Terhune, Cleeremans, Raz & Lynn, 2017). Although theories of hypnosis share the assumption that a hypnotic response feels like it is happening by itself, not all of them concur with the claim that the response is constructed by cognitive control processes. For instance, the *response expectancy theory* claims that expecting a behavior to happen can induce that particular behavior and so suggestions can be implemented without the involvement of intentional executive systems (Kirsch, 1985, 1997). Consequently, the theory assumes that behaviors following hypnotic suggestions feel involuntarily since they lack the intention and are driven by the expectation itself. This theory can partially

explain the effect of suggestions, however, it has been demonstrated that expectations cannot explain much variance in behavior (Benham, Woody, Wilson, & Nash, 2006), which may be due to measure unreliability but also leaves space for other, cognitive ability based models. Several theories, such as the sociocognitive (Comey & Kirsch, 1999; Spanos, 1986) and dissociation theories (Bowers, 1990; Hilgard, 1977, 1991; Kihlstrom, 1985), emphasize that the subjects of hypnotic suggestions play an active role in creating the hypnotic response. By synthesizing these theories, the *cold control* theory offers a solution to how a strategic and intentional behavior can be sensed as involuntary. It posits that the key to responding hypnotically is to alter one`s monitoring of one's own intentions, thus, the feeling of involuntariness arises from possessing inaccurate HOTs about the intentions (e.g., "I am not intending to raise my arm") and not from the lack of intentions (Barnier, Dienes & Mitchell, 2008; Dienes, 2012; Dienes & Perner, 2007). Hence, cold control theory postulates that a hypnotic response should be labelled as unconscious control. Although cold control theory provides a general explanation that suggestions are implemented by the involvement of executive functions, the mechanism by which control processes are activated, when the intentions are unconscious, is still unclear.

The application of the *word blindness* suggestion (Raz, Shapiro, Fan, & Posner, 2002; this term was coined by Parris, Dienes, & Hodgson, 2012) offers a unique opportunity to investigate how unconscious control can be activated. The word blindness suggestion is an instruction for highly suggestible people (henceforth "highs") that the words (i.e., stimuli of the Stroop task) will appear as a meaningless foreign script on the screen. The effect has been replicated by several independent research groups (Augustinova & Ferrand, 2012; Casiglia et al., 2010; Parris, Dienes, & Hodgson, 2012; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006), even with slightly modified wording (Zahedi et al., 2017), and it has been shown to be a reliable tool to halve the Stroop interference effect, as measured by the response time (RT) difference between incongruent and neutral trials (Parris, Dienes, & Hodgson, 2013; See Table 1. for the results of a meta-analysis). Importantly, this reduced interference is not a consequence of participants slowing down on neutral trials in the suggestion conditions compared to the no suggestion conditions but it is the result of speeded responses on incongruent trials, making it unlikely that the enhancement of performance is faked (MacLeod, 2011). These studies attest to the robustness of the word blindness effect, however, the debate on the

underlying mechanisms of the suggestion remained unsettled between two classes of models. Distinguishing between these two competing models has broader implications than a mere understanding of the underlying mechanism of the word blindness effect. This issue is closely related to the main question of the current paper about the role of conflict in the activation of unconscious control.[6]

The de-automatisation of reading model posits that the process of reading is not obligatory and it can be impaired or even hindered in a top-down way, by influencing the neural processing of input words (Raz et al., 2002, 2006; Raz & Campbell, 2011). Specifically, it assumes that the reduction of the Stroop interference effect, when the suggestion is active, is simply the result of a general impairing of the visual stimuli by which subjects cannot employ information about the meaning of the words, thereby, they experience congruent and incongruent trials as identical. Several studies provide evidence consistent with this model. First, it was demonstrated that the word blindness effect arises even when visual accommodation (e.g., visual blurring) and the usage of attentional strategies (e.g., looking away) are prevented (Raz et al., 2003) suggesting that a simple optical explanation cannot account for the phenomenon. Second, by applying the combination of neuroimaging methods (fMRI and EEG), Raz, Fan and Posner (2005) found that the suggestion mitigated the activity of the occipital cortex underscoring the idea that early visual information processing and not the operation of the visual sensory organs can be modulated by the application of the word blindness suggestion. In addition, a recent study using EEG showed evidence for increased theta-band and beta-band activity in the frontal-midline region in response to the activation of the word blindness suggestion, implying the increased involvement of top-down, cognitive control processes (Zahedi et al., 2017). This finding is in accord with the explanation of the de-automatisation model that the suggestion dampens early visual processing by a top-down mechanism. Nevertheless, these data are also in line with other cognitive control based models.

[6] Of note, the current study was not design to test the assumption of cold control theory (and several other theories of hypnosis) that responses to suggestions feel involuntary. We take this assumption for granted based on the abundant evidence from the field of hypnosis (e.g., Kihlstrom, 2008) and relying on another study demonstrating that highs and mediums report decreased control over the meaningfulness of the words in a suggestion condition compared to a no suggestion condition (Palfi, Parris, McLatchie, Kekecs & Dienes, 2018).

The response competition model proposes that the word blindness suggestion takes its effect at a later stage of the process, namely, at the level of response competition resolution (Augustinova & Ferrand, 2012; Parris, Dienes, & Hodgson, 2013). This model assumes that upstream information processing (e.g., semantic processing) remains mainly untouched by the suggestion and so word reading is not de-automatised. Hence, conflicts between word meaning and color can and will arise. According to the response competition driven model of word blindness suggestion effect, the resultant response competition is key in the application of the suggestion. In accord with the prediction that conflict activates conscious cognitive control processes as portrayed in the conflict-monitoring model (Botvinick et al., 2001; Botvinick, Cohen, & Carter, 2004), the response competition account argues that response competition triggers control processes required for the suggestion to take its effect. This is because the presence of response competition leads to an enhanced recruitment of control permitting the word blindness suggestion effect. Without response competition, for example when trials involve only semantic-associative conflict, the suggestion is not triggered, and the extra control permitted by the suggestion does not take effect.

This model is also in accordance with the finding that cognitive control related brain activity increases in the suggestion conditions compared to no suggestion conditions (e.g., Zahedi et al., 2011), as response selection is a primary function of cognitive control (Miller & Cohen, 2001). A handful of behavioral studies have also provided evidence supporting this model. First, Augustinova and Ferrand (2012) compared the effect of the suggestion on a traditional and a semantic version of the Stroop task (Neely & Kahan, 2001). The interference in the later version of the task is entirely the product of semantic factors (incongruent trials consist of words that are highly associated with colors but their font color does not match with it, such as *sky* presented in *green*) and it reduces conflict derived from response competition. Their results revealed that the word blindness suggestion could not reduce the interference effect on the semantic Stroop task, whereas it could mitigate the interference effect on the standard Stroop task, implying that not visual information processing but response competition is affected by the suggestion. Second, Parris, Dienes, and Hodgson (2013) analysed the distribution of the RTs of their earlier study by applying the ex-Gaussian approach[7] and their results were in line with

[7] The ex-Gaussian approach focuses on the distribution of the RTs and one of its parameters (μ, the mean of the Gaussian distribution) is known to be a useful tool for indexing response conflict (Steinhauser &

the view that the enhanced resolution of response competition is the underlying mechanism of the word blindness suggestion.

At root, the two portrayed models of the word blindness suggestion have an opposing stance on how the suggestion itself, more generally unconscious control, is activated. The de-automatisation of reading model proposes that the suggestion can reduce Stroop interference by preventing the conflict from happening: if the meaning of a word is not processed then the conflict between the meaning and the color cannot occur. Consequently, the model assumes that conflict does not play an essential role in the activation of the suggestion. Alternatively, the response competition model suggests that conflict between the meaning and the color still emerges and the increased activation of control processes permits the influence of the suggestion. The latter position implies that the word blindness suggestion effect operates by somehow piggy-backing on existing control mechanisms.

In order to disentangle these two models, we produced an experimental designs with altering levels of conflict from one block to another. Specifically, we manipulated the proportion of incongruent trials in the experimental blocks. By increasing the proportion of incongruent trials one can raise the level of response conflict (e.g. Logan & Zbrodoff, 1979). When the proportion of incongruent trials is high, response competition occurs more frequently increasing the recruitment of control processes; under such conditions, the Stroop effect is reduced. When the proportion of incongruent trials is low, response competition occurs less frequently reducing the activation of control processes; under such conditions the Stroop effect is increased. If the word blindness suggestion effect is the result of top-down control over input words (as per the wording of the suggestion and Raz and colleagues' conclusions) the proportion of incongruent trials should not affect the presence or magnitude of the effect. Alternatively, if the word blindness suggestion effect is dependent on the operation of mechanisms that control response competition, the effect should be larger when the proportion of incongruent trials is high (due to the enhanced activation of control processes).

Hübner, 2009; cf. White, Risko & Besner, 2016). The ex-Gaussian function helps better exploit RT data and reveal effects which would remain hidden for the traditional analysis process (i.e., the exclusion of trials with prolonged RTs and aggregation of RT data in each condition can result in information loss for positively skewed distributions).

To supplement the analyses of the objective measures (RTs), we also assessed the conscious experience of word meaninglessness of the participants, and the expectations of the participants about seeing the words as meaningless characters by employing subjective reports. In a nutshell, we aim to discriminate between two accounts (de-automatisation of reading vs. response competition) of the word blindness suggestion effect. If the response competition model is supported, the experiment addresses a key question raised earlier about the role of consciousness in cognitive control: When intentions are unconscious, what are the mechanisms by which control is triggered?

Methods

Participants

We recruited 23 (21 females, M = 19.13, SD = 0.81) highs from the subject pool of the University of Sussex to participate in our experiment. The hypnotisability of these students was measured with the Sussex Waterloo Scale of Hypnotisability (SWASH; Lush, Moga, McLatchie & Dienes, 2018) on group screening session at the university. This scale is a modified version of the WSGC (Bowers, 1993), it is adjusted to be ideal for screenings with larger groups of people and it contains items for each suggestion assessing the subjective experiences of the participants. Students, whose hypnotisability score is in the top fifteenth percentile based on their composite score (arithmetic mean of the subjective and objective scores), were recruited via email to attend the experiment (this threshold is comparable with those of employed by other researchers in the field; e.g., Anlló, Becchio, & Sackur, 2017; Barnier & McConkey, 2004). Based on the preregistration of the project (see details below), the stopping rule was to cease recruitment when the Bayes factor of the critical test (see details below) reached good enough sensitivity (either below 1/3 or above 3) or when had run 20 participants. All participants were informed about the nature of the study and all of them signed the consent form before participation. After finishing the experiment, the participants were debriefed and received a payment of £7 or course credit. The Ethical Committee of the University of Sussex (Sciences & Technology C-REC) approved the study.

Stimuli and apparatus

The stimulus set of the experiment closely followed those of the stimuli used by Raz et al. (2002). The stimuli contained 4 color words (RED, BLUE, GREEN, YELLOW) and 4 neutral non-color words (LOT, SHIP, KNIFE, FLOWER). The displayed color of

the color words can be red, blue, green or yellow depending on the congruency type of the current trial. Each neutral word can be only presented in the colors to whose name it is not matched for length. For instance, the word LOT can be coloured in blue, green and yellow, but not in red. The words were written in upper-case font and their vertical size was 0.6 cm (visual angle of 0.5º from 65 cm) and their horizontal size varied between 1.3 and 2.4 cm (visual angles of 1.146º - 2.115º). The stimuli were presented against a white background on a computer screen with a resolution of 1280x1024. The participants were able to indicate their answers about the displayed color of the words by pressing one of the following buttons on the keyboard: “V”, “B”, “N”, “M”. The Stroop task part of the experiment was run in Opensesame (Mathôt, Schreij, & Theeuwes, 2012).

Design

We employed a 2x2x3 within subjects design with independent variables: (i) presence of the word blindness suggestion (present vs. absent), (ii) proportion of incongruent trials in the current block (PI 10% vs. PI 80% where PI refers to the proportion of incongruent trials or simply proportion incongruence) and (iii) congruency type of the current trial (congruent vs. incongruent vs. neutral). The percentage of the neutral trials was held constant in all blocks (10%).

Procedure

Data collection was conducted in a small experimental room with the experimenter present. After providing informed consent, the participants were told that they will receive a hypnotic induction and a suggestion and that during the experiment they will be asked to report their subjective experiences and expectations about the effect of the suggestion. Next, they were informed that they will play a computer game where their task is to indicate the ink color of the word on the computer screen. They were asked to lay their left middle finger on V, left index finger on B, right index finger on N and right middle finger on M. The corresponding buttons of red, blue, green and yellow will be V, B, N and M, respectively. The buttons on the keyboard were neither color labelled nor marked. The importance of speed and accuracy were equally highlighted in the instructions (“Please respond as quickly and as accurately as you can.”). The participants were asked to focus on a black fixation cross that appeared in the middle of the screen and was replaced by a colored word after 500ms. They had 2000ms to indicate their answers, and after pressing one of the buttons, a feedback indicating the correctness of

their last answer appeared on the screen for 500ms (the word `CORRECT` or `INCORRECT` in black color). The interstimulus interval was 2000ms.

Figure 1 shows the process of the administration of the experiment. As a first step, the participants received 36 practice trials (12 trials from each congruency type). Performance was not assessed in the training session and all participants continued the experiment after finishing the practice block by receiving a standard hypnotic induction. .The hypnotic induction was followed by a single question to measure the `depth` of hypnosis (Hilgard & Tart, 1966; "How deeply hypnotised are you?"; 0 – normal state, 1 – relaxed, 2 – hypnotised, 3 – deeply hypnotised): . Next, we introduced the word blindness suggestion of Raz et al. (2002). To check the effect of the suggestion, participants were asked to rate the meaninglessness of a presented colored word on a 4 point Likert scale ("How strongly do you experience the word as meaningless?"; 1 – completely clear, 2 – little unclear, 3 – unclear, 4 – completely unclear). Participants answering 3 or 4 received the deactivation of the suggestion and the de-induction, while participants choosing 1 or 2 heard the following script from the experimenter before the deactivation of the suggestion and the de-induction to help them achieve the full potential of their response to the suggestion:

"Notice how as you look at the word on the screen, you can look at it with the meaning fading to the background of your mind. We have found even when people consciously experience some meaning after this suggestion, they still process the words differently at a deeper level. You know you are capable of not reading meaning fully, remember how you have zoned out while reading a book." (Raz et al., 2002)

After the de-induction, participants were asked again to report the depth of hypnosis they currently experience. Next, the participants engaged in four Stroop blocks: they received both types of blocks (PI 80% vs. PI 10%) with and without the suggestion. A clap activated and a double clap deactivated the suggestion. In each block, they received 120 trials. The order of the trials within the blocks was randomly generated and the order of the four blocks was counterbalanced between the participants. Before each Stroop block, the participants were asked to rate on a 5 point Likert scale how strongly they expected that the words on the screen would be meaningless. After each block, participants were asked to report their subjective experience about the effect of the word blindness suggestion by indicating the percentages of the trials with clear and meaningful

words. The subjective experience of meaninglessness was assessed with four items each time. In addition, we also asked them to indicate their experience about the depth of hypnosis they experienced during the last Stroop block. For the exact questions and choice options see Appendix A. These questions and the options were read out by the experimenter and were provided on a paper for the participants. Finally, the participants were thanked for attending and debriefed.

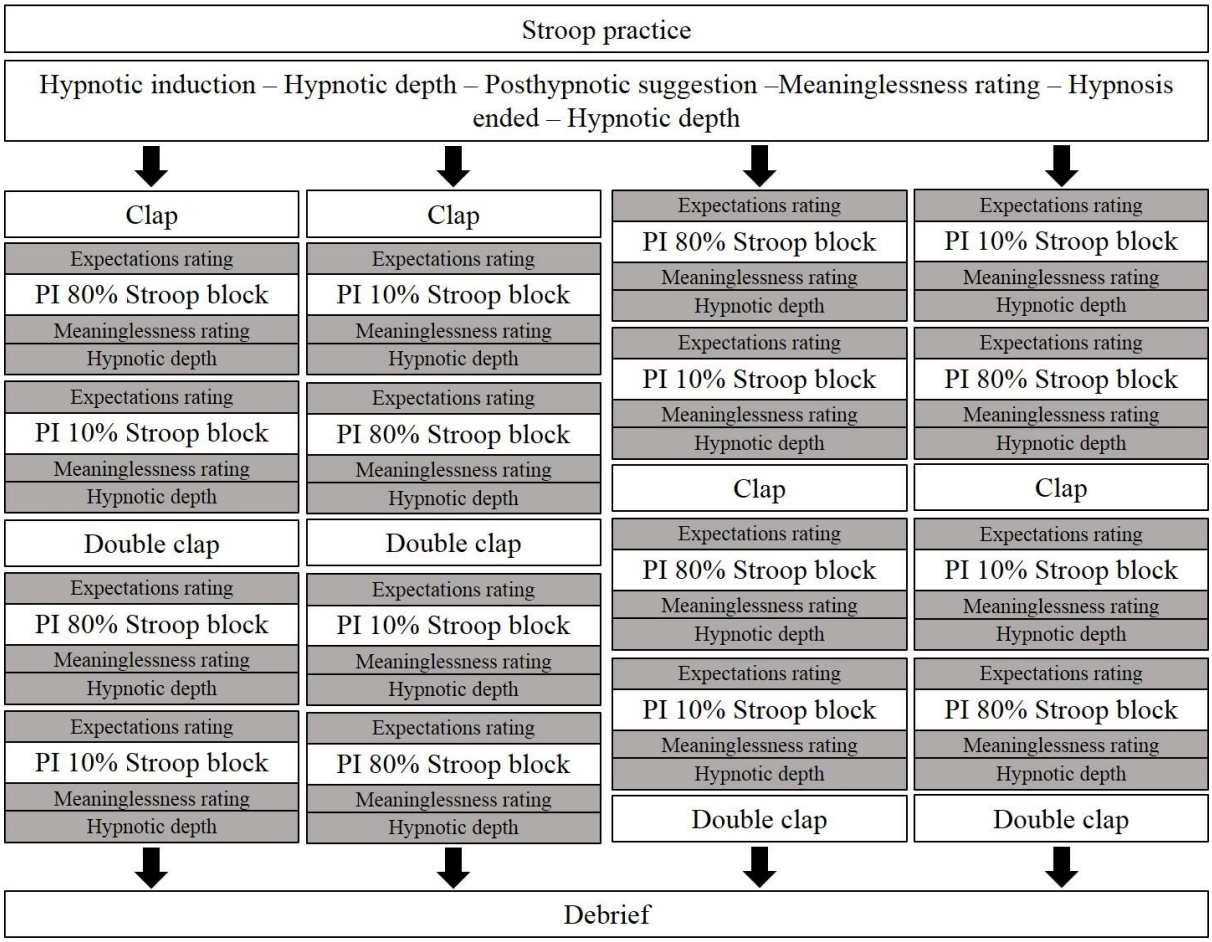

Figure 1. Flowchart depicting the four versions in which the experiment was conducted. The participants were randomly assigned to one of the four versions. The clap was used as a cue to activate the word blindness suggestion and the double clap was used to deactive the suggestion. The proportions of the incongruent, neutral and congruent trials in the PI80% and PI10% Stroop blocks were 80%, 10%, 10% and 10%, 10%, 80%, respectively.

Data analysis

Statistical analyses. We conducted the data transformation and ran all analyses with the statistical software R 3.3.1 (R Core Team, 2016). The R markdown script about data processing and the analyses can be retrieved from osf.io/pk2st. We had a within-

subjects design and, therefore, we simplified all of our tests to direct comparison of two conditions so that they could be conducted by running paired t-tests or simple linear regressions. We define Stroop interference as the RT difference between incongruent and neutral trials, and Stroop effect as the RT difference between incongruent and congruent trials. To explore the relationship between the RTs, the subjective reports and expectations, we ran linear regressions and tested the slopes against zero. In addition to the p-values, we calculated Bayes factors (Bs) and used them as a basis of decision making about the hypotheses. Moreover, to explore the extent to which the participants experienced being in a hypnotic state during the Stroop task, we conducted parameter estimation and reported the raw effect sizes (condition means) with the 95% Credibility Intervals (CIs). We calculated the 95% CIs by assuming a uniform prior distribution, thus, the values of the 95% CIs are numerically identical to the 95% Confidence Intervals.

Bayes factor. We used the Dienes and McLatchie (2018) calculator to compute the Bayes factor and extended it so that it can model the predictions of the alternative hypotheses (H1s) with a uniform distribution (See the R script of the calculator in Appendix B.). We employed t-likelihood function for each analysis and we set the degrees of freedom of the theory to 10^{10} so that the distribution of the likelihood function was comparable with a normal distribution. We report Bs as evidence for H1 over H0 in the following formats depending on the model of H1: $B_{H(0,\ X)}$ or $B_{U[0,\ X)]}$. The H in the subscript indicates that we used a half-normal distribution and the U indicates that we used a uniform distribution to model the predictions of the alternative hypotheses. For the half-normal models, the first character within the brackets indicates that the mode of the distribution was zero and X stands for the SD of the distribution, which varied among the analyses. For the uniform models, the first value within the brackets is the lower limit of the distribution, which was zero in all cases, and X marks the upper limit, which differed across the analyses.

We used half-normal models for the outcome neutral and crucial tests (i.e., analyses of RTs) as its application fits the scientific intuition that small effect sizes are more probable than large effect sizes; and we employed uniform distributions for the supporting and exploratory tests (all labelled as exploratory analyses in the preregistration) as we had no prior information about these questions. Nonetheless, a small effect size is more probable than a large one (Dienes, 2017), thus, we recalculated the *B*s based on uniform distributions to test the robustness of their conclusions to the

shape of the distribution modelling the predictions of H1 (See Supplementary Materials for results). To determine the parameters of the models, we used two strategies throughout the analyses. When we had estimates of effect sizes from the relevant literature, we applied those as the SDs of the half-normal models. For instance, for the analyses of the Stroop interference, the Stroop and the Suggestion effects (difference in the reduction of the interference effect between no suggestion and suggestion conditions), we informed our models based on the meta-analysis of Parris et al. (2013, Table 1) who summarized the results of 7 studies on the word blindness effect. We used the means of the Stroop interference, the Stroop and the Suggestion effects, which are 62ms, 90ms and 64ms, respectively. In the absence of prior information about the expected effect size, we derived the parameters of the models from the maximum effect size (Dienes, 2014). For the analysis with half-normal models, we took the half of the maximum effect size as the SD of the distribution. For instance, we halved the SD that we used to test the Stroop effect and employed it as the expected effect size for the test of the proportion congruency effect (that is, the Stroop effect is expected to be smaller in the blocks with higher proportion of incongruent trials). For the analyses with uniform models, we took the maximum effect size as the upper limit and zero as the lower limit of the distribution. For example, to test the effect of suggestion on the self-report scales, we simply used the maximum value of the scale as the upper limit, and the minimum value of the scale as the lower limit of the uniform distribution. For the regression analyses, the maximum of the possible effect size was the perfect linear relationship between the examined variables (from the minimum of each scale to the maximum, e.g., 100/4 for the test of the subjective measures and expectations: the ratio of scales heuristic), therefore, we used 100/4 as the upper and 0 as the lower limit of the model of H1s.

To distinguish good enough and insensitive evidence, we used the convention suggested by Jeffreys (1961), which is 3 for H1; and thus, by symmetry, 1/3 for H0. We also employed the evidence labels used by Lee & Wagenmakers (2013) to further differentiate between moderate ($3< B <10$ or $1/10< B <1/3$) and strong evidence ($B > 10$ or $B < 1/10$). Finally, as the value of a B depends on the features of the distribution we choose to model the predictions of a hypothesis (e.g., the SD of the half normal distribution; Rouder, Morey, Verhage, Province & Wagenmakers, 2016; Rouder, Morey & Wagenmakers, 2016), we report Bayesian robustness regions (RR) to assess the robustness of the conclusions of the acquired Bs to the value of the applied SDs. The

robustness regions are notated as RR[x1, x2] in which x1 is the smallest SD of the model that would produce a qualitatively identical conclusion as the chosen SD (e.g., if the *B* was greater than 3 then the *B* calculated with x1 equals to or is slightly greater than 3), whereas x2 is the largest SD that would bring us to the same conclusion.

Implementation of the preregistration

The protocol and the analysis plan were preregistered here https://osf.io/7ma4t/. The materials and procedure of the experiment closely followed those of that were determined in the preregistration. For the analysis, at a handful of steps, however, we deviated from this plan for either theoretical or practical reasons. First, consider our main hypothesis, which aimed to test whether a minimum amount of conflict is required to trigger the word blindness suggestion. We planned to use the extent of the Stroop effect as the index of conflict. However, based on a more thorough review of the literature (see introduction), we chose the proportion of incongruent trials as the index of conflict. Thus, the minimal required conflict hypothesis predicts that the suggestion should be more efficient in the high incongruence proportion blocks compared to the low incongruence proportion blocks. In addition, we tested the efficiency of the suggestion on the ability to reduce the Stroop interference effect rather than the Stroop effect as Parris et al. (2013) demonstrated in their meta-analysis that the suggestion affects the interference component of the Stroop effect robustly and the facilitation component (mean RT of neutral trials – mean RT of congruent trials) less often. In the main analyses. we also desisted from the usage of ratio scores (e.g., Stroop effect would have equalled to the congruent RTs divided by the incongruent RTs) to calculate the extent of the Stroop effect and employed the same procedure as other studies in the field of the word blindness effect, namely, the difference scores (e.g., Stroop effect equals to the difference of incongruent and congruent RTs, Stroop interference equals to the difference of incongruent and neutral RTs). Supplementary Materials presents the results with ratio scores; the pattern and conclusions are the same; in light of this, to aid the comparison of our results to other studies of the word blindness effect, we present the results of the difference score analyses here, and ratio scores in the Supplementary Materials. Second, we included an extra outcome neutral test to examine whether the manipulation of the proportion of incongruent trials successfully influenced the extent of the Stroop interference and Stroop effects in the different blocks (we expect larger effects in blocks with smaller incongruence proportions).

Results

Data transformation

Following Raz et al`s (2002) procedure, we excluded trials with errors and missed trials (8.09%). In addition, trials with response times (RTs) 3 standard deviations either above or below the mean (separately for each participant and experimental condition) were considered as outliers and omitted from all the further analyses (0.98%). In order to directly test our main hypothesis (three-way interaction of interference, incongruence proportion of the block and suggestion), we computed variables of difference scores. We calculated the extent of the Stroop interference (mean RT of Incongruent trials - mean RT of Neutral trials) in each block and suggestion condition.

Outcome neutral tests: Was there a Stroop interference effect and did the proportion of incongruent trials influence the extent of Stroop interference?

First, we confirmed that the latencies of the incongruent trials were the longest (*M* = 911 ms, *SD* = 146) followed by the neutral trials (*M* = 807 ms, *SD* = 113) and the RTs of the congruent trials were the shortest (*M* = 714 ms, *SD* = 112); see also panel A of Figure 2. There were Stroop interference ($t(22) = 5.82$, $p < .001$, $M_{diff} = 104$ ms, $d_z = 1.21$, $B_{H(0, 62)} = 6.97*10^3$, RR[8, $3.18*10^4$]) and Stroop effects ($t(22) = 13.42$, $p < .001$, $M_{diff} = 197$ ms, $d_z = 2.80$, $B_{H(0, 90)} = 3.80*10^9$, RR[10, $6.74*10^4$]). Furthermore, the extent of the Stroop interference effect was influenced by the type of the block ($t(22) = 2.94$, $p = .008$, $M_{diff} = 68$ ms, $d_z = 0.61$, $B_{H(0, 31)} = 11.78$, RR[13, 691]). Panel B of Figure 2 shows that the interference effect was larger in the low incongruence block (*M* = 138, *SD* = 124) than in the high incongruence block (*M* = 70, *SD* = 74). There was an interaction between the extent of the Stroop effect and the type of the block ($t(22) = 5.75$, $p < .001$, $M_{diff} = 125$ ms, $d_z = 1.20$, $B_{H(0, 45)} = 1.58*10^3$, RR[10, $3.79*10^4$]). Finally, the extent of the Stroop effect was larger in the low incongruence (*M* = 260, *SD* = 111) than in the high incongruence block (*M* = 134, *SD* = 55).

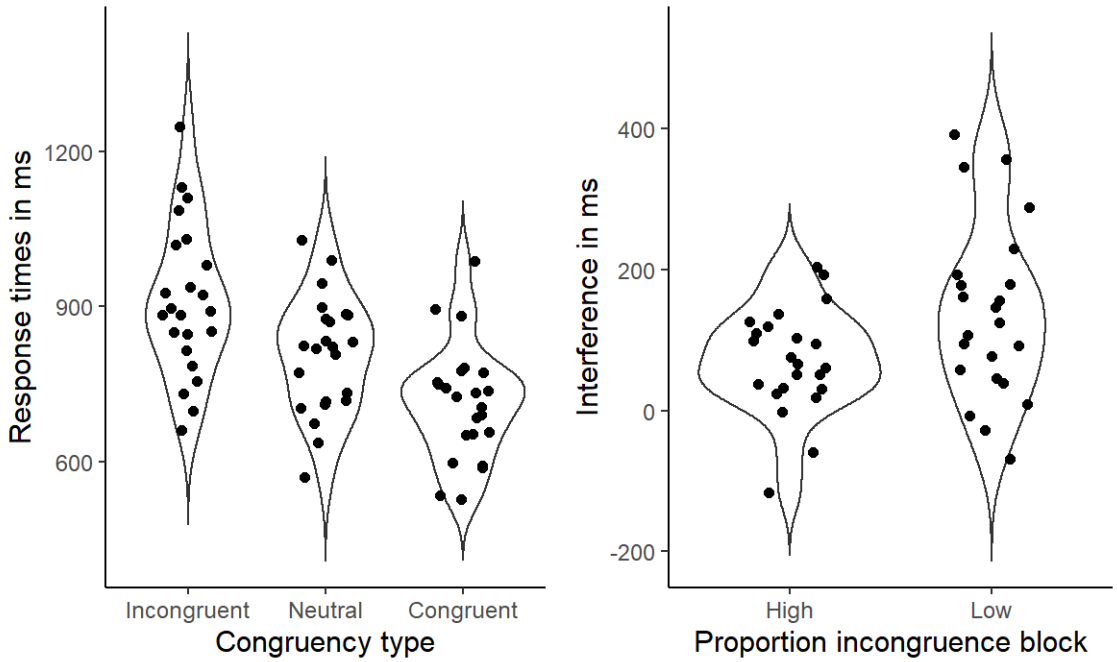

Figure 2. Panel A represents the distribution of the RTs broken down by the congruency type of the Stroop trials indicating the size of the Stroop interference (difference of incongruent and neutral condition) and the Stroop effects (difference of incongruent and congruent conditions). Panel B the extent of the Stroop interference effect separately for the high and low proportion incongruence blocks.

Crucial test: Is the suggestion more effective in the high incongruence than in the low incongruence block?

We examined the main question of interest by comparing the extent to which the suggestion reduced the interference effect in the high incongruence and low incongruence blocks. The results showed supporting evidence for the effect of the block ($t(22) = 1.86$, $p = .076$, $M_{diff} = 74$ms, $d_z = 0.39$, $B_{H(0, 64)} = 3.30$, RR[43, 101]) as the interference was diminished to a larger extent by the suggestion in the high incongruence block ($M = 58$ms, $SD = 113$) than in the low incongruence block one ($M = -16$ms, $SD = 161$). Separate analysis of the two blocks revealed that the suggestion worked only in the high incongruence block ($t(22) = 2.46$, $p = .022$, $M_{diff} = 58$ms, $d_z = 0.51$, $B_{H(0, 64)} = 7.79$, RR[15, 247]). There was anecdotal evidence that the suggestion did not lower the interference in the low incongruence block ($t(22) = -0.49$, $p = .629$, $M_{diff} = -16$ms, $d_z = -0.10$, $B_{H(0, 64)} = 0.34$, RR[0, Inf66]). Figure 3 shows a violin plot of the word blindness effect broken down by the type of the proportion incongruence block. For an exploratory analysis on the influence of the suggestion on the facilitation effect (mean RT of neutral trials - mean RT of congruent trials) see the Supplementary Materials. Table 1 depicts the

descriptive statistics broken down by suggestion conditions and proportion incongruence blocks.

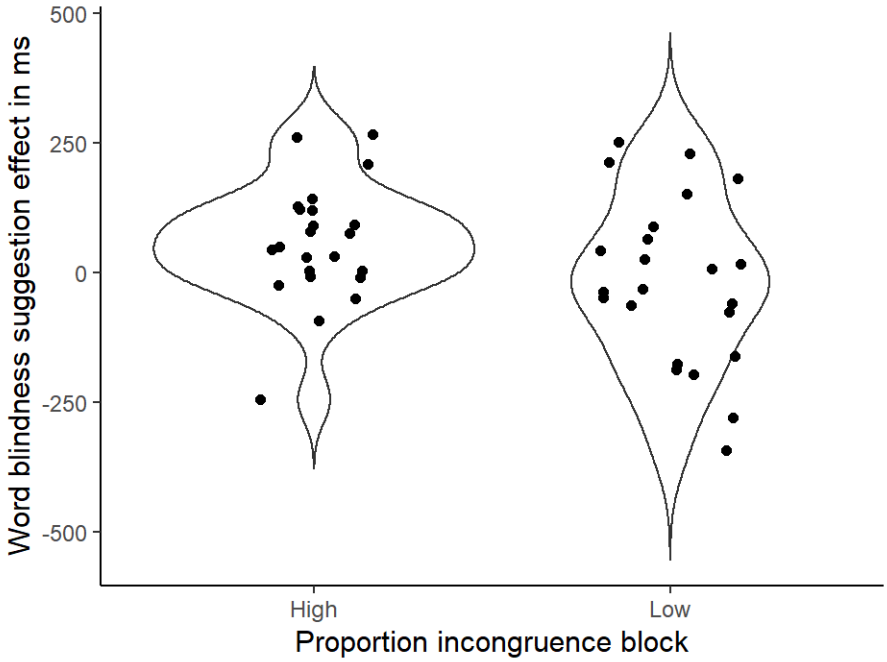

Figure 3. The violin plot shows the distribution of the extent of the word blindness effect in the high and low proportion incongruence blocks separately. The black dots on the plots indicate the average response of a particular participant.

Supporting test 1: Does the suggestion, the type of the block and their interaction influence subjects` experiences of word meaninglessness?

The participants experienced more meaningless words in the suggestion than in the no suggestion condition for all four items (all Bs > 4.15). There was strong evidence for the absence of the main effect of block for all items (all Bs < 0.08). Finally, the test of the interaction between the type of the block and the suggestion showed moderate evidence for the null for Q3 and Q4 (Bs < 0.17) and indicated data insensitivity for Q1 and Q2 (B equals to 1.13 and 0.47, respectively). Table 2 summarises the results of the statistical analyses.

Table 1.

Descriptive Statistics of the Response Times and Self-report Measures Broken Down by Suggestion Conditions and Proportion Incongruence Blocks

Category	Item (scale)	Low incongruence		High incongruence	
		No Suggestion	Suggestion	No Suggestion	Suggestion
Response times (RTs)	Incongruent (ms)	955 (225)	911 (187)	930 (178)	847 (152)
	Neutral (ms)	826 (150)	766 (131)	831 (149)	805 (152)
	Congruent (ms)	680 (121)	669 (92)	778 (171)	731 (127)
Expectations	Expecting the words to be meaningless (0-4)	0.32 (0.63)	1.35 (1.16)	0.34 (0.66)	1.08 (0.76)
Subjective experience of meaninglessness	Q1 (%)	80 (25)	61 (30)	86 (21)	55 (31)
	Q2 (%)	15 (23)	32 (27)	13 (24)	38 (30)
	Q3 (%)	17 (23)	32 (26)	16 (24)	36 (32)
	Q4 (%)	17 (26)	37 (30)	13 (19)	34 (30)
Depth of hypnosis	Experienced depth of hypnosis (0-3)	0.57 (0.72)	0.93 (0.74)	0.46 (0.69)	1.00 (0.90)

Note. The Standard Deviations (SD) of the means are shown within the brackets. Q1 is a reversed item compared to the other three subjective experience measures.

Table 2.

Results of the t-tests Investigating the Effect of Suggestion, Block and their Interaction on Subjective responses of meaninglessness

Item	Predictor	Statistics					
		M_{diff} (%)	d_z	t	p	$B_{U[0, 100]}$	RR
Q1	Suggestion	25	0.91	4.38	< .001	196	5, $6.51*10^3$
	Block	0	-0.02	-0.08	.936	0.04	13, Inf
	Suggestion*Block	12	0.43	2.09	.048	1.13	8, 339
Q2	Suggestion	22	0.88	4.22	< .001	120	5, $3.99*10^3$
	Block	2	0.17	0.83	.418	0.08	25, Inf
	Suggestion*Block	8	0.35	1.67	.110	0.47	0, 140
Q3	Suggestion	17	0.56	2.67	.014	4.15	7, 138
	Block	2	0.12	0.58	.565	0.06	20, Inf
	Suggestion*Block	5	0.17	0.82	.420	0.17	52, Inf
Q4	Suggestion	21	0.76	3.65	.001	33.66	5, $1.12*10^3$
	Block	-4	-0.22	-1.07	.294	0.02	7, Inf
	Suggestion*Block	0	0.01	0.05	.960	0.08	24, Inf

Supporting test 2: Does the suggestion affect subjects' expectations?

The participants expected to have a stronger experience of word meaninglessness in the suggestion condition (M = 1.22, SD = 0.81) than in the no suggestion condition (M = 0.39, SD = 0.64), ($t(36) = 3.85$, $p < .001$, $M_{diff} = 0.83$, $d_z = 0.80$, $B_{U[0, 4]} = 50.71$, RR[0.18, 67]).

Exploration 1: Is there any relationship between expectations and the subjective experiences of the suggestion?

In order to explore whether the expectations to experience the words as meaningless characters can predict participants` subjective experience of meaninglessness, we conducted linear regression analyses. We computed the raw regression slope in the suggestion and no suggestion conditions separately as well as the slope of the difference scores ($Expectation_{Suggestion}$ - $Expectation_{No\ Suggestion}$ as a predictor and Subjective $experience_{Suggestion}$ - Subjective $experience_{No\ Suggestion}$ as an outcome). In the no suggestion condition, the slope differed from zero for all items (all Bs > 3.97 and all bs > 13) apart from Q2 (B = 1.73, b = 10). In the suggestion condition there was strong evidence for Q1, Q2 and Q3 (Bs > $8.08*10^2$and bs > 23) and the results were insensitive for Q4 (B = 2.54, b = 12). Lastly, the analysis of the difference scores revealed a positive relationship between expectations and subjective experiences for all four items (Bs > $2.32*10^2$ and bs > 16). Table 3 depicts the results of the regression analyses.

Exploration 2: Is there any relationship between the objective responses and the expectations of the participants?

Next, we investigated the link between expectations and objective responses, specifically, the extent of the interference effect in the high incongruence blocks (as we only found evidence for the presence of the word blindness effect in the high incongruence blocks). Similar to the previous analysis, we calculated the raw regression slopes in the suggestion and in the no suggestion conditions as well as for their interaction by running the analysis on the difference scores. The test of the slopes against zero revealed insensitive evidence in all cases. For the no suggestion condition ($t(22) = 1.97$, $p < .001$, b = 57ms, $\beta = 0.39$, $B_{U[0, 25]} = 2.07$, RR[0, $1.38*10^3$]); for the suggestion condition ($t(22) = 0.89$, $p = .384$, b = 20ms, $\beta = 0.19$, $B_{U[0, 25]} = 1.35$, RR[0, 204]); and for the interaction of the conditions ($t(22) = -0.02$, $p = .987$, b = 0ms, $\beta = 0.00$, $B_{U[0, 25]} = 0.85$, RR[0, 94]).

Table 3.

Results of the Regression Models Predicting the Experience of Word Meaninglessness Based on the Subjects` Expectations

		Statistics					
Item	Condition	*b* (%)	*β*	*t*	*p*	$B_{U[0, 25]}$	RR
Q1	No Suggestion	28	0.86	7.70	< .001	$2.99*10^5$	4, $8.30*10^4$
	Suggestion	26	0.74	4.98	< .001	$1.39*10^3$	4, $2.76*10^4$
	Difference	22	0.82	6.68	< .001	$9.26*10^4$	3, $6.40*10^4$
Q2	No Suggestion	10	0.29	1.40	.177	1.73	0, 131
	Suggestion	23	0.71	4.63	< .001	$8.08*10^2$	4, $1.05*10^4$
	Difference	16	0.66	4.04	< .001	$2.32*10^2$	4, $1.95*10^3$
Q3	No Suggestion	19	0.57	3.14	.005	36.14	6, 352
	Suggestion	25	0.73	4.85	< .001	$1.15*10^3$	4, $1.84*10^4$
	Difference	24	0.78	5.78	< .001	$1.03*10^4$	4, $6.38*10^4$
Q4	No Suggestion	13	0.39	1.98	.063	3.97	10, 34
	Suggestion	12	0.34	1.68	.109	2.54	0, 196
	Difference	18	0.66	4.08	< .001	271	3, 573

Note. The raw slopes are indicated by *b*, whereas the standardised effect sizes are indicated by β.

Exploration 3: Is there any relationship between the subjective and objective responses of the participants?

The logic of the analysis of the question whether the subjective reports of meaninglessness can predict the extent of the Stroop interference effect in the different conditions closely followed that of Exploration 1 and 2. The regression slopes in the no suggestion condition suggest that seeing more words as meaningless characters predict faster RTs (note that positive *bs* indicate reduction in RTs owing to the increase of the experience of meaninglessness), whereas the slopes in the suggestion condition and the slopes calculated from the difference scores depicted mixed results. The statistical analyses showed that the evidence is not enough in any case to come to a decision whether the experience of word meaninglessness can predict RTs. Table 4 summarises the results of the regression models.

Table 4.

Results of the Regression Models Predicting Response Times Based on Subjects` Experience of Word Meaninglessness

		Statistics					
Item	Condition	*b* (ms)	β	*t*	*p*	$B_{U[0, 1]}$	RR
Q1	No Suggestion	0.57	0.11	0.53	.603	1.11	0, 6.61
	Suggestion	0.00	0.00	0.00	.998	0.65	0, 2.15
	Difference	0.07	0.02	0.11	.915	0.77	0, 2.72
Q2	No Suggestion	0.64	0.15	0.68	.505	1.19	0, 6.81
	Suggestion	0.25	0.10	0.44	.663	0.91	0, 3.24
	Difference	0.38	0.11	0.50	.622	1.04	0, 4.51
Q3	No Suggestion	0.93	0.21	1.00	.326	1.45	0, 9.82
	Suggestion	-0.17	-0.07	0.32	.753	0.52	0, 1.63
	Difference	0.14	0.05	0.24	.813	0.80	0, 2.77
Q4	No Suggestion	0.76	0.14	-0.64	.529	1.17	0, 1.27
	Suggestion	-0.76	-0.29	-1.37	.185	0.33	0.98, Inf
	Difference	-0.23	-0.07	-0.32	.754	0.63	0, 2.18

Note. The raw slopes are indicated by *b*, whereas the standardised effect sizes are indicated by β.

Discussion

In the current study, we manipulated the proportion of incongruent Stroop trials across blocks, in order to influence the level of conflict and so to test whether the activation of the word blindness suggestion requires a minimum level of conflict. Our results revealed that highly suggestible people can reduce the Stroop interference effect (respond to the suggestion) to a greater extent in a Stroop task with high levels of conflict (high incongruence block), compared to a Stroop task with low levels of conflict (low incongruence block). This finding supports the response competition model (Augustinova & Ferrand, 2012; Parris, Dienes, & Hodgson, 2013) of the word blindness suggestion, over the de-automatisation of reading model (Raz et al., 2002, 2006; Raz & Campbell, 2011). The former model predicts that a minimum amount of conflict is necessary for the activation of the suggestion, whereas the latter assumes that the operation of the suggestion is independent of the registered conflict.

Given the assumption of cold control theory (Barnier, Dienes & Mitchell, 2008; Dienes, 2012; Dienes & Perner, 2007) that the sole difference between hypnotic and non-hypnotic responding is the form of the accompanying HOT of intending, our findings

imply that conflict plays a crucial role in the activation of unconscious control. This indicates that the conflict-monitoring model (Botvinick et al., 2001; Botvinick, Cohen, & Carter, 2004) might be generalizable for unconscious control and the awareness of the intention may not be a requisite of this monitoring process. Coinciding with this view, Parris, Dienes, & Hodgson (2012) found that the word blindness suggestion is subject to the same time constraints as the CSE (Egner, Ely & Grinband, 2010), namely, shorter Response-Stimulus Interval (RSI, 500 ms) results in a stronger suggestion effect (stronger reduction in the interference effect) than a long RSI (3500 ms). To test the similarity of the mechanisms underlying the activation of regular control processes and the hypnotic response, they tested whether the congruency type of a trial can modulate the efficiency of the suggestion on a subsequent trial. They found no evidence for the idea that the suggestion reduces interference more strongly after an incongruent trial than after a neutral or congruent one. Nevertheless, the analysis of the traditional CSE effect revealed no evidence in either direction, implying that the design itself was not sufficient to test the effect of preceding conflict on the activation of control. Indeed, it has been demonstrated that in Stroop tasks that employ four colors and color words the emergence of a CSE effect is unlikely (Puccioni & Valessi, 2012). Future research could shed more light on whether hypnotic responding and so unconscious control can be triggered by recently registered conflict as well or it is only subject to the relative frequency of conflicting trials.

The application of multiple experimental manipulations that affect the extent of the Stroop interference effect in tandem (i.e., the proportion of incongruent trials within a block and the word blindness suggestion) can necessitate a different analysis plan than a simple calculation of difference scores. Blocks with low proportion of incongruent trials tend to have a greater baseline of Stroop interference than blocks with high proportion of incongruent trials (PCE). Therefore, a three-way interaction with difference scores can be found even when the suggestion, for instance, halves the interference effect in both of the blocks. In this case, a difference score analysis would give the false impression that the suggestion is more effective in the low incongruence than in the high incongruence block. Using ratio rather than difference scores overcomes this issue as the test of the ratio scores would only be sensitive if the suggestion is proportionately more effective in one block compared to the other (e.g., interference reduces to its half in one block and to its quarter in the other block). Nonetheless, it worked out in our case that the ratio score analysis

yielded the same conclusion as the difference score analysis, because the largest differences occurred for the smallest baselines.. Thus, we reported the classical difference score analysis in the main paper so that our results can be directly contrasted to other studies of the word blindness effect and the dependent variable can be unambiguously interpreted (See the Supplementary Materials for the analysis of the ratio scores). Choosing one analysis specification over another can increase the researchers' degrees of freedom exacerbating the rate of false-positive findings (Simmons, Nelson & Simonsohn, 2011). Multiverse analyses (Steegen, Tuerlinckx, Gelman & Vanpaemel, 2016) and specification curves (Simonsohn, Simmons & Nelson, 2015) can reduce the possibility of selective reporting and enhance the credibility of reported findings through an increased transparency of the consequences of the analytic choices.

The participants` subjective reports of word meaninglessness differed in the suggestion and no suggestion conditions but, interestingly, they did not discriminate between the high and low conflict environments, implying that the subjective experience of meaninglessness is unrelated to the level of Stroop interference (see also Parris, Dienes & Hodgson, 2012). Nonetheless, these results proved to be dependent on the shape of the distribution used to model the predictions of H1, and the evidence for the null is insensitive when a plausible model favouring small effect sizes is contrasted to H0 (Tables S2-S3, Supplementary Materials). Moreover, the direct tests of the link between the subjective experiences of meaninglessness and the objective scores of the Stroop interference reductions were all insensitive (with both uniform and half-Cauchy distributions) precluding any robust conclusion about the relation between the subjective experience of meaningless and Stroop reduction. Demonstrating support for the null hypothesis has proven to be challenging for experimental psychologists as the classical procedure of null hypothesis significance testing cannot provide evidence for the null even for a relatively large p-values (Fisher, 1935, Royal, 1997). Despite this, non-significant tests are still frequently used to support claims about the absence of effects (Aczel et al., 2018). The application of the Bayes factor overcomes this issue by allowing us to assess the relative strength of evidence for two competing models (Dienes, 2014, 2016), however, *B* depends on the shape and parameters of the distribution chosen to represent the predictions of H1 and H0 (Dienes, 2017; Rouder et al., 2016; Rouder, Morey & Wagenmakers, 2016). For instance, putting too much weight on large effect sizes when modelling the predictions of H1 could result in evidence for H0, but it is evidence against

only a vague model of H1, which precludes inference about theories (when there is information for the theory to be more precise). Thus, to make theoretically-motivated inferences based on *B*s, one needs to ensure that the predictions of the theories under comparison are represented fairly by the models. This is the main reason why we refrain from drawing specific conclusions based on the *B*s calculated with uniform distribution as a model of H1s.

The impact of expectations

The response expectancy theory (Kirsch, 1985, 1997, Kirsch & Lynn, 1997; Lynn, 1997), emphasise the causal role of expectancies in the activation and modulation of responses (e.g., subjective experiences, behaviors) to suggestions. Our results are in line with this theory insofar as participants expected to see more meaningless words in the suggestion condition, compared to the no suggestion condition. Moreover, the difference of the expectations strongly predicted the difference of the subjective experiences of meaninglessness between suggestion and no suggestion conditions (for all four items), implying that the participants were able to produce the anticipated subjective responses.

The evidence for the relationship between the expectations to see meaningless words and the capacity to reduce the interference effect was not decisive, since all regression analyses were insensitive. Nonetheless, it is important to note that our expectancy measures assessed word meaninglessness and not the experience of increased control. This difference might prove to be crucial as the current study provided further evidence for the notion that enhanced control over response competition, and not the de-automatisation of reading, is responsible for the reduction of the interference effect. Thus, future studies should consider including a new item measuring the expected ease of naming the color of the words during the Stroop task. Acquiring sensitive data here would help us understand whether placebo-suggestions and imaginative suggestions can both modulate behavior by influencing expectancies (Kirsch, 1985, 1999). For instance, a recent study demonstrated that the Stroop interference measured as the difference between the accuracy on incongruent and neutral trials can be modulated by a placebo-suggestion (persuading people that in the placebo condition the accuracy of their color discrimination will be enhanced), whereas an imaginative suggestion (a suggestion to imagine the ability of having an improved color perception) worded akin to the placebo-suggestion did not modulate the Stroop interference (Magalhães De Saldanha da Gama, Slama, Caspar, Gevers, & Cleeremans, 2013). These results seemingly count against the

response expectancy theory, however, the participants in this study were not screened for hypnotisability and so only a fraction of them were highly suggestible; and it has been shown that only highs and mediums can respond to imaginative suggestions that help them enhance the resolution of response competition (Parris & Dienes, 2013). Therefore, additional inquiry is needed to address whether expectancies are reliable predictors of responses (i.e., performance on the Stroop task) not just to placebo-suggestions but to imaginative suggestions as well.

Limitations

It is to be noted that the current study used a specific sample, namely highly suggestible individuals, and so the finding that unconscious control can be triggered by conflict may not be generalizable to non-highs. In addition, we applied a single interference task and suggestion, which may also restrict the generalizability of our findings and so future research could aim to implement conceptual replications of our study by employing suggestions claimed to reduce the Flanker compatibility (Iani, Ricci, Gherri, & Rubichi, 2006), the Simon (Iani, Ricci, Baroni, & Rubichi, 2009) or the McGurk effects (Lifshitz, Bonn, Fischer, Kashem, & Raz, 2013).

The most distinctive feature of a hypnotic response is the experience of involuntariness (Weitzenhoffer, 1980) and this sensation is also the key element of an unconscious control process (Dienes, 2012; Dienes & Perner, 2007). Although in the presented experiment, we did not gauge directly whether our participants were aware of their intentions to impair the meaningfulness of the words, in another study of ours, we have demonstrated that highs report reduced control over how meaningful the words appear to them when the word blindness suggestion was activated compared to the case when it was not (Palfi, Parris, McLatchie, Kekecs & Dienes, 2018). This finding underscores the idea that the implementation of this suggestion is accompanied by an inaccurate HOT about the intention and it is in accord with the interpretation of hypnotic responses to the word blindness suggestion as unconscious control processes. Future studies using (post)hypnotic suggestions as tools to examine unconscious control processes should consider collecting subjective reports measuring the extent to which participants experience involuntariness during the execution of the hypnotic response (e.g., Polito, Barnier, & Woody, 2013; Polito, Barnier, Woody, & Connors, 2014).

Conclusion

In sum, the moderating effect of conflict on the operation of the word blindness suggestion demonstrates compelling evidence for the notion that the suggestion takes its effect on response competition rather than on semantic processing. This finding supports also the idea that conflict plays a crucial role in the activation of the word blindness suggestion by enhancing the recruitment of control processes. Hence, given the assumption that hypnotic responding is intentional and it is accompanied by the feeling of involuntariness, this study is the first to indicate that cognitive control with and without the awareness of intentions can be activated by the very same process, namely, the monitoring and registration of conflict as outlined in the conflict-monitoring theory (Botvinick et al., 2001; Botvinick, Cohen, & Carter, 2004). This explanation is in line with the view about the strategic nature of hypnotic responding (Spanos & Barber, 1974; Comey & Kirsch, 1999; Hilgard, 1977; Spanos, 1986) as it depicts a model in which the extent of the exerted control is determined by the demands of the environment, namely, by the difficulty of the current proportion incongruence block. Of interest is that the present results suggest that the suggestion does not operate according to the instructions in the suggestion itself. Participants are processing the words to the level of meaning acquisition and even if they report to perceive them as words of a foreign language or as gibberish, this experience of meaninglessness is not a consequence of a deteriorated reading ability. Instead, participants seem to understand the desired outcome and unconsciously simulate that outcome by operating over existing control mechanisms conferring a level of control in the Stroop task difficult to achieve by conscious means.

Chapter III: Strategies that reduce Stroop interference

Introduction

An essential feature of the human cognitive system is its ability to attend to and utilise goal-related stimuli while it ignores the distractors of the environment. The Stroop task (Stroop, 1935; for a review see MacLeod, 1991) provides a window into selective attention and since its publication it has inspired many theories of attention and cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Cohen, Dunbar, & McClelland, 1990; Engle & Kane, 2004; Kalanthroff, Davelaar, Henik, Goldfarb, & Usher, 2018; MacLeod & MacDonald, 2000). This task requires participants to name the displayed color of the presented words while they should disregard the meaning of the words. People produce the quickest responses on congruent trials in which the meaning of the presented word is in accordance with its displayed color (e.g., RED displayed in *red*), followed by the neutral trials in which the meaning of the presented words is unrelated to colors (e.g., LOT displayed in *red*). The slowest response times (RTs) can be observed on incongruent trials where the displayed color and the meaning of the words are not in harmony (e.g., RED displayed in *blue*). Performance on the task can be assessed by computing the RT differences between these experimental conditions. The Standard Stroop effect is the RT difference of incongruent and congruent trials, and it can be broken down to two components. Namely, the Stroop interference effect which is the RT difference of incongruent and neutral trials, and the Stroop facilitation effect, which is the RT difference of the neutral and congruent trials.

The Stroop effect is remarkably large, and many report experiencing the accompanying cognitive conflict during an incongruent trial (MacLeod, 1991). A long line of research has demonstrated that the Stroop effect is very robust, it persists despite long term training (e.g., MacLeod, 1998) and bringing it under control through the application of deliberate strategies is difficult (MacLeod, 1991). Whilst methods have been reported that result in reduced Stroop effects (De Jong, Berendsen & Cools, 1999; Besner, Stolz, Boutillier, 1997; Parris, Sharma & Weekes, 2007) all involve a manipulation of the stimulus context (e.g. coloring a single letter instead of all letters or decreasing the response-stimulus interval) so as to provide exogenous support to control mechanisms, and are thus not likely the consequence of deliberate, top-down control. Even financial rewards offered to increase motivation to perform well results in either no

effect on reaction times other than a general speeding up on all trial types (Krebs et al., 2010; Veling & Aarts, 2010) or only small (~10ms) reductions of the Stroop effect (Krebs, Boehler, & Woldorff, 2010).

One of the few exceptions to the robustness of the Stroop effect may be provided the by the word blindness posthypnotic suggestion (Raz, Shapiro, Fan, & Posner, 2002; Parris, Dienes, & Hodgson, 2012). When the word blindness suggestion, a suggestion to see the words during the Stroop task as gibberish or meaningless characters, is given to highly suggestible people (henceforth highs), they can reduce and sometimes even eliminate the Stroop effect compared to the standard, no suggestion condition. This finding has been replicated by the original authors as well as independent labs (e.g., Augustinova & Ferrand 2012; Parris, Dienes, & Hodgson, 2012; Parris & Dienes, 2013; Raz et al., 2003, 2005; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006). The magnitude of the Stroop interference in the suggestion condition is usually half the size of the effect in the no suggestion condition (for a meta-analysis see Table 1 of Parris, Dienes & Hodgson, 2013). Nonetheless, the influence of the word blindness effect on the facilitation component of the Stroop effect appears to be more volatile. Importantly, responding to the suggestion speeds up the RTs of the incongruent trials compared to the no suggestion condition as well as compared to the control group of low suggestible people. Hence, the effect is an interesting use of cognitive control that is not produced simply by holding back on neutral and congruent trials (thereby equalising RTs on all trials; MacLeod, 2011).

The question arises of what exactly happens when highs respond to this posthypnotic suggestion? Many of the theories of hypnosis concur that responding to a hypnotic suggestion involves top-down cognitive control processes and that the feeling of involuntariness, which is the central feature of the hypnotic phenomena (Weitzenhoffer, 1974, 1980), is the result of a deteriorated or relinquished metacognition (Bowers, 1990; Dienes & Perner, 2007; Hilgard, 1991; Kihlstrom, 1985; Spanos 1986; for a review see Terhune, Cleeremans, Raz & Lynn, 2017)[8]. Cold control theory takes

[8] One exception to this is the response expectancy theory (Kirsch, 1985, Kirsch & Lynn, 1997), which provides a clear explanation of hypnotic responding that does not involve altered metacognitive processes. The theory postulates that expectations, produced by hypnotic suggestions, are enough by themselves to create the experiences and behaviour of hypnotic subjects. The subjects feel these responses involuntary due to the processes being truly unintentional, as there is no need to involve intentional cognitive control processes. This theory is not mutually exclusive with the theories involving cognitive control and metacognitive processes. However, measured expectations do not fully account for hypnotic responding

reduced metacognition as the fundamental process of hypnotic responding. Specifically, it asserts that hypnotic responding is implemented by intentional control, the subjects engage in strategies to create the experiences described in the suggestion but they are able to alter their monitoring over their intentions and make themselves believe that they are not acting deliberately (Barnier, Dienes & Mitchell, 2008; Dienes, 2012; Dienes & Perner, 2007). The theory draws on the higher order thought (HOT) theories of consciousness (Lau & Rosenthal, 2011; Rosenthal, 2005) according to which a mental state becomes conscious by virtue of a higher order state referring to it. For instance, to create the experience of a buzzing mosquito, one can form the following first-order intention: "imagine a buzzing mosquito". To be aware that one is engaged in imagination, one would need a second-order state that refers to the first-order state (i.e., "I intend to imagine a buzzing mosquito"). One can also create the experience of this noise without being aware of the first-order intention (i.e., "I do not intend to imagine a buzzing mosquito"), and in that case it would feel as if it happened by itself akin to the experience of hallucination. Importantly, this experience of involuntariness is what hypnotic subjects report about their behaviour when they respond to suggestions. Taken together, according to cold control theory, responding to a suggestion consists of engaging in a strategy to produce the experience described in the suggestion without being aware of using a strategy. From this assumption, it follows that the sole difference between a hypnotic and a non-hypnotic response is the form of the accompanying second-order state. Therefore, if one is capable of reducing the Stroop interference effect by responding to the word blindness suggestion, one should be able to do it by voluntary, non-hypnotic means as well, using the very same strategy that they used when they responded to the suggestion.

To create the experience of meaninglessness and reduce the Stroop interference effect, subjects may use a specific strategy or they might engage in different strategies to achieve the same outcome (Egner & Raz, 2007). We review four unique strategies here that have the potential to be regarded as an underlying mechanism of the word blindness suggestion. The most straightforward candidate is the looking-away strategy. Subjects may divert their attention from the word so that they can easily process the color but not the meaning of the word, which can result in a reduced interference. Indeed, it has been

(Benham, Woody, Wilson, & Nash, 2006; Shor, Pistole, Easton, & Kihlstrom, 1984). These findings may be due to measure unreliability but they also give rise to alternative accounts such as the metacognitive theories of hypnotic responding. Therefore, in this paper we focus on the explanation and predictions of the metacognitive theories to understand the underlying mechanism of the word blindness suggestion.

demonstrated that lows can reduce the Stroop interference by diverting their attention from the words (Raz et al., 2003). However, Raz et al. (2002, 2003) argued that it is unlikely that highs engage in this strategy when they respond to the suggestion. First, subjects reported that they observed the words in all instances and that they claimed that they did not engage in any attention related strategies. Second, the experimental sessions were videotaped and independent judges were unable to distinguish between highs and lows based on their eye-movement patterns. Nonetheless, these arguments are not bulletproof. As stated earlier, it is the essence of hypnosis that when subjects respond hypnotically, they can engage in strategies without being aware of doing so (Dienes & Perner, 2007; Spanos 1986), hence, asking them whether they used any strategies may not be a sensitive way to explore the underlying mechanism of the suggestion. Moreover, human judges might not be able to pick up on eye-movement patterns, rather an objective criterion based on, for instance, the fixation time outside of the area of interest defined around the words, which could provide a more severe test of the strategy.

A more subtle form of the looking-away strategy is when subjects focus their attention towards a single-letter or a portion of a letter of the word so that they can more easily name the color of the word. There is ample evidence that coloring only the last or the first letter of a Stroop word compared to the middle letter decreases the size of the Stroop interference effect (Augustinova, Flaudias & Ferrand, 2010; Besner et al., 1997; Parris et al., 2007; for a review see Flaudias & Llorca, 2014). Moreover, highs can respond more quickly on incongruent trials when this strategy is given to them in a hypnotic context (Sheehan, Donovan, & Macload, 1988; cf., Jamieson & Sheehan, 2004). Nonetheless, the Sheehan et al. study lacked a non-hypnotic strategy condition, hence, it is unclear whether the inclusion of hypnosis in the strategy condition did not increase the motivation and expectations of highs compared to the non-hypnotic baseline condition. The lack of appropriate control could create a “hold back” effect (Spanos, 1986; Zamansky, Scharf & Brightbill, 1964) in the non-hypnotic baseline condition as a way of satisfying demand characteristics.

Another vision-related candidate strategy is blurring. Subjects may adjust visual accommodation (e.g., by relaxing of the muscles around their eyes) so that the image of the word does not fall on the retina. Blurring may help them prioritise the color of the word over the meaning. Raz et al. (2003) provided a test of this strategy by giving a pharmacological agent to highs to disrupt visual accommodation, or in other words induce

the state of cycloplegia. The subjects were exposed to 2 drops of 1% cyclopentolate hydrochloride and their vision was corrected by lenses so that they saw the words crisply during the Stroop task. Highs still decreased the Stroop interference effect when they responded to the suggestion compared to the no suggestion condition. One might therefore conclude that highs achieved the reduction by means other than visual blurring. However, this conclusion is conditional on the participants being in a state of complete cycloplegia. There was no outcome neutral test testing whether the participants had completely lost their ability to accommodate. The authors point out that residual accommodation can still occur, especially for younger participants, when this particular agent is used.

Finally, there is evidence that subjects spontaneously resort to a strategy that involves the rehearsal of the task instructions, such as “focus on the color of the word” (Sheehan, Donovan, & Macload, 1988). Goal-maintenance has been shown to play a critical role in task performance in the Stroop task, therefore, a strategy that sustains an active goal representation might help participants mitigate Stroop interference (De Jong et al., 1999; Kane & Engle, 2003; Parris, Bate, Brown & Hodgson, 2012).

The purpose of this project is to explore the underlying mechanism of the word blindness suggestion by testing whether any of these four strategies (looking-away, visual blurring, single-letter focus and goal-maintenance) could be one that highs use when they respond to the suggestion. To this aim, we designed a fully within subjects experiment in which participants undertook the Stroop task in five blocks: in four blocks they were explicitly asked to use one of the mentioned strategies and in one block they were told to not use any of these strategies. According to the cold control theory, if a strategy can be applied hypnotically to reduce the Stroop effect, it should be available and applicable non-hypnotically as well. Hence, the experiment was administered outside of the hypnotic context; in fact, no reference was made to hypnosis or to the word blindness suggestion. The prime test was whether each strategy could reduce Stroop interference.

As a secondary analysis, we were able to test whether or not the efficiency of a specific strategy is related to hypnotisability beyond the effect of expectations and motivations conditional on a hypnotic context. Cold control theory postulates that individual differences in hypnotisability are grounded in differential metacognitive skills (which may or may not be limited to the domain of intentions). Consequently, lows and

mediums should be able to use a specific strategy just as efficiently as highs, when they are sufficiently motivated. If the results revealed a positive relationship between hypnotisability and strategy usage outside of the hypnotic context, the purely metacognitive account of hypnosis would need to be revised. To test this assumption, we recruited participants from a subject pool where the majority of the people were screened for hypnotic hypnotisability so that we would not need to disclose this hypothesis to the participants. The consent to link results to hypnotisability scores was acquired after the experiment; therefore, it is unlikely that they could associate the current experiment in any way with hypnosis or hypnotisability.

The word blindness effect is a unique phenomenon in which highs can reduce the Stroop interference in a way that is not fully understood yet. Unravelling its mechanisms should allow testing of theories of hypnotic responding as well as theories of the Stroop effect and more broadly theories of cognitive control.

Experiment 1

Methods

Participants. We recruited 78 participants from which 57 (mean age = 19.61, SD = 1.47, females = 51) had been screened for hypnotisability with the Sussex-Waterloo Scale of Hypnotisability (SWASH; Lush, Moga, McLatchie & Dienes, 2018). As we specified in the pre-registration, we excluded the data of those who did not have a SWASH score from all of the analyses. The experiment was advertised for first and second year psychology students of the University of Sussex who finished a module earlier in which they had the opportunity to participate in a hypnosis screening session. High and low hypnotisability were defined as scoring in the top and bottom 15% of the SWASH, respectively. We calculated the cut-off a priori based on the composite (objective and subjective) SWASH scores of all the first and second year students in our database. The cut-off for highs was 5.35 whereas the cut-off for lows was 2.00 (on a scale of 0 to 10). From the 57 participants, 10 were high, 39 medium and 8 low hypnotisables. The participants were proficient readers of English and they attended the experiment in exchange for course credits. All participants gave their informed consent before the experiment as well as after the experiment when we revealed that we wished to correlate their performance with their hypnotisability scores. The Ethical Committee of the University of Sussex approved the study.

Stimuli and apparatus. In order to aid comparability of the current experiment and the experiment of Raz et al. (2002), the stimuli closely followed those used by Raz et al. (2002). The stimulus set included 4 types of color words (RED, BLUE, GREEN, and YELLOW) and 4 types of neutral words (LOT, SHIP, KNIFE, and FLOWER). The congruent trials consisted of color words presented in colours matching the meaning of the words (e.g., RED in the color red). The incongruent trials were color words displayed in colours mismatching the meaning of the word (e.g., RED in the color blue). The color and the neutral words were frequency and length matched. All words were written in upper-case font and presented against a white background. The vertical visual angle of the stimuli was 0.5º, while the horizontal visual angle of the stimuli lied between 1.3º and 1.9º depending on the length of the word. The distance between the participants` eyes and the computer screen was approximately 65cm. The response keys used in the experiment were "V", "B", "N", "M" for the colours red, blue, green and yellow, respectively. The keyboard buttons were not colour-labelled. The experiment was produced in and run by the software Opensesame (Mathôt, Schreij, & Theeuwes, 2012) on a computer with a screen resolution of 1366 x 768.

Design and procedure. The study had a 3x5x3 mixed design with the independent variables of the congruency type of the trial (congruent vs. neutral vs. incongruent), the strategy used in the conditions (no strategy, looking-away, blurring, single-letter focus, goal-maintenance) and hypnotisability (low, medium or high)[9]. The proportion of congruent, neutral and incongruent trials was equal (33%) in each. The order of conditions as well as the order of the Stroop trials (144 per condition) were randomised across participants.

The experiment took place in a dimly lit room with the experimenter present and only one participant at a time. After providing their informed consent to the study, the participants engaged in a practice Stroop task (36 trials). The participants were instructed to lay their left middle finger on "V", left index finger on "B", right index finger on "N" and right middle finger on "M" while undertaking the Stroop task. The participants were instructed to focus at fixation cross and retain their focus the centre of the screen during the Stroop task. After 1500ms, the fixation cross was replaced by one of the Stroop words and remained on the screen for 2000ms. Finally, a feedback ("CORRECT" or

[9] Note that hypnotisability was measured as a continuous variable and we created groups using cut-off described in the Participants subsection of the Methods section.

"INCORRECT") flashed in black on the screen and then a new trial started with the fixation cross. The response to stimulus interval was 2000ms. This sequence remained constant across all conditions.

Next, the participants undertook the five experimental conditions. The order of the conditions was randomly generated for each participant. In the no strategy condition, the participants were asked to not use any of the mentioned strategies, and to respond as fast and as accurately as they could. All of the strategy conditions started with a screen explaining the strategy they are asked to use on each trial. For the visual strategies, an example word was presented so that the participants could practice the strategy (See the Appendix C for exact instructions). Before the start of the condition, the experimenter asked the participants whether they had understood how to use the strategy and provided clarification on request. After each strategy condition, the participants were asked to report the percentage of the trials on which they managed to use the strategy ("What do you think, on what percentage of the trials did you use the strategy? Please answer with a number between 0 and 100."). After finishing the last condition, the participants were thanked and debriefed.

Data analysis

Statistical analyses. We conducted all of our analyses with the statistical software R 3.3.1 (R Core Team, 2016). We calculated difference scores for the RTs so that we were able to directly test all of our hypotheses with Bayesian paired t-tests (comparing two conditions or testing whether a regression slope is different from zero) or Bayesian independent t-tests. Note that we did not run any omnibus tests (e.g., F test including all five conditions at a time) as it would not be informative in respect of hypotheses of the current project. We reported p-values for each statistical test, but we used the Bayes factor (B) to draw conclusions.

Bayes factor. We applied the R script of Dienes and McLatchie (2018) to calculate the Bayes factors. This calculator has a t-distribution as a likelihood function for the data as well as for the model of H1. We set the degrees of freedom of the model of H1 to 10,000 in each analysis to have a likelihood function for the theory following a normal distribution. To calculate the B, one also needs to specify the prediction of the two models (H1 and H0) under comparison. Every tested hypotheses had directional prediction, hence we applied a half normal distribution with a mode of zero to model the

predictions of H1. We specified the distribution as a half-normal since it is in line with the assumption that smaller effects are more probable than larger effects (Dienes, 2014). We report Bs as $B_{H(0, X)}$, in which H indicates that the model is half-normal, the first parameter (0) indicates the mode of the distribution and the second parameter (X) represents the SD of the distribution. We used various strategies to define the SDs of the different H1s.

Concerning the outcome neutral tests of the Stroop interference and the Stroop effects, we informed the SD of the model based on results of the baseline condition of a recent study of ours that used identical Stroop materials (Palfi, Parris, McLatchie, Kekecs, & Dienes, 2018). That is, the SD of the models were 60ms and 105ms, respectively. For the critical analysis, testing the efficiency of the strategies, we used 30ms, which is the half of the baseline Stroop interference. This value is based on the finding that the word blindness suggestion usually halves the baseline Stroop interference and we expect that a successful strategy should produce about the same effect size (Parris, Dienes & Hodgson, 2013). Incidentally, this value is exactly the same that we would obtain by using the room-to-move heuristic to define the maximum possible effect size, provided that the baseline Stroop interference is 60 ms (Dienes, 2019). The SD of the model predicting a positive relationship between hypnotisability and reduction in Stroop interference by strategy application was 5ms and it was based on the findings of Parris and Dienes (2013) who demonstrated a positive link between hypnotisability and the imaginative word blindness effect. In other words, H1 predicts that one unit increase on the SWASH aids the ability to reduce the Stroop interference using one of the strategies with about 5 ms.

In order to draw conclusions about the compared models, we used the convention of $B > 3$ to distinguish between insensitive and good enough evidence for the alternative hypotheses (Jeffreys, 1961). By symmetry, we used the cut-off of $B < 1/3$ to identify good enough evidence for the null hypothesis. To evaluate the robustness of our Bayesian conclusions to the SDs of the H1 models, we report a robustness region for each B, providing the range of SDs of the half-normal models that qualitatively support the same conclusion (using the threshold of 3 for moderate evidence for H1 and ⅓ for moderate evidence for H0) as the chosen SD[10]. The robustness regions are reported as: RR [x1, x2] where x1 is the smallest and x2 is the largest SD that gives the same conclusion.

[10] Thanks to Balazs Aczel for this suggestion

Pre-registration

The design and analysis plan of this experiment was pre-registered at https://osf.io/4z3xu. We closely followed the steps of the pre-registration when running the experiment and the analysis. Nonetheless, we added an analysis to the set of the crucial tests (Crucial test 1): the test of the efficiency of the strategies with all participants who had SWASH scores. This analysis is critical to demonstrate whether or not there is a main effect of successful strategy application irrespective of the participants` hypnotisability. We also included an auxiliary explorative analysis (Exploration 1). For the justification and interpretation of this analysis, see the discussion section of the first experiment. Finally, exploratory analyses assessing the strength of the relationships between self-reports of strategy usage and reduction in Stroop interference are reported in the Supplementary Materials.

Results

Data processing. We excluded the trials with errors from the analyses (8.2% in total from which 1.3% from the no strategy, 2.1% from the looking-away, 1.6% from the blurring, 1.7% from the single letter focus and 1.5% from the goal-maintenance conditions). Following the outlier exclusion criterion of Raz et al. (2002), we omitted trials with RTs that were 3 standard deviations either above or below the mean (1.2% of the correct trials from which 0.2% from the no strategy, 0.3% from the looking-away, 0.3% from the blurring, 0.2% from the single letter focus and 0.2% from the goal-maintenance conditions).

Outcome neutral checks 1 (non-preregistered): On what percentage of the trials did the participants use the strategies? The conditions in descending order based on the means of the reported percentages of strategy usage: goal-maintenance (M = 86%, 95% CI [82%, 90%]); looking-away (M = 83%, 95% CI [80%, 87%]); blurring (M = 73%, 95% CI [68%, 78%]); and single-letter focus conditions (M = 66%, 95% CI [61%, 71%]).

Outcome neutral tests 2: Is there a Stroop interference effect in the No strategy condition? As anticipated, the RTs in the no strategy condition were the fastest in the congruent trials followed by the neutral trials and then the incongruent trials (See Table 1 for condition means and SDs). The comparison of the incongruent and neutral trials yielded evidence for the Stroop interference effect ($t(56) = 7.74$, $p < .001$, M_{diff} =

78 ms, d_z = 1.03 , $B_{H(0, 60)}$ = 1.49*10^8, RR[3, 2.76*10^4]). The contrast of the incongruent and congruent trials revealed evidence in support of the Stroop effect (t(56) = 11.73, p < .001, M_{diff} = 126 ms, d_z = 1.55, $B_{H(0, 105)}$ = 2.23*10^{14}, RR[4, 4.62*10^4]).

Table 1

Summary Table about the Means of the RTs (ms) in the five Strategy Conditions

	Congruency type		
Strategy condition	Incongruent	Neutral	Congruent
No strategy	808 (127)	730 (101)	682 (94)
Looking-away	815 (94)	802 (94)	771 (97)
Blurring	821 (121)	776 (119)	739 (114)
Single letter focus	880 (157)	812 (133)	766 (130)
Goal-maintenance	804 (142)	726 (107)	689 (90)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Crucial test 1 (non-preregistered): Are the strategies effective in reducing the Stroop interference effect? Using the data of all the participants we tested whether any of the four strategies decreased Stroop interference. Comparing the no strategy and strategy conditions revealed evidence for the effectiveness of the looking-away (t(56) = 4.99, p < .001, M_{diff} = 65 ms, d_z = 0.66, $B_{H(0, 30)}$ = 3.93*10^3, RR[5, 2.05*10^4]) and the blurring (t(56) = 2.85, p = .006, M_{diff} = 33 ms, d_z = 0.38, $B_{H(0, 30)}$ = 20.05, RR[6, 365]) strategies. There was anecdotal evidence for no difference between no strategy and the single-letter focus (t(56) = 0.73, p = .469, M_{diff} = 9 ms, d_z = 0.10, $B_{H(0, 30)}$ = 0.73, RR[0, 74]), and between the no strategy and goal-maintenance strategies (t(56) = 0.01, p = .993, M_{diff} = 0 ms, d_z = 0.00, $B_{H(0, 30)}$ = 0.38, RR[0, 34]). The Bayes factor of the latter two tests did not reach the level of good enough evidence. See Figure 1 for the distribution of the Stroop interference scores broken down by the experimental conditions.

Interestingly, the mean RTs of incongruent trials in the looking-away and blurring conditions were numerically higher than that of the no strategy condition. We found evidence that neither of the looking-away (t(56) = 0.46, p = .647, M_{diff} = -7 ms, d_z = 0.06, $B_{H(0, 30)}$ =0.34, RR[0, 30]) nor the blurring strategies (t(56) = 0.86, p = .392, M_{diff} = -13 ms, d_z = 0.11, $B_{H(0, 30)}$ = 0.27, RR[23, ∞]) reduced the RTs of incongruent trials.

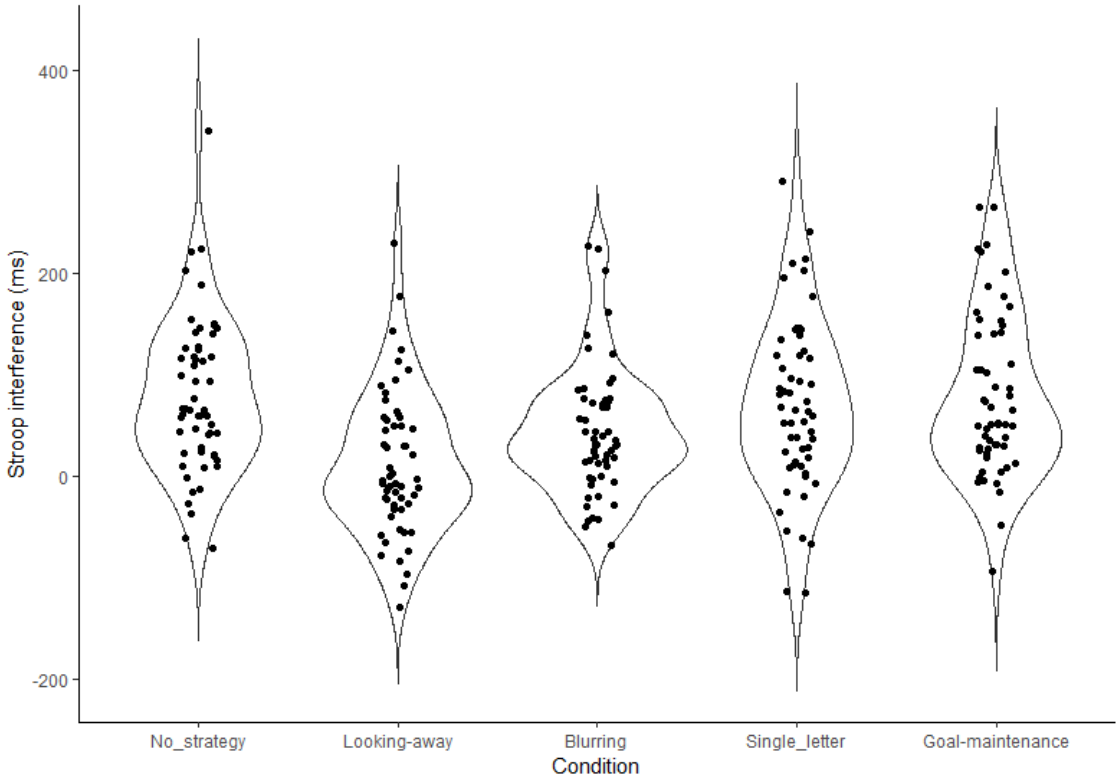

Figure 1. Violin plot depicting the distribution of Stroop interference scores (ms) in the no strategy and in the four strategy conditions. Each black dot represent the Stroop interference score of a single participant.

Crucial test 2: Can highs and lows employ the strategies to reduce the Stroop interference effect? Is there any difference between the groups in this aspect? Highs managed to reduce the Stroop interference by using the looking-away ($t(9) = 3.50$, $p = .007$, $M_{diff} = 104$ ms, $d_z = 1.11$, $B_{H(0, 30)} = 19.66$, RR[20, $1.49*10^3$]) and the blurring strategies ($t(9) = 2.22$, $p = .054$, $M_{diff} = 81$ ms, $d_z = 0.70$, $B_{H(0, 30)} = 2.97$, RR[0, 30]), although the latter test did not reach the conventional threshold of good enough evidence. We found data insensitivity regarding the effectiveness of the single-letter focus ($t(9) = 1.26$, $p = .238$, $M_{diff} = 45$ ms, $d_z = 0.40$, $B_{H(0, 30)} = 1.69$, RR[0, 431]) and the goal-maintenance strategies ($t(9) = 0.54$, $p = .605$, $M_{diff} = 20$ ms, $d_z = 0.17$, $B_{H(0, 30)} = 1.05$, RR[0, 182]). The pattern of results was similar for lows. Namely, there was evidence for the effect of the blurring strategy ($t(7) = 3.37$, $p = .012$, $M_{diff} = 70$ ms, $d_z = 1.19$, $B_{H(0, 30)} = 9.41$, RR[15, 669]) and the raw effect size of the looking-away strategy was comparable to that of the highs ($t(7) = 1.60$, $p = .154$, $M_{diff} = 75$ ms, $d_z = 0.56$, $B_{H(0, 30)} = 1.88$, RR[0, 929]). The level of evidence regarding the single-letter focus ($t(7) = 1.06$, $p = .324$, $M_{diff} = 36$ ms, $d_z = 0.38$, $B_{H(0, 30)} = 1.48$, RR[0, 307]) and goal-maintenance strategies ($t(7) =$

1.07, p = .322, M_{diff} = 23 ms, d_z = 0.38, $B_{H(0, 30)}$ = 1.40, RR[0, 202]) remained insensitive for the lows as well. Importantly, comparing lows and highs revealed data insensitivity in all four cases: looking-away (t(12.24) = 0.53, p = .605, M_{diff} = 29 ms, d_z = 0.25, $B_{H(0, 30)}$ = 1.09, RR[0, 261]), blurring (t(13.90) = 0.26, p = .797, M_{diff} = 11 ms, d_z = 0.12, $B_{H(0, 30)}$ = 0.92, RR[0, 149]), single-letter focus (t(15.91) = 0.19, p = .856, M_{diff} = 9 ms, d_z = 0.09, $B_{H(0, 30)}$ = 0.92, RR[0, 165]), and goal-maintenance (t(14.09) = - 0.08, p = .941, M_{diff} = -3 ms, d_z = 0.04, $B_{H(0, 30)}$ = 0.79, RR[0, 117]).[11]

Crucial test 3: Is there a relationship between hypnotisability and the extent to which people can reduce the Stroop interference by the tested strategies? To this aim, we regressed the SWASH scores on the extent of the reduction in the Stroop interference by the strategies and tested the regression slopes against zero. Even though the raw regression slopes are comparable to zero, we did not gain good enough evidence for the null in any case. The raw regression slopes in descending order: blurring (t(55) = 0.91, p = . 368, b = 1.74 ms/SWASH unit, β = 0.03, $B_{H(0, 5)}$ = 0.91, RR[0, 24]), single-letter focus (t(55) = 0.11, p = . 920, b = 0.79 ms/SWASH unit, β = 0.01, $B_{H(0, 5)}$ = 0.92, RR[0, 23]), looking-away (t(55) = 0.06, p = .950, b = 0.49 ms/SWASH unit, β = 0.01, $B_{H(0, 5)}$ = 0.86, RR[0, 23]) and goal-maintenance strategy (t(55) = -0.11, p = . 911, b = -0.81 ms/SWASH unit, β = -0.2, $B_{H(0, 5)}$ = 0.78, RR[0, 18]). Figure 2 depicts the scatterplots, regression slopes and their 95% Confidence Intervals for each strategy separately.

[11] Note that we used the Welch t-test for these four comparisons as the groups had non-equal samples.

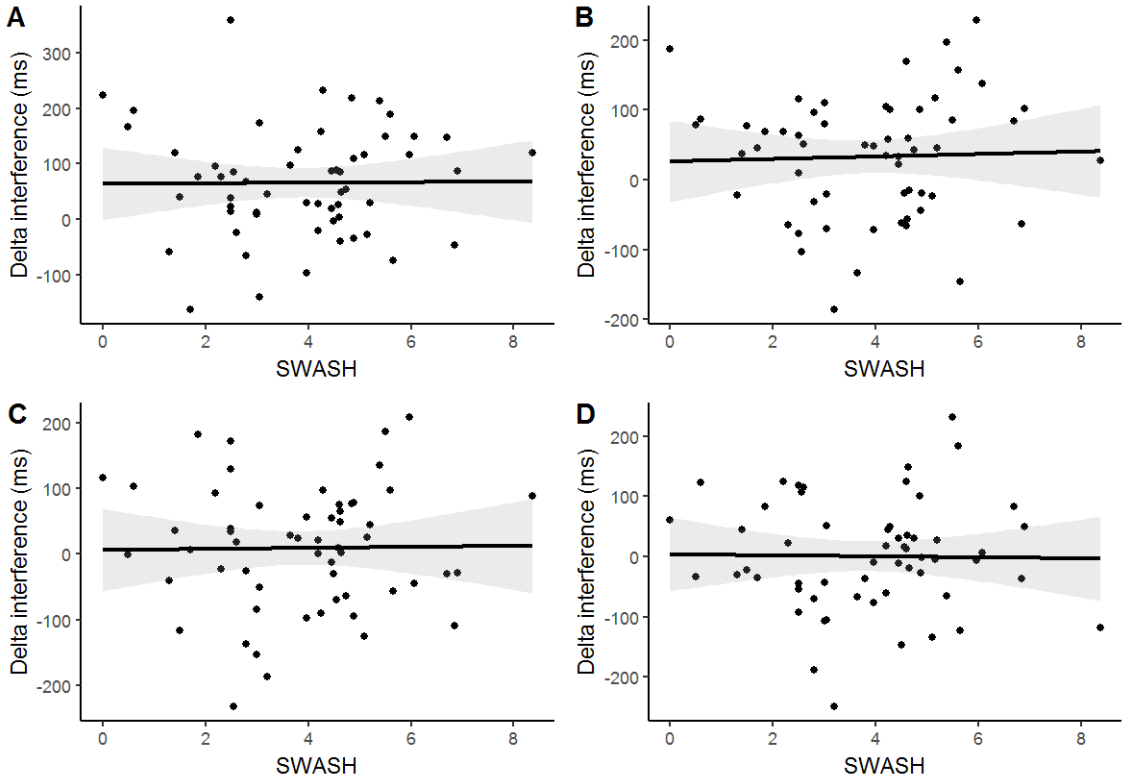

Figure 2. Scatterplots showing the relationship between hypnotisability (measured by the SWASH) and the reduction in the Stroop interference induced by the four strategies. The four panels indicate the Looking-away (Panel A), Blurring (Panel B), Single-letter focus (Panel C) and Goal-maintenance (Panel D) strategies.

Exploration (non-preregistered): What is the strength of the relationship between the extent of the Stroop interference and the general speed of the responses in each condition? One way strategies might reduce Stroop interference is simply by changing overall RTs. Pratte, Rouder, Morey, and Feng (2010) found that Stroop effects became smaller as RTs became smaller. Thus, a strategy may work by moving subjects along the regression line of Stroop interference (incongruent - neutral RTs) against RT (e.g. incongruent plus neutral RT). The estimate of the slope in the no strategy condition was moderately strong and the values in the 95% CI show a positive rather than negative relationship between general speed and the magnitude of the interference (b = 0.13ms, 95% CI = [0.04, 0.21], β = 0.36). Do the other strategies simply move along this line? The estimates in the looking-away (b = -0.003ms, 95% CI = [-0.11, 0.11], β = -0.01) and blurring strategy conditions (b = 0.01ms, 95% CI = [-0.07, 0.08], β = 0.03) were close to zero. Figure 2 depicts the three regression slopes separately with their 95% CIs. To compare the slopes, we conducted multilevel linear regression analyses estimating the

interaction of condition and general speed, while allowing different intercept for every subject. Contrasting the no strategy and looking-away conditions yielded a negative rather than positive slope and a fairly wide 95% CI nearly including zero (*b* = -0.13ms, 95% CI = [-0.26, -0.01]). The comparison of the no strategy and the blurring strategy yielded a negative slope as well (*b* = -0.11ms, 95% CI = [-0.22, -0.01]). Finally, the comparison of the two strategies revealed the slope to be reasonably close to zero (*b* = -0.01ms, 95% CI = [-0.14, 0.12]).

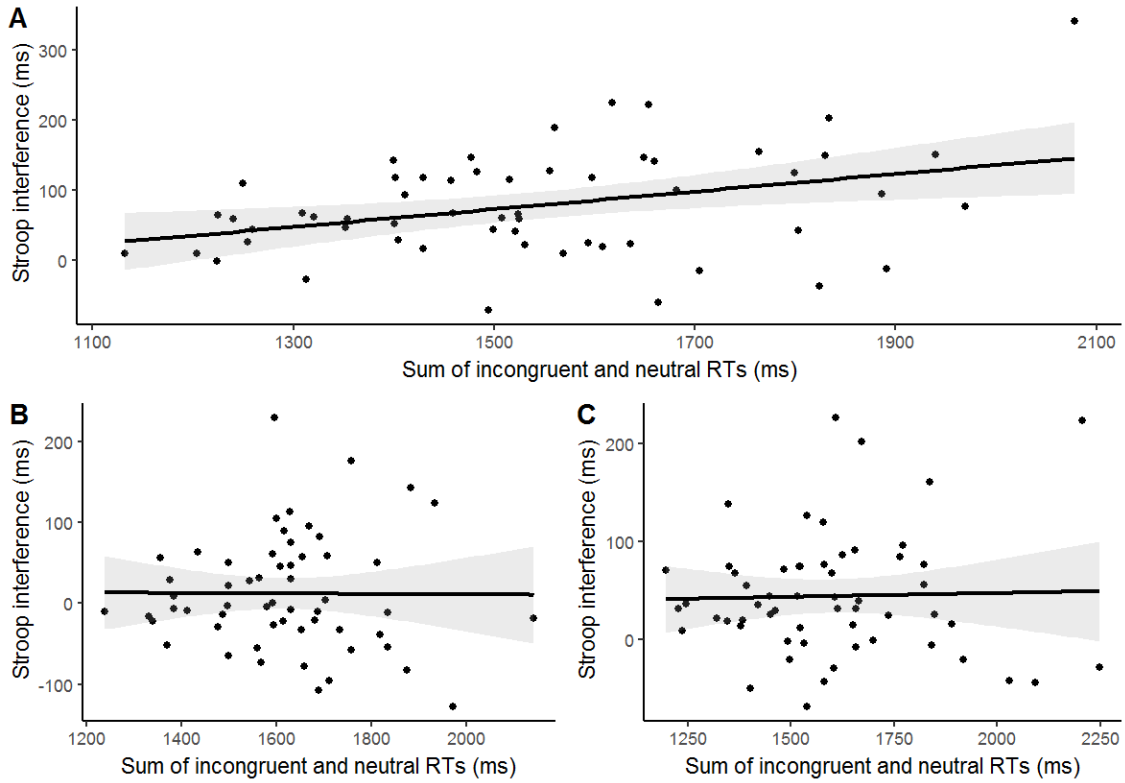

Figure 2. Scatterplot depicting the relationship between general response speed (i.e., sum of incongruent and neutral RTs) and the extent of the Stroop interference effect separately for the no strategy condition (Panel A) and the two conditions in which the strategies managed to reduce the interference. Looking-away (Panel B) and blurring strategies (Panel B).

Discussion

In this experiment, we tested four strategies that putatively reduce the Stroop interference effect to examine whether any of these strategies can be the underlying mechanism of the word blindness suggestion. The crucial test of the strategies provided insufficient evidence either way for whether the single-letter focus or the goal-maintenance strategies could mitigate the extent of the interference. On the other hand,

the looking-away and the visual blurring strategies passed the crucial tests and so there is evidence for them as plausible underlying mechanisms of the word blindness suggestion. These strategies aided the subjects to decrease the extent of the interference drastically for all levels of hypnotisability. Moreover, the blurring strategy approximately halved the extent of the Stroop interference (reduction of 33ms from the baseline of 78ms), which is precisely what the word blindness suggestion achieves in general (Parris, Dienes & Hodgson, 2013). However, as mentioned earlier, the word blindness effect has another distinctive feature: it realizes the reduction of the interference effect by reducing the RTs of incongruent trials (MacLeod, 2011; See Supplementary Materials Table S1 for a meta-analysis of studies demonstrating the word blindness effect). Surprisingly, our results do not match this pattern, there is evidence that neither of the strategies managed to decrease the RTs of the incongruent trials (for the looking-away strategy as the corresponding $B = 0.34$ was just above the conventional cut-off of $B < 1/3$). If this finding is robust, it challenges the idea that these strategies are the underlying mechanisms of the suggestion. Therefore, in the next experiment, we preregistered reduction in incongruent RTs as a test of the strategies.

Another key characteristic of the word blindness suggestion is that it seems to reduce interference by attenuating response competition and not by de-automatising reading per se (Augustinova & Ferrand, 2012; Palfi et al., 2018; Parris, Dienes & Hodgson, 2013; contrast Raz, Fan & Posner, 2005). Hence, the strategy that underlies the suggestion should not dampen the visual input of the meaning of the words; rather it should aid the subjects to handle response conflict between the competing response options. It is not clear, however, how looking-away or visual blurring would be in accordance with this notion. Nonetheless, the way with which the strategy takes its effect is testable by an experimental design in which the semantic and response conflict components of the interference effect are distinguishable. If the extent of the semantic conflict is identical in the no strategy and strategy conditions then one can conclude that the strategy leaves semantic processing and so reading itself untouched (for the original argument see Augustinova & Ferrand, 2012). Thus, we specify this outcome as a condition that needs to be met in the next experiment by the strategies to be deemed as plausible underlying mechanisms of the suggestion.

Finally, in many cases, the word blindness suggestion impacts the RTs of neutral trials as well, and surprisingly, it reduces them (e.g., Augustinova & Ferrand, 2012; Parris et

al., 2012; Raz et al., 2002). This feature of the suggestion is completely in harmony with a strategy that condenses the interference by simply speeding up all responses. To test this notion, one can, for instance, compare the conditions in terms of the patterns of the relationship between the general speed of responses and the magnitude of the interference effect (cf., De Jong, Liang, & Lauber, 1994; Pratte, Rouder, Morey, & Feng, 2010). However, data of the current experiment suggest that this relationship is weak in the looking-away and blurring conditions, and it is strong in the no strategy condition. Nonetheless, our analysis was not pre-registered and so the findings need to be replicated.

Experiment 2

In this experiment, we aim to test whether the beneficial effects of the looking-away and visual blurring strategies on the mitigation of Stroop interference can be replicated. Furthermore, we defined two conditions that the strategies ought to meet to be considered as appropriate underlying mechanisms of the word blindness suggestion: they need to reduce incongruent RTs and they should alleviate response conflict rather than semantic conflict. In order to test the latter assumption, we added non-response set incongruent trials to all of the experimental conditions. These trials consist of color words that are not part of the response set (e.g., brown) displayed in one of the colors of the response set. Therefore, responding to these types of trials should not involve response competition, and the non-response set interference (RT difference between non-response set incongruent and neutral trials) can be taken as an index of conflict that occurs during semantic processing (Klein, 1964; Sharma & McKenna, 1998).[12] Henceforth, we refer to the non-response set interference effect simply as semantic conflict or semantic interference effect. Finally, we repeated the exploratory analyses of the first experiment intending to test whether or not there is a relationship between general speed of responses and the extent of the Stroop interference in any of the conditions. Moreover, we were interested whether or not the slopes in the strategy conditions are qualitatively identical to the one in the no strategy condition. We defined these analyses merely as supporting

[12] To distinguish between the semantic and response conflict components of the Stroop interference effect, one can also use color-associated words (e.g., sky) that tend to produce longer RTs than neutral words but shorter RTs than response set incongruent trials (Klein, 1964). For instance, Augustinova and Ferrand (2012) applied color-associated words in their experiments to assess the magnitude of semantic conflict and to present evidence that the word blindness suggestion influences solely the response conflict component of the interference effect. Nonetheless, their experiments employed vocal responses and when it comes to manual responses, the color-associated interference effect is volatile (Sharma & McKenna, 1998; Kinoshita, Mills, & Norris, 2018).

tests of interest as it is not an established feature of the word blindness suggestion that it produces an identical slope as a control condition.

Methods

Participants. We recruited 35 participants, however, one of the participants claimed that they did not follow the instructions closely and used visual blurring in the no strategy condition. We excluded the data of that participant and all analyses were run on the data of 34 participants (mean age = 21.82, SD = 4.38, females = 27). The participants received either course credits or payment (£5) in exchange for attending the study.

Stimuli and apparatus. The materials of the registered experiment closely followed those of the first experiment. We added four color words to the stimulus set (BROWN, PINK, GREY, ORANGE) so that either these four or the original four color words could be used in the non-response set incongruent trials. We ran the experiment in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) and the resolution of the computer screen was 1920x1080.

Design and procedure. There were three major changes in this experiment: we did not include the single-letter and goal-maintenance strategy conditions; there were more trials in each condition as we included non-response set trials as well; we did not take into account the hypnotisability of the participants. The experiment had a 4x3x2 mixed design with congruency type (congruent, neutral, incongruent non-response set, incongruent response set) and strategy condition (no strategy, looking-away, visual blurring), and non-response set groups (response set being equivalent [A] vs not equivalent to the first experiment [B]) as independent variables. The participants were assigned to response set groups A or B based on the parity of their subject number. Group membership determined whether the colors of A or B would have corresponding response buttons. For instance, if someone was assigned to group B, then the colors brown, pink, grey and orange had the corresponding response buttons of “V”,”B”,”N” and “M”, respectively. In this case, none of the words were displayed in red, blue, green or yellow. Apart from this, the procedure of the experiment was identical to that of the first experiment.

Data analysis

The steps of the data analysis are in line with those of the first experiment, including the exclusion criterion regarding RT data and how we drew conclusions based on the results of the Bayes factors. We informed the parameters of the model predicting the presence of the semantic interference effect based on the findings of Augustinova and Ferrand (2012), who found in two experiments that the size of the semantic interference (using color-associated words) was about 20ms. We expect that an intervention impacting semantic processing should approximately halve this effect. For the test of the regressions slopes investigating the relationship of general response speed and the extent of the Stroop effect, the model parameters of H1 were stemmed from the finding that the slope was 0.13 ms in the no strategy condition in the first experiment. We used this value as the SD of H1 for the tests of the slopes against zero as well as for their comparisons.

Pre-registration

The design and analysis plan of the experiment was pre-registered and they can be accessed at https://osf.io/gbsaf. We closely followed the steps of the design and of the analysis plan. There was one deviation in terms of the implementation of the comparison of the regression slopes. We pre-registered that we will compare regression slopes by calculating difference scores of the dependent variable and running the regression analysis on these difference scores. However, this implementation is incorrect and does not produce the intended outcome of the comparison of the slopes. Therefore, we ran multilevel linear regressions that allow the comparison of slopes of dependent data by the test of the interaction.

Results

Data processing. First, we omitted trials with errors from the analyses (10.4% in total from which 2.3% from the no strategy, 4.4% from the looking-away, 3.7% from the blurring conditions). Next, we eliminated trials with RTs that were 3 standard deviations either above or below the mean (1.2% of all correct trials from which 0.5% from the no strategy, 0.4% from the looking-away, 0.3% from the blurring conditions).

Outcome neutral checks 1 (non-preregistered): On what percentage of the trials did the participants use the strategies? The participants reported that, on average, they used on 80% (95% CI [75%, 85%]) of the trials the looking-away strategy, and on 73% (95% CI [66%, 81%]) of the trials the blurring strategy.

Outcome neutral tests 2: Is there a difference between the two response set groups regarding the magnitude of the Stroop interference and the sematic Stroop effect (in the No strategy condition)? Before collapsing the data across response set groups, we compared the two groups in terms of the extent of the Stroop interference and semantic Stroop effects. The size of the Stroop interference effect was comparable in the two response set groups (M_A = 78 ms, M_B = 79 ms) and there is some evidence in favour of the model predicting no difference (t(30.66) = -0.05, p = .958, M_{diff} = 1 ms, d_z =, $B_{N(0, 60)}$ =0.38, RR[0, 69]), however the strength of evidence did not reach the conventional cut-off of good enough evidence. The size of the semantic Stroop effect was numerically larger in the group with the response set of the first experiment (M_A = 49 ms, M_B = 15 ms), however the analysis yielded data insensitivity (t(29.46) = 1.27, p = .212, M_{diff} = 35 ms, d_z =, $B_{N(0, 20)}$ = 1.08, RR[0, 179]). Consequently, we decided to conduct all of the subsequent analyses on the collapsed data.

Outcome neutral tests 3: Is there a Stroop interference and a semantic Stroop effect in the No strategy condition? As in the first experiment, the RTs in the no strategy condition were the fastest in the congruent trials followed by the neutral trials. The RTs of the non-response set incongruent trials were slower than those of the neutral trials, and the longest RTs were observed in the incongruent trials (See Table 2 for condition means and SDs). The analyses revealed strong evidence for Stroop interference (t(33) = 6.56, p < .001, M_{diff} = 79 ms, d_z = 1.12 , $B_{H(0, 60)}$ = $2.48*10^5$, RR[5, $2.6*10^4$]) as well as for the Stroop effect (t(33) = 10.16, p < .001, M_{diff} = 130 ms, d_z = 1.74, $B_{H(0, 105)}$ = $3.36*10^9$, RR[6, $4.57*10^4$]). Moreover, the contrast of the non-response set incongruent and the neutral trials yielded evidence for the semantic Stroop interference effect (t(33) = 2.53, p = .016, M_{diff} = 34 ms, d_z = 0.43 , $B_{H(0, 20)}$ = 8.29, RR[8, 177]).

Table 2

Summary Table about the Means of the RTs (ms) in the three Strategy Conditions

Strategy condition	Congruency			
	Incongruent	Incongruent non-response set	Neutral	Congruent
No strategy	791 (131)	746 (112)	712 (97)	661 (81)
Looking-away	838 (126)	822 (126)	830 (127)	790 (118)
Blurring	822 (130)	812 (130)	786 (128)	737 (119)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Crucial test 1: Are the strategies effective in reducing the Stroop interference effect? First, we examined whether or not the beneficial effect of the looking-away and blurring strategies replicated in the current experiment. We found strong evidence that both of the looking-away ($t(33) = 4.42$, $p < .001$, $M_{diff} = 71$ ms, $d_z = 0.76$, $B_{H(0, 30)} = 297.77$, RR[7, $1.93*10^4$]) and the blurring strategies ($t(33) = 3.05$, $p = .005$, $M_{diff} = 43$ ms, $d_z = 0.52$, $B_{H(0, 30)} = 24.93$, RR[7, 632]) helped the participants to reduce the Stroop interference compared to the no strategy condition. Figure 1 depicts the distribution of the Stroop interference scores broken down by the strategy conditions.

As an additional analysis, we tested whether the strategies reduced the response conflict component (incongruent RTs – non-response set RTs) of the Stroop interference effect so that our results can be compared to those of Augustinova and Ferrand (2012). The analyses revealed moderate evidence supporting that the blurring strategy reduced response conflict ($t(33) = 1.98$, $p = .056$, $M_{diff} = 34$ ms, $d_z = 0.34$, $B_{H(0, 30)} = 3.94$, RR[16, 64]) and anecdotal evidence that looking-away strategy reduced response conflict ($t(33) = 1.61$, $p = .117$, $M_{diff} = 29$ ms, $d_z = 0.27$, $B_{H(0, 30)} = 2.40$, RR[0, 364]) compared to the no strategy condition. Note these two latter tests were not pre-registered.

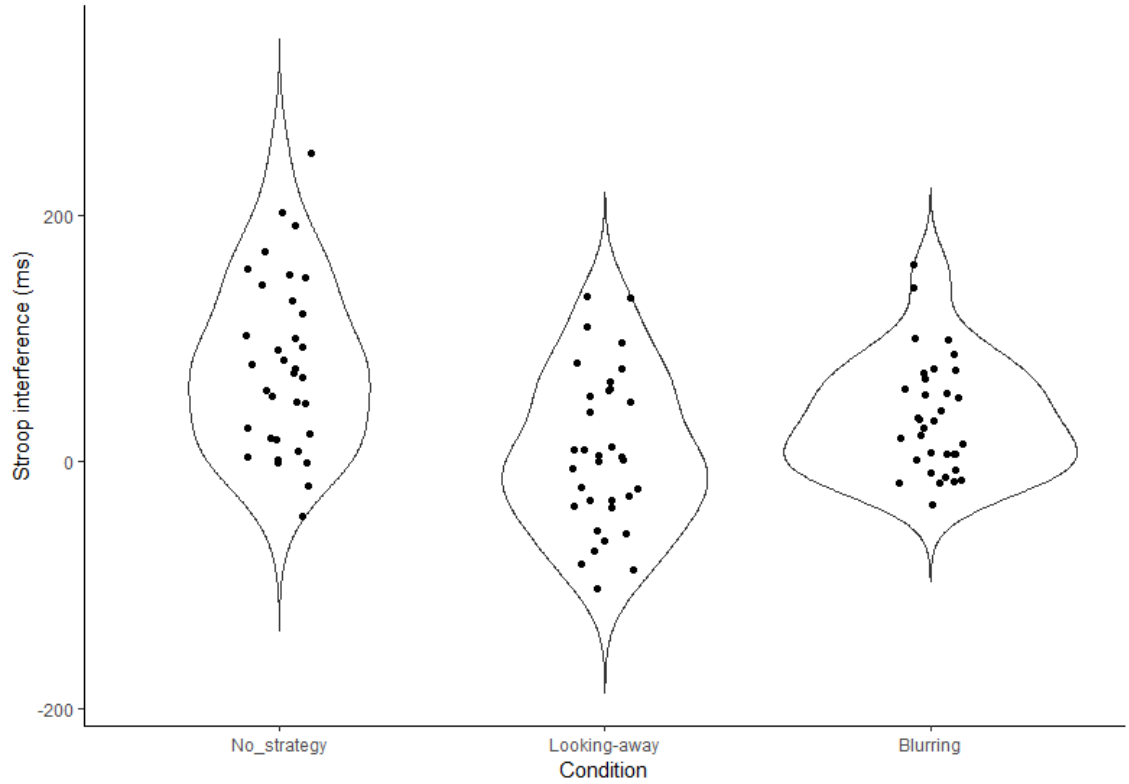

Figure 4. Violin plot portraying the distribution of Stroop interference scores (ms) in the no strategy and in the two strategy conditions. Each black dot represent the Stroop interference score of a single participant.

Crucial test 2: Do the strategies diminish the RTs of the incongruent trials? We found moderate evidence supporting the claim that neither the looking-away ($t(33) = -2.35$, $p = .025$, $M_{diff} = -47$ ms, $d_z = -0.40$, $B_{H(0, 30)} = 0.22$, RR[19, ∞]) nor the blurring strategy ($t(33) = -1.99$, $p = .055$, $M_{diff} = -31$ ms, $d_z = -0.34$, $B_{H(0, 30)} = 0.19$, RR[16, ∞]) reduced the incongruent RTs compared to the no strategy condition.

Crucial test 3: Do the strategies influence the magnitude of the semantic Stroop interference effect? There was anecdotal evidence that the looking-away strategy reduced the semantic Stroop interference effect ($t(33) = 2.41$, $p = .022$, $M_{diff} = 42$ ms, $d_z = 0.41$, $B_{H(0, 10)} = 2.80$, RR[0, 11]). In fact, the strategy eliminated the semantic Stroop effect in the looking-away strategy condition ($t(33) = -1.06$, $p = .296$, $M_{diff} = -8$ ms, $d_z = -0.18$, $B_{H(0, 20)} = 0.20$, RR[12, ∞]). In case of the blurring strategy, there was no evidence either way for whether or not semantic Stroop interference was reduced ($t(33) = 0.50$, $p = .617$, $M_{diff} = 8$ ms, $d_z = 0.09$, $B_{H(0, 10)} = 1.07$, RR[0, 74]).

Supporting test of interest: Is the relationship between the extent of Stroop interference and the general speed of the responses different from zero in each

condition? We found strong evidence in the no strategy condition that there was a positive relationship between general speed and the magnitude of the interference ($t(33) = 3.33$, $p = .002$, $b = 0.16$ms, $\beta = 0.51$, $B_{H(0, 0.13)} = 49.05$, RR[0.03, 4.46]). As expected based on the results of the first experiment, there was good enough evidence supporting the notion that there is no relationship between general speed of the responses and the extent of the interference both in the looking-away ($t(33) = -0.10$, $p = .920$, $b = -0.004$ms, $\beta = -0.02$, $B_{H(0, 0.13)} = 0.31$, RR[0.12, ∞]) and blurring strategy conditions ($t(33) = 0.27$, $p = .792$, $b = 0.01$ms, $\beta = 0.05$, $B_{H(0, 0.13)} = 0.30$, RR[0.12, ∞]). Inspecting the raw effect sizes reveals that both of the slopes were virtually zero (Figure 2 depicts the three regression slopes separately with their 95% CIs.). The comparison of the no strategy and looking-away conditions yielded strong evidence supporting their difference ($t = 2.71$, $b = 0.15$ms, $B_{H(0, 0.13)} = 13.50$, RR[0.04, 0.80]). Contrasting the no strategy and the blurring strategy revealed good enough evidence for their difference ($t = 2.49$, $b = 0.17$ms, $B_{H(0, 0.13)} = 8.81$, RR[0.04, 1.14]). Finally, we found anecdotal evidence that the slopes in the looking-away and blurring conditions were identical ($t = 0.14$, $b = 0.01$ms, $B_{N(0, 0.13)} = 0.39$, RR[0, 0.15])[13].

[13] Note that for this comparison, we modelled the predictions of H1 with a normal distribution (rather than a half-normal) as there was no directional hypothesis. Hence the "N" instead of "H" in the subscript of B.

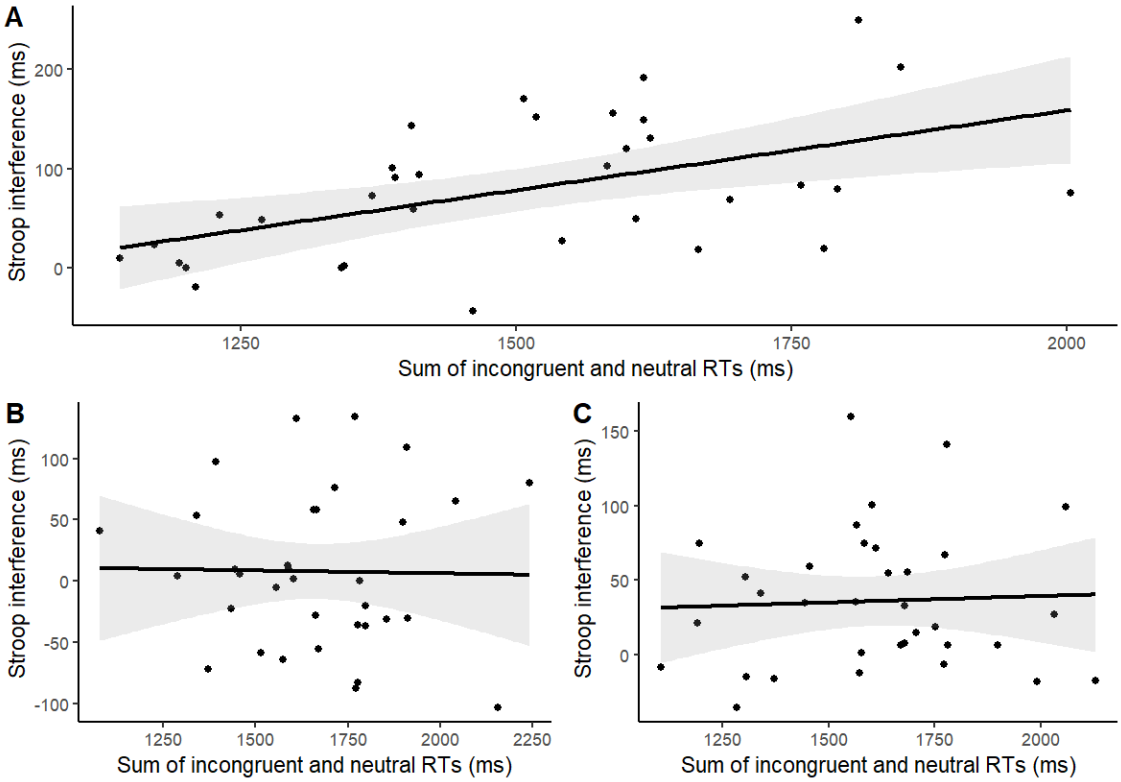

Figure 5. Scatterplot depicting the relationship between general response speed (i.e., sum of incongruent and neutral RTs) and the extent of the Stroop interference effect separately for the no strategy condition (Panel A) and the looking-away (Panel B) and blurring (Panel B) strategy conditions.

Discussion

Once more, both of the looking-away and blurring strategies were demonstrated to be useful in reducing Stroop interference, and the blurring strategy approximately halved the Stroop interference effect as the word blindness suggestion tends to do when it is given to highly suggestible people. We also replicated the finding that neither of the strategies can speed up responses on the incongruent trials. By introducing non-response set incongruent trials, we were able to distinguish the semantic and response conflict component of the interference effect, and we found that some evidence that the looking-away strategy works by alleviating both sources of conflicts, whereas for the blurring effect, the evidence is not clear whether it solely reduces response conflict or it may diminish semantic conflict as well. Importantly, we specified these two latter analyses as severe tests that can disconfirm the idea that looking-away or blurring could be the underlying mechanism of the word blindness effect. Consequently, we ought to conclude that none of the strategies have met the criteria and are unlikely to be the strategies that

highs resort to when they respond to the word blindness suggestion. Finally, the differential relationship between the general speed of the responses and the size of the interference effect was replicated as well. We found evidence for a positive relationship in the no strategy condition, and evidence for no relationship in the looking-away and blurring conditions, as well as evidence for the difference between no strategy and the strategy conditions.

General Discussion

The purpose of the project was to investigate whether or not simple strategies, such as blurring one`s vision can attenuate the Stroop interference effect. According to cold-control theory of hypnotic responding (Dienes & Perner, 2007), people use strategies to create the experience that was described to them in the suggestion. Hence, identifying successful strategies is crucial to understand how highs can manage to reduce the interference effect when they respond to the word blindness suggestion, and by this we can unravel the mystery surrounding the suggestion. Importantly, the ability of highs to respond hypnotically (with the feeling of involuntariness) seems to be independent of their first-order executive functions, such as cognitive inhibition (Dienes et al., 2009) and selective attention (Varga, Németh, & Szekely, 2011), that could help them overcome cognitive conflict during the Stroop task (see Parris, 2017, for a review). We found no evidence one way or the other for a correlation between hypnotisability and the extent to which any of the strategies could decrease Stroop interference. Reaching good enough evidence for the null usually demands many participants in correlational designs. For instance, assuming that the estimates of the slopes are zero, the standard deviation of hypnotisability and of the reduction in Stroop interference by the given strategy remains identical to those of that we observed in the first experiment (and the correlations of these variables are zero), we would need to recruit 825 and 1052 participants for the blurring and looking-away conditions respectively to have a 50% probability of obtaining Bs smaller than 1/3.

Next, we probed the efficiency of the four strategies: looking-away, visual blurring, single letter focus, and goal-maintenance. Importantly, looking-away and blurring strategies were shown to be useful in diminishing the interference effect in both of the experiments substantiating that participants are able to reduce the Stroop interference by consciously engaging in simple strategies, a finding that has been rarely

demonstrated in the Stroop literature (cf. Raz et al., 2003). Nonetheless, none of these strategies should be considered as likely candidates of being the underlying mechanism of the word blindness suggestion, as they did not meet other criteria such as reducing the RT of incongruent trials.

The idea that goal-maintenance plays a crucial role in responding quickly and accurately to a Stroop word is well established (e.g., De Jong et al., 1999; Kane & Engle, 2003) and it is embedded in many of the cognitive control models (e.g., Cohen, Dunbar, & McClelland; Roelofs, 2003). It is important to note that our findings do not challenge this idea. In this project, we solely aimed to test whether a simple way to update one`s goal (i.e., rehearsal of the target) is sufficient to improve performance in the Stroop task. We did not provide good enough evidence one way or the other for whether highs achieve the reduction of the Stroop interference, when they respond to the word blindness suggestion, by the internal rehearsal of the task instructions. However, it is still possible that the strategy with which highs reduce Stroop interference facilitates goal-maintenance. In fact, based on the finding that the word blindness suggestion operates better when the response-stimulus interval is short (500 ms) than when it is long (3500 ms), we can assume that the strategy that highs employ influences processes related to goal-maintenance (Parris, Dienes, & Hodgson, 2012).

We found in both of the experiments that when people applied the blurring and looking-away strategies, then the relationship between general speed and interference disappeared. These results deem it unlikely that either the blurring or the looking-away strategy would work by speeding up responses and condensing the Stroop interference effect simply by moving along a fixed interference-overall RT slope. Another possibility is that moving along the interference-overall RT slope is a strategy in itself. For example, a simple model of motivation is that it moves people along this slope, speeding up overall RT and hence reducing Stroop interference (cf. Pratte et al., 2010). Indeed, enhanced motivation has most commonly lead to an overall speeding up of responses (Krebs et al., 2010; Locke & Braver, 2008; Veling & Aarts, 2010). Nonetheless, the introduction of a reward has not often produced large reductions in Stroop effects (Krebs et al., 2010; Veling & Aarts, 2010). More promising, setting up competition for reward in the presence of a competitive other has been shown to result in a greater than 50% reduction in Stroop

interference (Huguet, Dumas & Monteil, 2004)[14]. One might argue that the hypnotic context provides stronger motivation for highs than monetary reward by itself or combined with competition. However, the re-analysis of an earlier study of ours that had an identical design to the current experiment in terms of the Stroop test, but used the word blindness suggestion, revealed evidence for a raw slope of zero (B = 0.19) between Stroop interference and overall RT (sum of RTs of incongruent and neutral trials) in the suggestion condition (Pilot study of Palfi, Parris, Mclatchie, Kekecs, & Dienes, 2018). That is, it does not appear that in the suggestion condition people simply move along a fixed slope, generally speeding up and thereby reducing interference. Instead, people typically reduce the RT in especially the incongruent condition when responding to the suggestion. A proper understanding of the relation of motivation to the word blindness suggestion remains to be explored.

One strategy still remains that was not tested in the current experiment. When highs are suggested to see meaningless words throughout the Stroop task, perhaps, they take the instructions literally, and they create the experience of meaninglessness by imagining a counterfactual world in which words are truly meaningless. Imagining a counterfactual world might influence top-down cognitive control processes in a way that helps subjects reduce Stroop interference. There are two reasons why this notion is plausible. First, imagination can have an impact on behaviour as well as on cognitive processes. For instance, mental practice can improve one`s performance in golf (Ploszay, Gentner, Skinner, & Wrisberg, 2006). Imagination can advance self-regulation (Taylor, Pham, Rivkin, & Armor, 1998), confirm or in some cases challenge and mitigate prejudice (Slusher & Anderson, 1987), create false autobiographical memories (Mazzoni & Memon, 2003), and, finally, even enhance performance of visual search (Davoli, & Abrams, 2009; Witt & Proffitt, 2008). Second, cognitive penetrability is not completely unprecedented in the Stroop task. For instance, expectations modulated by placebo-suggestion were shown to influence performance, measured by accuracy (Magalhães De Saldanha da Gama, Slama, Caspar, Gevers, & Cleeremans, 2013), though such placebo Stroop reduction does not appear to match the word blindness suggestion in reducing Stroop interference in RTs (contrast response expectancy theory, Kirsch, 1985).

[14] But note in the Huguet et al. study the baseline level of interference (and reaction times) were unusually large, resulting in reduced manual response Stroop interference values still greater than 70ms, considerably larger than in the typical word blindness suggestion (about 35 ms).

Depending on the instructions of the placebo-suggestion, it can either enhance or impair the accuracy of the responses. There is, however, evidence from independent labs that a prime to deteriorate one`s reading abilities by imagining what is it like to have dyslexia, can help people reduce the Stroop interference effect compared to a baseline condition with a neutral prime that has no reference to reading (Augustinova & Ferrand, 2014b; Goldfarb, Aisenberg, & Henik, 2011).

Interestingly, the dyslexia prime and word blindness suggestion phenomena share many properties. They both substantially decrease the interference effect by speeding up the RT of incongruent trials compared to no suggestion / no prime baseline conditions when the response mode is manual (See Experiment 1 of Augustinova & Ferrand, 2014b; and Experiment 1 of Goldfarb et al., 2011). The dyslexia prime, similarly to the word blindness suggestion, affects the response competition component of the interference while it leaves the semantic conflict component untouched (Augustinova & Ferrand, 2012, 2014). This latter feature of the dyslexia prime is particularly important in challenging the initially proposed mechanism, namely the de-automatisation of reading account that putatively underlies these phenomena. An even more remarkable similarity between the instructions of the dyslexia prime and the word blindness suggestion experiments is that both invite the participants to think about disrupting one`s reading abilities. One could further this line of thought and propose that both of these effects are achieved via deliberate strategy engagement, specifically the imagination of a counterfactual world in which words are meaningless. Theories of social priming argue that responses to primes are unintentional and purely triggered by the activation of a specific social concept (Bargh, Chen, & Burrows, 1996; Dijksterhuis, & van Knippenberg, 1998). However, there are many reasons to be sceptical about the unintentional nature of the responses to social primes, such as the presence of clear demand characteristics or the absence of valid and reliable outcome neutral tests demonstrating that the participants were not aware of the link between the social prime and the dependent variable of the experiment (Doyen, Klein, Pichon, & Cleeremans, 2012; Newell & Shanks, 2014; Shanks et al., 2013). These criticisms apply to the dyslexia studies as well, deeming it plausible that the participants reduced the Stroop interference via intentional strategy usage rather than via the unintentional or automatic activation of the concept of dyslexia.

Nonetheless, the idea that imagining that one is unable to take meaning from the Stroop words facilitates the resolution of response competition is a conjecture that needs to be tested. Currently, a Registered Report is undertaking such a test by requesting highs to voluntarily imagine the words during the Stroop task as meaningless characters so that they can reduce the Stroop interference compared to a baseline condition in which they are asked to not engage in imagery strategies (Palfi et al., 2018). Provided that highs can decrease the Stroop interference by voluntarily imagining that the words are meaningless, it would still need to be explored whether everyone can use this strategy to alleviate the interference, and whether the power of imagination could be generalised to other cognitive tasks, such as the flanker (Eriksen & Eriksen, 1974) or Simon task (Simon & Wolf, 1963).

In sum, reducing interference in the Stroop task via intentional means is difficult and the current study provided compelling evidence that there are at least two strategies, looking-away from the target word and visual blurring, that any subject can apply. Sadly, none of these strategies met the criteria to be considered as potential underlying mechanisms of the word blindness suggestion deferring the resolution of the question of how exactly highly suggestible people diminish the interference when they respond to the suggestion. Although these findings further the mystery surrounding the word blindness suggestion, we hypothesize that imagination (i.e., imagining that the Stroop words are meaningless) may be the key strategy with which subjects reset top-down cognitive processing to comply with the request of the suggestion, and lead to the reduction of the Stroop interference.

Chapter IV: Can unconscious intentions be more effective than conscious intentions? Test of the role of metacognition in hypnotic response

Introduction

The cornerstone of hypnotic responding is the feeling of involuntariness that accompanies an otherwise goal-directed behaviour (Weitzenhoffer, 1974, 1980; Terhune, Cleeremans, Raz & Lynn, 2017). Responses to hypnotic suggestions vary widely in terms of their difficulty. Some motor actions can be done by almost everyone (e.g., feeling a magnetic power between the palms that is pulling them towards each other), whereas the imaginative exercise to produce vivid hallucinations of noises can only be performed by a minority of the population (usually highly suggestible people, henceforth highs). The question of how these alterations in cognition can be implemented with a disrupted sense of agency has been the focus of scientific endeavour for decades. Theories from the sociocognitive tradition of hypnosis stress the role of demand characteristics in forming the subjective experiences involved with hypnotic responding and often highlight the strategic nature of the action as appropriate to the specific context (Comey & Kirsch, 1999; Spanos, 1986). The cognitive approach also often underscores the active role of the participants in creating an altered sense of reality (e.g., Kihlstrom, 1998) and several theories of this tradition, the dissociation theories (Bowers, 1990; Hilgard, 1977, 1991; Kihlstrom, 1985), propose that the sense of involuntariness can emerge by dampening the monitoring of one's own control processes. Recently, a theoretical attempt has been made to synthesize these views by depicting a metacognitive account of hypnosis, namely the cold control theory (Barnier, Dienes & Mitchell, 2008; Dienes, 2012; Dienes & Perner, 2007), that draws from the higher-order thought theories (Lau & Rosenthal, 2011; Rosenthal, 2005) of consciousness.

The (higher-order thought) HOT theory of consciousness postulates that a representation or a state is only conscious if one becomes aware of its content by the existence of a "higher-order thought" or state (HOT) that refers to it (Rosenthal, 2005). A higher order mental state is a mental state not just about the world (which is first order) but about a mental state. For instance, imagine having a first-order state about the world (e.g., "there is a tree"). According to HOT theories, one has no conscious experience of the world unless one possesses in addition a second-order thought about that first-order

state (e.g., “I see that there is a tree”). HOTs are not restricted to perception, thus they can refer to any mental state, including those with control functions.

Cold control theory stresses that the mechanism by which hypnotic responding (behavior accompanied by the feeling of involuntariness) emerges is a process that allows people to replace the HOTs about their intentions with inaccurate ones (Dienes, 2012). For instance, to create the experience of the hallucination of a noise (e.g., the buzz of a mosquito), one has to have an intention about imagining that particular noise (e.g., an intention with the content “imagine a buzzing mosquito”), while forming an inaccurate HOT about the intention (e.g. with the content, “I`m not intending to imagine a buzzing mosquito”). The theory also claims that the intention to produce the appropriate response is formed by the actor and it is implemented by regular cognitive control processes implying that the behaviour will be in accord with the goals of the actor and that a hypnotic response cannot be more efficient than a non-hypnotic one (if a simple first-order intention is logically sufficient to produce the response). In a nutshell, cold control theory posits that hypnosis is solely a metacognitive phenomenon, and, in the simplest version of cold control, the theory assumes that hypnosis is targeting specifically the HOT of intending thereby leaves the first-order states untouched.[15] The latter assumption coincides with the idea that HOTs, or in other words, being conscious of mental states, have only limited or zero function in terms of influencing first-order states (Lau & Rosenthal, 2011; Rosenthal, 2008)[16]. In addition, this assumption also implies that the theory deems hypnotic responding as a form of unconscious control as it claims that responses to suggestions are intended (implemented by executive control processes) while the intention to act is unconscious by virtue of possessing an inaccurate HOT that the intention does not exist.

The notion that hypnotic responses are produced by a strategically relinquished metacognition of one's intentions has gained some support. For instance, individual differences in metacognition, particularly the tendency to generate inaccurate HOTs of intending, are moderately associated with hypnotisability (Dienes, 2012). Further,

[15] Cold control theory assumes that hypnotic response involves strategic changes in HOTs about solely intentions. This claim is independent of whether or not there is domain specificity in metacognitive abilities (e.g. Fleming, Ryu, Golfinos & Blackmon, 2014), or whether hypnotisability involves alterations in metacognition over other domains as well, such as perception.

[16] Of note, a special case of HOT theories, the cross-order integration theory (COI; Kriegel, 2007), stresses that first-order states and HOTs can causally influence each other by binding together to a unified conscious representation, which can, for instance, enhance cognitive functioning.

experimental evidence suggests that the temporary disruption of the dorsolateral prefrontal cortex (DLPFC), which has a vital role in the functioning of metacognition (Lau & Passingham, 2006; Rounis, Maniscalco, Rothwell, Passingham & Lau, 2010), with rTMS (Coltheart et al, 2018; Dienes & Hutton, 2013) or alcohol (Semmens-Wheeler, Dienes & Duka, 2013) facilitates hypnotic responding[17]. Another line of research also corroborates the idea that hypnotic responding is the product of a purely metacognitive process by revealing that behaviors created by hypnotic suggestions are not related to first-order abilities of cognitive functioning (apart from metacognition). For example, several studies have presented evidence that performance on tasks involving first-order abilities of executive functioning such as inhibition (Dienes et al., 2009) or sustained attention (Jamieson & Sheehan, 2002) do not predict hypnotisability. Moreover, evidence counts against the claim that responses to hypnotic suggestion can enhance first-order abilities compared to responses that are non-hypnotic. For instance, there is no evidence for the superiority of hypnotic suggestions in recollection (Erdelyi, 1994; Nogrady, McConkey, & Perry, 1985), (more controversially) analgesia (Milling, Kirsch, Meunier & Levine, 2002; Spanos, 1986; for a counter-argument: e.g. Derbyshire, Whalley & Oakley, 2009; Hilgard, 1977; Miller & Bowers, 1993) and endurance (Barber, 1966, Levitt & Brady, 1964). However, an experimental finding, the word blindness effect (Raz, Shapiro, Fan & Posner, 2002; the term was first used by Parris, Dienes & Hodgson, 2012), calls into question the key statement of the theory, as it suggests that highs can acquire abilities through hypnosis that they do not possess when responding non-hypnotically.

The word blindness phenomenon can be induced by suggesting to highs that they will see words as meaningless characters, or as words of a foreign language, while they are engaged in a color naming Stroop task (Stroop, 1935). Generally, the suggestion is applied post-hypnotically, which means that it is provided during a hypnotic induction prior to the Stroop task and only later activated by a clue (e.g., a clap). It has been shown by various independent laboratories that when this suggestion is given to highs, they can

[17] It is to be noted that none of the experimental manipulations were exclusive in a sense that they might impaired cognitive functions aside from metacognition (Dienes, 2012), allowing for theories focusing on the role of disrupted executive functioning (Woody & Sadler, 2008) to account for the data. Moreover, a recent replication failure of the Rounis et al. (2010) study suggests that the stimulation of the DLPFC with rTMS might not impair visual awareness (Bor, Schwartzman, Barrett & Seth, 2017); the meaning of these findings is a matter of ongoing debate (Ruby, Maniscalco & Peter, 2018; c.f., Bor, Barrett, Schwartzman & Seth, 2018).

lower the interference and the Stroop effects (as measured by the difference in response times (RTs) between the incongruent and neutral, and the incongruent and congruent trials, respectively) compared to their own performance in a non-hypnotic condition (Augustinova & Ferrand, 2012; Parris et al., 2012; Raz et al., 2002; Raz, Kirsch, Pollard & Nitkin-Kaner, 2006). Moreover, low suggestible people cannot reproduce this improvement in performance (Casiglia et al., 2010; Raz & Campbell, 2011; Raz et al., 2002, 2003) further underlining the notion that hypnosis and so the ability to respond hypnotically can have a causal influence on first-order states. It has been proposed that the word-blindness suggestion allows people to gain control over otherwise automatic processes (i.e., reading), specifically, by being able to dampen the processing of input words (Raz et al., 2002; Raz et al., 2006; Raz, Fan & Posner, 2005).

Overall, these findings cast doubt on the idea that a response by becoming hypnotic only impacts HOTs of intending and cannot alter first order abilities, but the findings do not refute the cold control theory *per se*. First, the cold control theory postulates that to produce the word blindness effect, one has to have a first-order intention to create the experience of the script as being meaningless by using a strategy at the disposal of the person without having an accurate HOT about intending to do so. Consequently, cold control theory asserts that the mere comparison of a suggestion and a no suggestion (Stroop task under normal circumstances) condition overlooks the fact that people have been (implicitly) instructed to create an experience of meaninglessness in the former case but they were told not do so in the latter one (Dienes, 2012). Therefore, this contrast cannot inform us whether the power of imagination (i.e., creating a counterfactual model of reality in which meaning cannot be extracted from the script) depends on the form of the accompanying HOT. Second, individual differences between highs and lows in the ability to create word blindness can account for the disparity in their performance, and indeed, it has been found that highs and not lows can produce the word blindness effect as a response to suggestions in the absence of a hypnotic induction (Parris & Dienes, 2013). This latter finding may seem to settle the matter in favour of cold control: subjects have no more first-order abilities responding hypnotically as non-hypnotically. However, this conclusion depends on hypnotic responding being entirely conditional on a previous hypnotic induction. If a subject can, without an induction, respond hypnotically, then the mere presence or absence of a hypnotic induction is irrelevant to theory testing. Indeed, it has been shown that highs can for example produce

hallucinations in response to suggestion, or dramatically relieve pain, without a previous hypnotic induction (e.g., Kirsch et al., 2008; Milling et al., 2002). Moreover, it has been demonstrated that the induction procedure might be irrelevant to the production of the feeling of involuntariness, which is the core feature of a hypnotic response; for example, highs reported comparable levels of involuntariness after a suggestion to experience a sex change with and without a prior induction (McConkey, Szeps & Barnier, 2001). Thus, the use of an induction or not is not relevant to testing the prediction of cold control theory. What is relevant is requesting subjects to have the same first-order intentions while having an accurate or inaccurate HOT about the intention. That is, a clear test of the key prediction of cold control theory necessitates the contrast of the control of highs experienced as voluntary with the control of highs experienced as involuntary (henceforth voluntary and involuntary control) in the capacity of reducing the Stroop interference effect while asked to achieve this by having the same first-order intention. By this, we could investigate whether hypnosis is purely a metacognitive phenomenon. Cold control theory defines hypnotic responding as nothing more nor less than acting intentionally while having the inaccurate HOT that one is not intending to perform that action. This is perhaps one of the simplest theories of hypnosis one could have: the essence of a response being hypnotic lies only in a type of metacognitive monitoring. Thus, critically testing the theory is important: Is a more complex theory needed or not?

One might argue that former research has already tested the core claim of cold control theory in studies investigating the efficiency of imagination compared to hypnotic responding to suggestions and that the theory has been disconfirmed. For instance, there is evidence that the fusiform activation of highs is bilateral when they are responding to a hypnotic suggestion to hallucinate colours whereas only the right fusiform shows activation when they are requested to imagine a grey-scale pattern in color, indicating that voluntary imagination might not produce the same visual experience as hypnotic responding (Kosslyn, Thompson, Costantini-Ferrando, Alpert & Spiegel, 2000). In addition, it has been shown that highs produce stronger pain experience of heat when responding to a hypnotic suggestion contrasted with a request to imagine the same type of pain (Derbyshire, Whalley, Stenger & Oakley, 2004). However, in both of these studies other factors than a mere change in monitoring of the HOT of intending might have been in play to produce varying experiential and neuropsychological responses. For example, if the wording is not carefully phrased in the non-hypnotic and hypnotic conditions than

it can create demand characteristics resulting in a “hold back” effect (Spanos, 1986; Zamansky, Scharf & Brightbill, 1964) or stronger expectations in the hypnotic condition (Braffman & Kirsch, 1999), as the participants aim to please the experimenter or they do not believe that their non-hypnotic response can be as effective as the hypnotic one. We argue that none of these studies provide an unequivocal test of the prediction of cold control theory as the expectations of the subjects were not measured in any of them. The wording of the conditions in Kosslyn et al.`s (2000) experiment were not designed to convince the participants that they can and should try to create comparable responses in the different conditions. Further, and crucially, it was not demonstrated in these examples that the imagination condition involved greater feelings of voluntariness than the hypnotic condition; thus cold control may have been the mechanism in both conditions. Therefore, a genuine test of the prediction of cold control theory need to possess a volitional request that can create equal level of expectations about the efficiency of non-hypnotic and hypnotic responses ensuring that the participants expect to perform the same with and without the HOT of intending.

To address this issue, we constructed a fully within-subjects design experiment in which the performance of involuntary and voluntary control can be directly compared. We employed three experimental conditions using highly suggestible subjects. In the posthypnotic suggestion condition (henceforth simply “suggestion” condition), we used the word-blindness posthypnotic suggestion to see the words as meaningless characters during the Stroop task. In the volition condition, we told the participants to reproduce the effect of the word blindness suggestion by responding to our volitional request to imagine the words as meaningless characters while doing the Stroop task. In the no suggestion condition, we asked the participants to undertake the Stroop task with the instruction of not imagining the words as meaningless so that we can measure their baseline performance. In this scenario, the cold control theory predicts that people can overcome the Stroop interference to the same extent in the suggestion and volition conditions when compared to the no suggestion condition. Therefore, if the results show a stronger reduction of the interference effect in the suggestion compared to the volition condition then one has to conclude that there is more to hypnosis then the strategic relinquishment of metacognitive monitoring in the form of accurate HOTs of intending. The experiment is testing a core prediction of the simplest version of cold control theory and so if it is disconfirmed, we need to revise the theory to fit the data. The key assumption of cold

control theory is that the difference between hypnotic versus non-hypnotic responding is just the difference between having and not having a HOT; if this assumption is retained, the finding of a greater Stroop reduction in the suggestion rather than volition condition would imply that the HOT of an intention can have a causal influence on first order states by hindering cognitive control processes (a rare finding of conscious executive processing being less effective than unconscious, contrast Cleeremans, 2006).

A key relevant outcome neutral test is that subjects experienced the word blindness effect as more volitional in the volition condition than in the suggestion condition. This would be the evidence that there was a difference in the presence of relevant HOTs of intending. To the best of our knowledge, this study is the first that measured the subjects' conscious experience of control over 'word meaninglessness' to unravel whether such an experience feels like something that has been intentionally imagined or merely perceived. Investigating the phenomenological level of the participants' cognition can inform us whether their behaviour felt involuntary when the suggestion was active compared to the volitional control. Moreover, controlling the potentially confounding role of expectations is imperative (Braffman & Kirsch, 1999), so we implemented a self-report measure to gauge the participants` expectations about seeing the words as meaningless characters. If subjects reported different levels of expectation for producing a word blindness effect in the suggestion than in the volition condition, expectations alone may explain differences in Stroop reduction in the two conditions (Magalhães De Saldanha da Gama et al, 2013). In addition, we took the participants` subjective experiences of `word meaningfulness` to explore the extent to which voluntary and involuntary control can alter the conscious experiences of the world. The measures reflect the extent to which subjects subjectively responded to the suggestions and to the volitional request; they could therefore constitute the crucial test of whether suggestions and volitional requests are equally effective. However, as the apparent problem for cold control lies with the objective measure of Stroop reduction, it is the RT measures that form the crucial test. Finally, we measured the `depth` of hypnosis to shed more light on the nature of the experienced state that accompanies the implementation of both types of control. This is an exploration, a sideline from the main point of the experiment, testing the assumption that the experience as of being in a hypnotic state, as interpreted by the participants, does not accompany post-hypnotic suggestion (e.g., Terhune, Luke & Cohen Kadosh, 2017).

First, we report a pilot experiment using this procedure. While the results yielded moderate evidence against cold control theory, the procedure and analyses were not pre-registered. Further, there was not a strict stopping rule (albeit Bayesian analyses were used). Thus, the pilot study will serve as a basis for a proper pre-registered test of cold control theory.

Pilot Experiment

Methods

Participants. Thirty-three highly suggestible students of the University of Sussex, all proficient readers of English, attended the experiment in exchange for course credits or payment. Eleven participants were recruited in 2013 and twenty-two students were recruited in 2014. The students had been screened in group sessions for being highly suggestible prior to the study. Students scoring 9 or higher on the Waterloo-Stanford Group Scale of Hypnotic Susceptibility, Form C (WSGC; Bowers, 1993) were recruited to the study. The participants granted their informed consent before participation and the Ethical Committee of the University of Sussex has approved the study.

Stimuli and apparatus. The stimuli of the experiment closely followed those used by Raz et al. (2002). The stimuli consisted of 4 types of color words (RED, BLUE, GREEN, and YELLOW) and 4 types of neutral words (LOT, SHIP, KNIFE, and FLOWER). The stimuli set of the congruent condition included the color words presented in colours matching the meaning of the words (e.g., RED in the color red). The incongruent items were color words displayed in colours mismatching the meaning of the word (e.g., RED in the color blue). The neutral words were length-matched to the color words and so all items had their corresponding presentation color (e.g., LOT presented always in red). All words were written in upper-case font and presented against a white background. The vertical visual angle of the stimuli was 0.5º, while the horizontal visual angle of the stimuli lied between 1.3º and 1.9º depending on the length of the word. The distance between the participants` eyes and the computer screen was approximately 65cm. The response keys used in the experiment were “V”, “B”, “N”, “M” for the colours red, blue, green and yellow, respectively. The keyboard buttons were not colour-labelled. The experiment was produced in and run by the software Experiment Builder (SR Research Ltd, Ottawa, ON, Canada).

Design and procedure. The study had a 3x3 within subjects design with the independent variables of the congruency type of the trial (congruent vs. neutral vs. incongruent) and the experimental condition (no suggestion, suggestion, volition). The proportion of congruent, neutral and incongruent trials was equal (33%) in each condition and the presentation of color and neutral words was frequency and length matched. The conditions were counterbalanced across participants and the Stroop trials (144 per condition) were displayed in a random order within each condition.

The experiment took place in a small room with the experimenter present and only one participant at a time. After providing their informed consent to the study, the participants engaged in a practice Stroop task for 5 minutes. The participants were asked to lay their left middle finger on "V", left index finger on "B", right index finger on "N" and right middle finger on "M" while undertaking the Stroop task. The participants were told to focus on the middle of the screen during the Stroop task, where a black fixation cross appeared for 1500ms at the beginning of each trial. The fixation cross was replaced by one of the Stroop stimuli and remained on the screen until response. Finally, a feedback ("CORRECT" or "INCORRECT") flashed in black on the screen and then a new trial started with the fixation cross. The response to stimulus interval was 2000ms. This sequence remained constant among the experimental conditions.

Next, a hypnotic induction[18] with the post-hypnotic suggestion to see the words as meaningless characters (Raz et al., 2002) was delivered by the experimenter and the participants were told that a clap would activate and a double clap would deactivate this suggestion. To test the effectiveness of the suggestion, the experimenter activated it by the clap and asked the participant to rate the meaningfulness of a presented coloured word on the following scale: 1 - completely clear, 2 - little unclear, 3 - unclear, 4 - completely unclear. Those who reported to see the word completely clear or little unclear received an additional instruction: "Notice how as you look at the word on the screen, you can look at it with the meaning fading to the background of your mind. We have found even when people consciously experience some meaning after this suggestion, they still process the words differently at a deeper level. You know you are capable of not reading meaning fully, remember how you have zoned out while reading a book.". Finally, the suggestion

[18] Although, according to the cold control theory, the usage of the induction procedure is not necessary to produce a hypnotic response to a suggestion, we included the induction in the protocol to make sure that the responses of the subjects can unambiguously be considered as hypnotic.

has been deactivated, and the participants have been brought back to wakefulness by a deinduction. For the exact wording of the protocol, see Appendix D.

Subsequently, the participants undertook the three experimental conditions in a random order. In the no suggestion condition, the participants were told to respond as fast and as accurately as they could, and they were asked not to make any attempt to see the words as gibberish or words of a foreign language. The suggestion condition started with a clap accompanied by a sentence highlighting that the suggestion had been activated. At the end of the condition, the suggestion was deactivated by the double clap. In the volition condition, the participants were requested to voluntarily reduce the Stroop interference:

"Highly hypnotisable individuals such as you have been shown to be able to eliminate the interference from the irrelevant word when under the influence of the post-hypnotic suggestion and even when the suggestion is given without hypnosis. We would like you to voluntarily strongly and clearly imagine the irrelevant words as gibberish, words of a foreign language so that no meaning can be taken from them. This is not a hypnotic suggestion and we have not hypnotised you for this part of the task. You'll notice we have not initiated a suggestion by clapping or giving any other cue. You have the ability to do that anytime you please, under your control, as effectively as you just did. Please now voluntarily remove meaning from the words. You can do this so that it is under your control, just by exercising your imagination. You can be aware it is your imagination at the same time as it produces powerful effects."

Throughout the experiment, we administered several self-report measures, and in each case, the experimenter read out loud the question and provided the answer options on a sheet for the participants. Before the start of the Stroop task in each condition, the participants reported their expectations on how certain they are that the words will be meaningless. When they finished the Stroop task, the following measures were taken: four items assessing the subjective experience of the meaningfulness of the words; a task to recall the words they have seen[19]; depth of hypnosis scale (Hilgard & Tart, 1966); an item gauging the experienced control over the meaningfulness of the words; a dichotomous question whether they perceived or imagined the words as meaningless[20].

[19] The data regarding this question have not been utilised for this project.
[20] This question was omitted from the no suggestion condition

For the exact questions and answer options see Appendix E. After finishing the last condition, the participants were thanked and debriefed.

Data analysis

Statistical analyses. We conducted all of our analyses with the statistical software R 3.3.1 (R Core Team, 2016). Since we had a fully within subjects design, we calculated difference scores so that we were able to test directly all of our hypotheses with Bayesian paired t-tests (we only conducted direct contrasts; i.e. not an omnibus *F* or *B* comparing the three conditions as the omnibus statistic would not be informative in terms of our hypotheses). Along with frequentist statistics, we calculated the corresponding Bayes factor (B) which was used as the basis of decision making in respect of the compared hypotheses.

Bayes factor. The Dienes and McLatchie (2018) calculator in the R environment was used to calculate the Bayes factors, which has a t-distribution as a likelihood function for the data, and we set the degrees of freedom of the theory to 10,000 in each analysis to have a likelihood function for the theory close to normal. The computation of the B requires the specification of the prediction of the two models that we intend to compare. We applied a half normal distribution with a mode of zero to model the predictions of the alternative hypotheses, as the tested hypotheses have directional predictions and assume that smaller effects are more probable than larger effects (Dienes, 2014). We report Bs in the following format: $B_{H(0, X)}$, in which H indicates that the model is half-normal, the first parameter (0) stands for the mode of the distribution and the second parameter (X) is the SD of the distribution. To specify the standard deviation of the alternative models, we applied the following strategies. Based on the meta-analysis of Parris, Dienes & Hodgson (2013) who have found that the word blindness suggestion generally halves the interference effect of the baseline (no suggestion) condition, we employed half of the interference effect observed in the no suggestion condition as the SD of all models testing the difference between the suggestion and volition and the no suggestion conditions. In order to test the traditional Stroop and interference effects, we used the average of the Stroop and interference effects found among studies containing the word blindness suggestion (See Table 1 of Parris et al., 2013). Concerning the self-report measures, we applied the rule of thumb of Dienes (2014) that suggests, in the absence of prior information, to halve the scale of measurement and use it as the SD of the one-sided

model (if that matches scientific intuitions closely enough: In this case a population mean difference anywhere on the scales is not completely unreasonable).

Although, B is a continuous measure of evidence by definition, we used the convention of 3 and 1/3 to distinguish between no evidence and good enough evidence for the alternative and null hypotheses, respectively (Jeffreys, 1961). Moreover, we use the label of moderate evidence for the values between 3-10 or 1/3-1/10, and the label of strong evidence for Bs greater than 10 or smaller than 1/10, in order to highlight the strength of the evidence (Lee & Wagenmakers, 2013).

A Bayes factor is the strength of evidence for one model over another and thus depends on what the models are (Rouder, Morey, Verhage, Province & Wagenmakers, 2016; Rouder, Morey & Wagenmakers, 2016). We have endeavoured to keep the models simple and otherwise scientifically informed; nonetheless, the chosen parameters (e.g., the SD of a half-normal distribution) could be motivated in different ways. Therefore, to ascertain the robustness of our Bayesian conclusions to the SDs of the H1 models, we report a robustness region for each B, providing the range of SDs of the half-normal models that qualitatively support the same conclusion (using the threshold of 3 for moderate evidence for H1 and ⅓ for moderate evidence for H0) as the chosen SD[21]. The robustness regions are notated as: RR [x1, x2] where x1 is the smallest and x2 is the largest SD that gives the same conclusion.

Bayesian parameter estimation with 95% Credibility intervals. To explore the extent to which the post-hypnotic suggestion or the voluntary control reactivates a hypnotic trance, we applied parameter estimation rather than hypothesis testing. To conduct the estimation, we report the condition means of the depth of hypnosis with the 95% Credibility Intervals (CI). Note that the 95% CIs are numerically identical to the 95% Confidence Intervals as we employed uniform prior distributions.

Results

Data transformation. The data of three participants were partially missing (one participant had only response time data whereas two participants had only self-reported data), and therefore they were excluded from the analyses. Trials with errors were omitted from the analysis of the response times (RTs) data (4.7% in total from which 1.4% from

[21] Thanks to Balazs Aczel for this suggestion

the no suggestion, 1.9% from the suggestion and 1.5% from the volition conditions)[22]. Moreover, using the outlier exclusion criterion of Raz et al. (2002), we deleted RTs that were 3 standard deviations either above or below the mean (1% of the correct trials from which 0.2% from the no suggestion, 0.3% from the suggestion and 0.4% from the volition conditions). In order to test the congruency related effects, we computed new variables. We calculated the extent of interference effect (RT incongruent – RT neutral) in the different suggestion conditions for each participant. The interference effect was specifically identified by Parris et al. (2013) as the Stroop component most reliably affected by the word blindness suggestion.

Outcome neutral tests 1: Was there a Stroop effect and did the suggestion work? As expected, the RTs were the longest in the incongruent (M = 811, SD = 182) followed by the neutral trials (M = 766, SD = 177) and the fastest in the congruent trials (M = 729, SD = 173). Comparing the conditions revealed support for the Stroop interference ($t(29) = 6.34$, $p < .001$, $M_{diff} = 45$ ms, $d_z = 1.16$, $B_{H(0, 62)} = 8.1*10^3$, RR[3, $1.47*10^5$]) and the Stroop effects ($t(29) = 8.09$, $p < .001$, $M_{diff} = 82$ ms, $d_z = 1.48$, $B_{H(0, 90)} = 7.4*10^5$, RR[5, $2.79*10^5$]). Also importantly, we found moderately strong evidence for the classical word blindness effect ($t(29) = 1.99$, $p = .056$, $M_{diff} = 34$ ms, $d_z = 0.36$, $B_{H(0, 30)} = 3.99$, RR[15, 63]), as the extent of the Stroop interference was reduced from the baseline of 60 ms to 26 ms in the suggestion condition.

Outcome neutral tests 2: Did suggestion and volition conditions differ in experienced degree of control? The analysis of the experienced level of control over the meaningfulness of the words indicated that the instruction to imagine the word as meaningless characters triggered a process experienced as more controlled than the suggestion ($t(29) = 5.34$, $p < .001$, $M_{diff} = 0.9$, $d_z = 0.98$, $B_{H(0, 1.5)} = 5.4*10^3$, RR[0.07, $2.75*10^2$]). Although, the participants tended to report that they *perceived* the script as meaningless in the suggestion condition (64% of the participants reported that they perceived rather than imagined the meaninglessness) and they rather *imagined* it in the volition condition (57% of the participants reported that they imagined and not perceived the meaninglessness), the results remained insensitive concerning whether the two

[22] Note that we do not possess the raw data collected in 2013 anymore (only the RTs averaged across trials and within conditions and participants), therefore, these percentages have been based on the data collected from 22 participants in 2014.

procedures are different in nature ($t(25) = 2.00$, $p = .056$, $M_{diff} = 0.23$, $d_z = 0.39$, $B_{H(0, 0.5)} = 2.78$, RR[0.45, 4.8]).[23]

Crucial test: Is the suggestion equally effective for suggestion and volition conditions? Next, we tested the key prediction of the cold control theory by comparing the suggestion and volition conditions in terms of the RTs of interference effects, and the analysis yielded supporting evidence of a smaller interference effect in the suggestion condition ($t(29) = 2.03$, $p = .052$, $M_{diff} = 25$ ms, $d_z = 0.37$, $B_{H(0, 30)} = 4.00$, RR[11, 50]). The participants managed to decrease the interference by 34 ms in the suggestion condition and only by 9 ms in the volition one compared to the no suggestion condition. However, the evidence regarding the difference between the volition and no suggestion conditions remained insensitive ($t(29) = 0.51$, $p = .611$, $M_{diff} = 9$ ms, $d_z = 0.09$, $B_{H(0, 30)} = .74$, RR[0, 81]). Table 1 displays the descriptive statistics of the RTs in the congruency conditions broken down by the experimental conditions. Figure 1 depicts the distribution of the Interference scores in the three experimental conditions.

[23] Note that the corresponding item of the questionnaire had only two levels (either imagination or perception), but we analysed the data as a continuous variable to make it comparable with the measure we will use in the pre-registered experiment. The Supplementary Materials include an analysis of these data that considers this item as a dichotomous variable and aims to estimate the effect size. The results are in accordance with those in the main text, namely, the estimation revealed that the effect size lies within a broad range covering values larger as well as smaller than 1(OR = 4, 95% CI[0.64-25.02]).

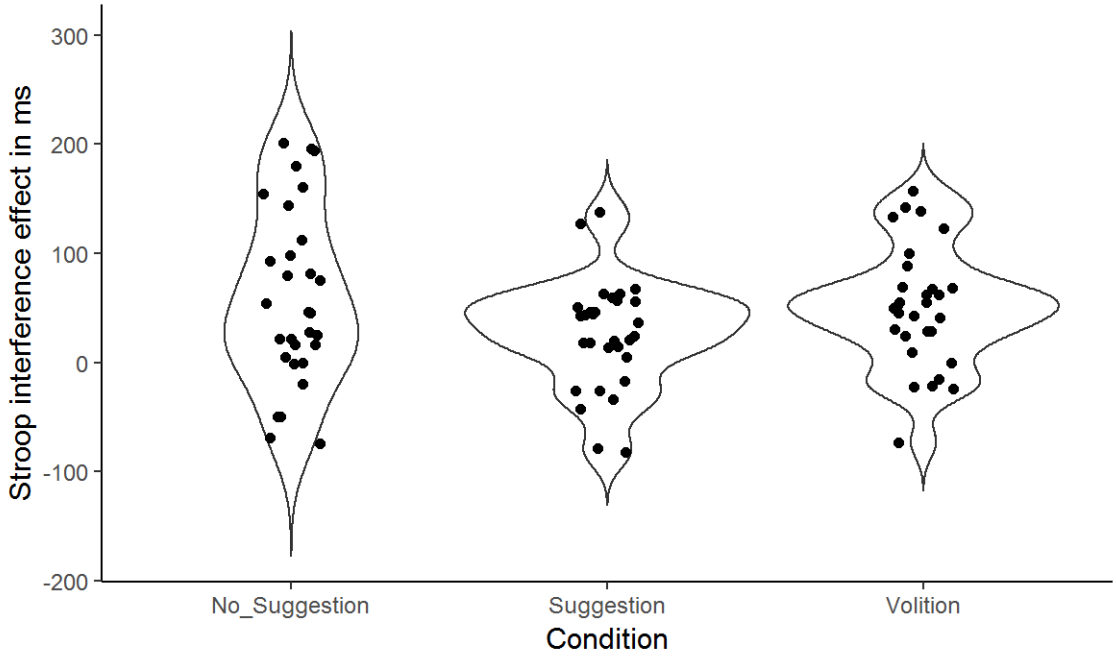

Figure 1. Violin plot portraying the distribution of interference scores in the three experimental conditions. The black dots indicate individual data points; one dot represents the interference score of a single participant.

The expectations to see the words as meaningless characters were raised in both of the suggestion ($t(29) = 5.99$, $p < .001$, $M_{diff} = 1.7$, $d_z = 1.09$,$B_{H(0, 2.5)} = 3.19*10^4$, RR[0.11, $5.42*10^2$]) and the volition conditions ($t(29) = 5.65$, $p < .001$, $M_{diff} = 1.58$, $d_z = 1.03$, $B_{H(0, 2.5)} = 1.27*10^4$, RR[0.11, $4.93*10^2$]) compared to the no suggestion condition. Yet, these increments were comparable (See Table 1) and there is evidence for no difference between the suggestion and volition conditions ($t(29) = 0.38$, $p = .710$, $M_{diff} = 0.12$, $d_z = 0.07$, $B_{H(0, 2.5)} = 0.18$, RR[1.33, Inf]) implying that the the suggestion effect was enhanced hypnotically versus volitionally beyond the impact of expectations (we have not regarded this as an outcome neutral test, in that if there had been an difference in expectancies we still could have conducted a version of the crucial test by partialling out expectancy effects).

Supporting test of interest: Do suggestions and volitional requests produce the same subjective response? We investigated whether the manipulation of the meaninglessness of the words was successful on the subjective level of the participants, and whether the posthypnotic suggestion and volitional request produced similar subjective responses. The descriptive statistics about the subjective experience of meaninglessness are shown in Table 2 for each question and condition separately. Note

that the first question was phrased reversed compared to the other three questions, thus, smaller values indicate stronger experience of meaninglessness in that case. The results of the phenomenological data on how the meaninglessness was sensed were only partly in line with the findings of the RTs. Statistically speaking, we found strong evidence favoring suggestion and volition over no-suggestion in creating vivid experiences of meaninglessness. Although, the participants reported that they sensed more meaningless words in the suggestion than in the volition condition (in case of three measures from the four) the evidence regarding the advantage of the posthypnotic suggestion over volition remained insensitive in each case.

Q1. We found strong evidence for a difference between no-suggestion and each of the suggestion ($t(29) = 5.78$, $p < .001$, $M_{diff} = 45$, $d_z = 1.05$, $B_{H(0, 50)} = 2.04*10^4$, RR[3, $1.42*10^4$]) and volition conditions ($t(29) = 4.29$, $p < .001$, $M_{diff} = 36.8$, $d_z = 0.78$, $B_{H(0, 50)} = 4.15*10^2$, RR[4, $9.45*10^3$]). However, the evidence is insensitive in respect of the difference between these latter two conditions($t(29) = 1.56$, $p = .13$, $M_{diff} = 8.2$, $d_z = 0.29$, $B_{H(0, 50)} = 0.65$, RR[0, 99]).

Q2. The results revealed strong evidence in favour of the difference between no suggestion and suggestion conditions ($t(29) = 4.69$, $p < .001$, $M_{diff} = 34.1$, $d_z = 0.86$, $B_{H(0, 50)} = 1.10*10^3$, RR[3, $9.65*10^3$]) and moderate evidence favouring a difference between no suggestion and volition conditions: ($t(29) = 2.64$, $p = .013$, $M_{diff} = 19.6$, $d_z = 0.48$, $B_{H(0, 50)} = 6.90$, RR[5, $1.23*10^2$]). The data show insensitivity whether the suggestion and volition conditions differ ($t(29) = 2.04$, $p = .0502$, $M_{diff} = 14.5$, $d_z = 0.37$, $B_{H(0, 50)} = 1.99$, RR[31, $3.14*10^2$]).

Q3. The results indicate strong evidence favouring a difference between no suggestion and suggestion conditions ($t(29) = 3.66$, $p < .001$, $M_{diff} = 24.5$, $d_z = 0.67$, $B_{H(0, 50)} = 71.47$, RR[3, $1.35*10^3$]) and between no suggestion and volition conditions ($t(29) = 3.32$, $p = .002$, $M_{diff} = 28.7$, $d_z = 0.61$, $B_{H(0, 50)} = 36.9$, RR[4, $7.33*10^2$]). Moreover, we have strong evidence that the suggestion and volition do not differ ($t(29) = -.59$, $p = .557$, $M_{diff} = -4.3$, $d_z = -0.11$, $B_{H(0, 50)} = 0.09$, RR[14, Inf]).

Q4. Similarly to Q2, we found strong evidence for a difference between the no suggestion and suggestion conditions ($t(29) = 5.31$, $p < .001$, $M_{diff} = 34.4$, $d_z = 0.97$, $B_{H(0, 50)} = 5.51*10^3$, RR[3, $1.05*10^4$]), moderate evidence for a difference between no suggestion and the volition condition ($t(28) = 2.60$, $p = .015$, $M_{diff} = 17$, $d_z = 0.48$, $B_{H(0,$

$_{50)}$ = 5.71, RR[4, 100]) and insensitive evidence for the difference between suggestion and volition conditions ($t(28) = 2.05$, $p = .049$, $M_{diff} = 15.2$, $d_z = 0.38$, $B_{H(0, 50)} = 2.11$, RR[33, $3.35*10^2$]).

Exploration: Do post-hypnotic suggestions produce a hypnotic trance at the time of activating the suggestion? The suggestion might not be truly *post*-hypnotic as the participants reported being relaxed or even hypnotised ($M = 1.37$, 95% CI[1.06 – 1.67]) when the post hypnotic suggestion was triggered, indicating that a hypnotic state might have been experienced. Seemingly, voluntary control does not involve experiencing this hypnotic state, as the upper bound of the 95% CI ($M = 0.8$, 95% CI [0.52 - 1.08]) barely exceeded the level of being relaxed.

Table 1

Summary Table about the Means of the RTs and Self-report Measures in the three Experimental Conditions

		Experimental condition		
Category	Item (scale)	No Suggestion	Suggestion	Volition
Response times (RTs)	Incongruent (ms)	853 (187)	775 (207)	805 (213)
	Neutral (ms)	793 (183)	749 (205)	755 (198)
	Congruent (ms)	748 (141)	712 (212)	726 (214)
Expectations	Expecting the words to be meaningless (0-5)	0.59 (1.03)	2.29 (1.40)	2.17 (1.28)
Experienced Control	Control over meaningfulness (0-3)	2.33 (0.84)	1.1 (0.69)	2 (0.71)
	Perception vs. Imagination (% of perception)	-	64% (49)	43% (50)
Depth of hypnosis	Depth of hypnosis during the task (0-3)	0.43 (0.5)	1.37 (0.81)	0.8 (0.76)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Table 2

Summary Table of the four Items Measuring the Subjective Experience of Meaninglessness

Item	Experimental condition		
	No suggestion	Suggestion	Volition
Q1: „Was the meaning of the words on the screen completely clear to you"	81.8% (25.9)	36.8% (28.5)	45% (35)
Q2: „Were you aware of only an unclear meaning of the words on the screen"	13.7% (25.7)	47.8% (34.3)	33.3% (30.6)
Q3: „Were you just aware of the color and had no idea of what script of the words were written in"	20.9% (29.3)	45.3% (31.8)	49.6% (33.7)
Q4: „Were the words on the screen written in a clear yet meaningless script"	12.6% (23.1)	47% (30.5)	30% (31)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Discussion

In this experiment we aimed to discover whether highly suggestible people can produce the word-blindness effect outside of hypnotic context by voluntarily imagining the words as meaningless. The results provided moderate to strong evidence supporting the successfulness of the experimental manipulations in outcome neutral tests. Most importantly, the classical word-blindness effect was replicated and the volitionally induced meaninglessness was experienced as voluntary compared to its post-hypnotic counterpart. Although, the second measure assessing the nature of control was not sensitive, the amount of evidence was close to the convention of 3 (B = 2.78), suggesting that the process of meaninglessness was experienced as imagined in the volitional condition and as perceived in the suggestion condition. This difference between the two measures of control might be due to the fact that the latter item was only dichotomous and so not sensitive enough to capture the mild difference in how people sensed the meaninglessness. Therefore, a continuous item assessing the nature of control would be more appropriate. In sum the outcome neutral tests were satisfied and we can proceed with the crucial test.

The main results revealed that volitionally induced control by imagining the words as meaningless characters did not enhance performance on the Stroop task to the same

extent as the post-hypnotic suggestion. The evidence remained insensitive regarding the efficiency of voluntary control. Theories of hypnosis that regard the unique hypnotic nature of a response is constituted simply by a change in monitoring HOTs, such as the simplest versions of the cold control theory, cannot account for these data as it seems that the suggestion allowed highly suggestible people to more efficiently resolve conflict than it was possible for them through non-hypnotic means. Thus, retaining the assumption of cold control that hypnotic vs non-hypnotic action differ primarily in accurate HOTs of intending, it seems HOTs of intention, at least the intention to create the experience of meaninglessness, can disrupt task performance; thus, HOTs can have causal effects on first-order states (cf. Rosenthal, 2008). Incidentally, this finding depicts a counterexample for the concept that conscious cognitive control processes are superior to unconscious ones (Cleeremans, 2006), given the assumption that the hypnotic and volitional processes differ in the conscious status of the intentions.

A plausible candidate that can influence the two types of control to produce different results is the expectation about their efficiency, which is a well-known predictor of behaviors elicited by suggestions (Kirsch, 1985; Braffman & Kirsch, 1999). Our scale measuring the participants` expectations emphasised the experience of `word meaninglessness` and the results derived from these data indicate that expectancy was the same in the volitional and the suggestion condition, implying that expectancies of meaninglessness alone cannot account for the difference in the effectiveness in producing the word blindness effect in the two conditions. However, the underlying mechanism of the word blindness suggestion may not be related or restricted to visual processing, which would call into question the relevance of the scale we used in gauging expectations. Recent behavioral and neural studies of the word blindness suggestion provide evidence for the notion that the suggestion affects cognitive control processes rather than the visual input stream, thus, the successfulness of the suggestion might lie in the enhanced conflict resolution and not in the dampened perception of the meaning of the words (Casiglia et al., 2010; Augustinova & Ferrand, 2012; Parris et al., 2013; Zahedi, Stuermer, Hatami, Rostami & Sommer, 2017). In line with this view, the suggestion, in our experiment, seemed to influence the performance mostly on the incongruent trials by reducing it compared to the no suggestion and volition conditions. Therefore, a measure of the expectations should aim to assess the beliefs of people about the efficiency of voluntary and involuntary control and not solely focus on the experience of meaningfulness.

Incidentally, this new design will also allow us to critically evaluate the simplest form of the response expectancy theory (Kirsch, 1985), which claims that expectations are the single driving factors of hypnotic responses. Were the extent of interference different in volition and suggestion conditions while the expectations to see the words as meaningless characters, and to exert control be comparable in the two conditions, the response expectancy theory will need to be revised.

Cold control theory asserts that to create the experience of meaninglessness the subjects need to have a first-order intention to produce it by engaging in an active strategy. The exact mechanism of this strategy is a mystery currently, but several empirical studies have been conducted on this issue that can help us exclude possible explanations. For instance, as mentioned above, it has been demonstrated that neither the dampening of the visual input (Raz et al., 2003), nor the inhibition of meaning processing can be responsible for the whole word blindness effect (Augustinova & Ferrand, 2012; Parris et al., 2013; Zahedi, et al., 2017). These findings are in consonance with the fact that a posthypnotic suggestion that specifically requires highs to lose the ability of reading did not result in the reduction of the Stroop interference effect; it appears that the suggestion needs to include a phrase such as "words are meaningless gibberish" to be successful in enhancing performance (MacLeod, 2011). Consequently, the meaning of the words must be processed to some extent even in the suggestion condition indicating that the information that the scripts of the stimuli are in fact meaningful is available to the participants. This strikes a chord with the idea that highs need to hold two models of reality in the suggestion condition as they do in the volition one. In one model, the meaning can be extracted from the words, as they are meaningful, whereas in the other counterfactual model, this is not possible. Entertaining multiple models is the basis of pretence and imagination (Perner, 1991); not being aware that one intended to engage in pretence or imagination would, according to cold control theory, lead to the experience of hallucination or delusion.

Pre-registered Experiment

In this experiment, we intend to replicate the pilot experiment as a multi-lab pre-registered replication project to increase the evidential value (by virtue of a larger sample) of our data. Moreover, we introduce a new item measuring the participants` expectations to how easily they can overcome the interference in each condition. With this, we aim to address

more thoroughly whether involuntary control can be more efficient than the voluntary counterpart beyond the influence of expectations.

The questions that will be addressed are those of the pilot study: (a) Outcome neutral test: Was there a Stroop effect and did the suggestion work?; (b) Outcome neutral test: Did suggestion and volition conditions differ in experienced degree of control?; (c) Crucial test: Is the suggestion equally effective for suggestion and volition conditions in reducing Stroop interference as measured in RTs (taking into account expectations)?; (d) Supporting test of interest: Do suggestions and volitional requests produce the same subjective response?; (e) Side interest: Do post-hypnotic suggestions produce a hypnotic trance at the time of activating the suggestion? Is the depth of hypnosis different for the suggestion and volition conditions? In addition, we will explore whether the post-hypnotic suggestion produces a subjective experience of being in hypnosis while it is activated: Traditionally it is assumed that a post-hypnotic suggestion, by virtue of being post-hypnotic, does not involve the experience of being in a hypnotic state at the time of responding (Terhune, Luke & Cohen Kadosh, 2017). Furthermore, we will estimate the extent to which self-reported measures of the feeling of voluntariness converge to assess their validity. Finally, we will run an exploratory correlation analysis to estimate the extent to which the participants are engaging in the same cognitive strategy in the volition and suggestion conditions. The results of this analysis can be used to estimate the sample size of a future study that aims to reach a good enough evidence supporting the idea that the underlying mechanisms of the responses are either the same or different in the two conditions.[24]

Methods

Participants. Labs from the following institutions recruited participants throughout the academic year of 2018-19: University of Sussex, School of Psychology (US) Lancaster University Department of Psychology (LaU).[25] We invited highly suggestible students who are proficient readers of English to attend the experiment in exchange for course credits or payment. The amount of payment and course credits will be in line with the regulations of the local universities (£10 at US, £7 at LaU). The suggestibility of the students was gauged by the Sussex Waterloo Susceptibility to

[24] Thanks to an anonymous reviewer for suggesting this analysis.
[25] We planned to recruit participants at the University of Bournemouth (UB) and at Lund University (LuU); however, for various reasons we did not manage to collect data at these universities.

Hypnosis scale (SWASH; Lush, Moga, McLatchie & Dienes, 2018) prior to participation and the threshold of highly suggestibility was based on the composite SWASH score (top 15% of the population) of the first year psychology students at Sussex (year 2018)(matching typical percentages used to define "high" in the literature; Barnier & McConkey, 2004; Anlló, Becchio & Sackur, 2017). This score was 5.35 on a scale from 0 to 10. To reduce the cost of screening at LaU, where possible we planned toinvite participants to undertake the SWASH who were previously identified as highs with other measures. The participants were asked to read an information sheet about the study and consent to the terms of participation before starting the experiment. The local Ethical Committees have approved the study.

Since we relied solely on Bayes factors to draw statistical inference, we useed optional stopping (Rouder, 2014). The minimum sample size was set at 20 and then we conducted all of the crucial analyses after roughly every subject (as different labs are involved there were some clumping). We stop collecting the data when all outcome neutral tests provide at least moderate evidence supporting that they have been successful, or else have failed, and when the main test of the study, comparing volition and suggestion conditions, also become sensitive (i.e. the B either larger than 3 or smaller than 1/3). A sample size estimation based on the data of the pilot study suggests that we need around 40 participants to show supporting evidence for the null, if the difference between the samples is 0 ms and if the standard deviation of the crucial measure is the same in this study as observed in the pilot study (See Supplementary Materials for details of the analysis). Should any of the four analyses remain insensitive with 60 participants, we desist from recruiting more participants. We began to recruit participants after the date of in-principle acceptance (autumn term of 2018) and stop if all of the specified analyses reach sensitivity, if we have 60 participants or if the spring term of 2019 finishes (end of May).

In total, we recruited 46 highly suggestible people from which we needed to exclude 9 participants as we did not use the appropriate item to measure their experienced control over their response. This step was approved by the editor (for the final items of the control measure see Appendix G.). Moreover, we needed to omit one more person as their RT data file was corrupted. All of the analyses are done on the data of the remaining 36 participants (27 females, 1 unknown, M_{age} = 20.72, SD_{age} = 4.55) from which 33 were recruited at US and 3 at LaU.

Stimuli and apparatus. The materials of the registered experiment were identical to those used in the pilot. We employed OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) to compile and run the Stroop task part of our experiment. The resolution of the applied computer screens was either 1280x1024 or adjusted to these values so that the size of the presented stimuli remained constant across labs.

Design and procedure. The design of the registered experiment was in accordance with those of the pilot experiment. To ensure that none of our participants possesses color vision deficiency, we included a statement in the recruitment letter that only people with intact color vision can attend the study. In addition, we made three modifications in the instruction of the volition condition. Namely, we put the sentence "You have the ability to do that anytime you please, under your control, as effectively as you just did." before the following two sentences "You have the ability to do that anytime you please, under your control, as effectively as you just did. You'll notice we have not initiated a suggestion by clapping or giving any other cue.", in order to avoid the implication that the participants have the ability to activate the suggestion without the clap even in the volition condition. Moreover, we replaced the "as effectively as you just did" part with "as effectively as you did it during the hypnotic induction" to make it clear that we refer to the word blindness test that was done during the induction procedure. In addition, we add an extra sentence highlighting that the effect of a suggestion can be achieved through voluntary means. See Appendix F for the final instruction of the volition condition.

In addition, we introduced four amendments in the self-report measures: (a) we included a new item at the beginning of each Stroop condition measuring the expectations about the efficiency to control the interfering information; (b) we replaced the dichotomous answer option of the question measuring the experienced nature of meaninglessness by a continuous scale; (c) we omitted the question concerning the recall of the words; (d) we replaced the item measuring the depth of hypnosis to the one which is used in the SWASH (2018). See Appendix F for the new items.

Data analysis

The steps of the data analysis closely followed those of the pilot experiment, including the exclusion criterion regarding RT data and how we draw conclusions based on the results of the Bayes factors (e.g., outcome neutral tests and the crucial test).

In terms of Bayes factor calculation, we retained the parameters of the H1 models of the analyses with RTs. However, to increase the sensitivity of our tests with the self-report measures in comparing the suggestion and volition conditions, we were informed by the results of the pilot experiment. Specifically, given the score of pilot subjects in the volition condition we can determine the maximum predicted change allowed with respect to the suggestion condition. For example on a 0-3 scale of experienced control (0 = no control, 3 = complete control), the volition condition in the pilot study scored 2.0, so the suggestion condition could experience up to 2 rating units of less control (as it is expected to be smaller than the mean of the volition condition). The maximum difference between conditions was thus estimated as about 2 for the new experiment, and the SD of the half normal was set as max/2 = 1 rating unit (Dienes, 2014). For expectations (both questions), the SD was set at 1.4 by this process, and the SD for four items assessing subjective experiences as meaningless was set at 30.

We had three outcome neutral tests to ensure that our experiment is able to test the proposed question. All of these tests had to provide evidence favouring the alternative hypotheses to allow us to carry on with the main analyses. We tested the presence of the Stroop interference effect while ignoring the influence of the type of the control. We tested that the experienced degrees of control is higher in the volition than the suggestion conditions. Finally, we assessed whether the suggestion reproduced the word blindness effect by reducing the extent of Stroop interference in the suggestion condition compared to the no suggestion condition.

The crucial test of the experiment was the comparison of the suggestion and volition conditions in terms of the extent of Stroop interference. Thus, we based our final conclusion on this statistical test. In addition, we planned to run a further analysis to control for the effect of expectations, conditional on the test of difference in expectations between the volition and suggestion conditions. If the evidence does not reach ⅓ to support the claim that the beliefs about the efficiency of suggestion and volition are identical, we would conduct the following secondary test. We would use a regression model with the difference in the interference score between conditions (suggestion vs. volition) as the dependent variable and the difference for expectations (suggestion vs. volition) as independent variables (if none of the expectation measures provide evidence for the null then the outlined analysis would be done as a multiple regression with both of the measures as predictors in the model). To conduct the crucial analysis while

partialling out the effect of expectations, we would test the intercept of the regression line against zero. By this, we could examine the difference between the suggestion and volition effects while controlling for the effect of expectations. The parameters of this Bayes factor analysis would be the same as the one testing the main question of the study.

The following are not the main point of the experiment and are thus of secondary interest. We tested whether post-hypnotic suggestion and volitional request produce the same subjective responses in exactly the same way as was done in the pilot. We estimated hypnotic depth for no suggestion, suggestion and volition conditionswith 95% CIs assuming a uniform prior over the scale range. To explore whether these conditions differ in hypnotic depth, we calculated the Bayes factor for the following comparisons: difference between no suggestion, suggestion and volition conditions. We modelled H1 with a half-normal, and SD of 0.86 rating unit based on the difference between suggestion and volition conditions in the pilot (we used the difference between the volition and suggestion condition means after adjusting it according to the lengths of the new and old scales).

To estimate the convergent validity of the self-report measures of involuntariness, we calculated the correlation and 95% CIs of the "level of control" and "experienced nature of meaninglessness" items on the difference scores of the volition and suggestion conditions. We can assess whether people changed the conscious status of the intention to imagine by the difference between volition and suggestion conditions in the experienced nature of meaninglessness item (i.e. experienced as imagination vs perception). As "imagination" is not mentioned in the volition instructions (unlike in the pilot), this tests whether subjects report a change that was not directly instructed, but should still occur according to cold control theory. As this item has a 4-point scale as the degree of control scale does, we tested with the same model of H1 (i.e. SD = 1 unit). We calculated the correlation and 95% CI between the extent to which subjects can reduce the interference in the suggestion and volition conditions.

Results

Data transformation. Following the steps of the analysis of the pilot experiment, trials with errors were excluded from the analysis of the response times (RTs) data (6.9% in total from which 2.4% from the no suggestion, 2.2% from the suggestion and 2.3% from the volition conditions). Again, we omitted RTs that were 3 standard deviations

either above or below the mean (0.1% of the correct trials from which 0.3% from the no suggestion, 0.3% from the suggestion and 0.3% from the volition conditions).

Outcome neutral tests 1: Was there a Stroop interference effect and did the suggestion work? The pattern of the RTs followed that of the pilot experiment. The RTs were the longest in the incongruent (M = 830, SD = 144) followed by the neutral trials (M = 749, SD = 108) and the quickest in the congruent trials (M = 707, SD = 109). We found evidence both for the Stroop interference ($t(35) = 8.80$, $p < .001$, M_{diff} = 80 ms, d_z = 1.47 , $B_{H(0, 62)} = 1.70*10^8$, RR[4, $2.79*10^4$]) and the Stroop effects ($t(35) = 12.03$, $p < .001$, M_{diff} = 123 ms, d_z = 2.00, $B_{H(0, 90)} = 5.52*10^{11}$, RR[5, $4.38*10^4$]). Comparing the extent of the Stroop interference effect in the no suggestion and suggestion conditions revealed moderately strong evidence for the classical word blindness effect ($t(35) = 2.20$, p = .034, M_{diff} = 31 ms, d_z = 0.37, $B_{H(0, 30)}$ = 5.56, RR[10, 90]).The Stroop interference was 98 ms in the former and 67 ms in the later condition.

Outcome neutral tests 2: Did suggestion and volition conditions differ in experienced degree of control? We found strong evidence supporting that the participants experienced more control over the meaningfulness of the words when they responded to the volitional request than when they responded to the suggestion ($t(35)$ = 4.38, $p < .001$, M_{diff} = 0.61, d_z = 0., $B_{H(0, 1)} = 6.0*10^2$, RR[0.06, 166]). Moreover, the participants experienced the effect of the meaninglessness in the suggestion condition mostly as *perception*, whereas they experienced it in the volition condition mostly as *imagination* (See Table 3 for descriptive statistics). Importantly, we found evidence for the difference of the conditions ($t(35) = 2.68$, p = .011, M_{diff} = 0.23, d_z = 0., $B_{H(0, 1)}$ = 7.65, RR[0.08, 2.74]). Finally, we estimated the correlation of the two measures using the Kendall`s τ method as it is robust to the violation of the assumption of normality. We computed the 95% Credibility Interval of the estimate with the *credibleIntervalKendallTau* R function (van Doorn, Ly, Marsman, and Wagenmakers, 2016). The two items measuring whether the participants experienced control over their responses were associated only to a small extent and the possible population effect sizes lie within a rather wide range of values (τ =.10, 95% CI [-.13, .30]).

Crucial test: Is the suggestion equally effective for suggestion and volition conditions? To test the key prediction of the cold control theory, we compared the extent of the Stroop interference effect between the suggestion and volition conditions. The test

yielded data insensitivity with anecdotal evidence supporting the model predicting no difference ($t(35) = 0.66$, $p = .514$, $M_{diff} = 8$ ms, $d_z = 0.11$, $B_{H(0, 30)} = 0.65$, RR[0, 64]). The evidence regarding the difference between the volition and no suggestion conditions remained insensitive with anecdotal evidence supporting their difference ($t(29) = 1.46$, $p = .154$, $M_{diff} = 23$ ms, $d_z = 0.24$, $B_{H(0, 30)} = 1.93$, RR[0, 256]). Table 1 shows the means and SDs of the RTs in the congruency conditions broken down by the experimental conditions. Figure 1 depicts the distribution of the interference scores in the three experimental conditions.

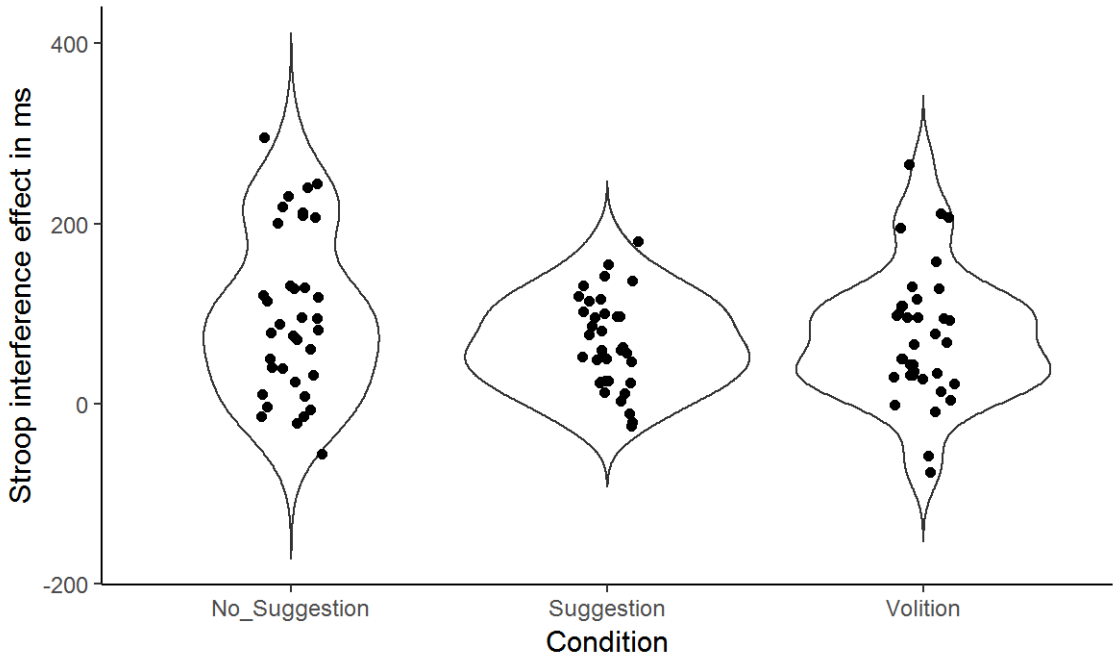

Figure 2. Violin plot depicting the distribution of interference scores in the three experimental conditions.

Importantly, we did not need to partial out the effect of the expectancies as for both of the measures we found supporting evidence for no difference between suggestion and volition. For the item measuring the expectations to experience the words as meaningless: $t(35) = -0.50$, $p = .618$, $M_{diff} = -0.11$, $d_z =$, $B_{H(0, 1.4)} = 0.11$, RR[]. For the item gauging the expectations that naming the color of the words will be easy: $t(35) = -0.50$, $p = .543$, $M_{diff} = -0.14$, $d_z =$, $B_{H(0, 1.4)} = 0.10$, RR[0.43, Inf]. For the first item, both of the suggestion ($t(35) = 6.25$, $p < .001$, $M_{diff} = 1.58$, $d_z =$, $B_{H(0, 1.4)} = 1.42*10^6$, RR[0.09, 521]) and the volitional request ($t(35) = 7.92$, $p < .001$, $M_{diff} = 1.69$, $d_z =$, $B_{H(0, 1.4)} = 1.60*10^7$, RR[0.08, 583]) elevated the expectations compared to the no suggestion

condition. See Table 1 for the descriptive statistics. For the second item, neither the suggestion ($t(35) = 0.47$, $p = .642$, $M_{diff} = 0.14$, $d_z =$, $B_{H(0, 1.4)} = 0.31$, RR[1.32, Inf]) nor the volitional request ($t(35) = 1.24$, $p = .223$, $M_{diff} = 0.28$, $d_z =$, $B_{H(0, 1.4)} = 0.60$, RR[0, 2.59]) managed to raise the expectations compared to the no suggestion condition.

Supporting test of interest 1: To what extent does the magnitude of the reduction in the Stroop interference by the suggestion and by the volitional request correlate? The extent by which the participants managed to reduce the Stroop interference while responding to the suggestion correlated strongly with their response to the volitional request ($r = .70$, 95% CI [.48, .83]). The plausible population effect sizes range among the large effect sizes as one can be 97.5% confident that the effect size is not smaller than .48.

Supporting test of interest 2: Do suggestions and volitional requests produce the same subjective response? We tested whether responding to the suggestion and to the volitional request was accompanied by a comparable phenomenology, namely the subjective experience of meaninglessness. The results of the subjective reports were mostly in line with the RTs as we found strong evidence for all four measures that the experience of meaninglessness was elevated by the suggestion and volitional request compared to the no suggestion baseline. The tests revealed evidence supporting the claim that one experiences the same level of meaninglessness while responding to the suggestion and the volitional request. ***Q1.*** We found strong evidence for a difference between no-suggestion and both of the suggestion ($t(35) = 4.91$, $p < .001$, $M_{diff} = 28.1$, $d_z =$, $B_{H(0, 30)} = 3.12*10^3$, RR[3, $8.39*10^3$]) and volition conditions ($t(35) = 5.78$, $p < .001$, $M_{diff} = 30.1$, $d_z =$, $B_{H(0, 30)} = 3.64*10^4$, RR[2, $9.69*10^3$]). Moreover, there was evidence supporting the null effect in respect of the difference between the suggestion and volition conditions ($t(35) = -0.61$, $p = .549$, $M_{diff} = -2.1$, $d_z =$, $B_{H(0, 30)} = 0.20$, RR[18, Inf]).

Q2. Again, the results indicate strong evidence in favour of the difference between no suggestion and suggestion conditions ($t(35) = 4.43$, $p < .001$, $M_{diff} = 24.9$, $d_z = 0.$, $B_{H(0, 30)} = 804$, RR[3, $6.86*10^3$]) and between no suggestion and volition conditions ($t(35) = 5.95$, $p < .001$, $M_{diff} = 28.2$, $d_z = 0.$, $B_{H(0, 30)} = 5.90*10^4$, RR[2, $9.15*10^3$]). We found strong evidence supporting that suggestion and volition conditions do not differ ($t(35) = -0.80$, $p = .429$, $M_{diff} = -3.3$, $d_z = 0.$, $B_{H(0, 30)} = 0.08$, RR[7, Inf]).

Q3. The comparisons revealed strong evidence favouring a difference between no suggestion and suggestion conditions ($t(35) = 3.08$, $p = .004$, $M_{diff} = 19.7$, $d_z = 0.$, $B_{H(0, 30)} = 25.19$, RR[3, 316]) and between no suggestion and volition conditions ($t(35) = 3.66$, $p < .001$, $M_{diff} = 20.4$, $d_z = 0.$, $B_{H(0, 30)} = 99.60$, RR[3, $1.26*10^3$]). We found moderate evidence for no difference between the suggestion and volition conditions ($t(35) = -0.18$, $p = .862$, $M_{diff} = -0.7$, $d_z = -0.$, $B_{H(0, 30)} = 0.12$, RR[10, Inf]).

Q4. Similarly to the other measures, there was strong evidence supporting the difference between the no suggestion and suggestion conditions ($t(35) = 4.23$, $p < .001$, $M_{diff} = 27.2$, $d_z = 0.$, $B_{H(0, 30)} = 486$, RR[3, $7.07*10^3$]), moderate evidence for a difference between no suggestion and the volition condition ($t(35) = 4.45$, $p < .001$, $M_{diff} = 25.1$, $d_z = 0.$, $B_{H(0, 30)} = 852$, RR[3, $6.95*10^3$]) and insensitive evidence for the difference between suggestion and volition conditions ($t(35) = 0.51$, $p = .611$, $M_{diff} = 2.1$, $d_z = 0.$, $B_{H(0, 30)} = 0.22$, RR[19, Inf]).

Exploration 1: Do post-hypnotic suggestions produce a hypnotic trance at the time of activating the suggestion? Applying the new measure of hypnotic depth, we found again that the reported level of hypnotic depth was the lowest in the no suggestion condition ($M = 0.93$, 95% CI[0.55 – 1.31]), followed by the volition condition ($M = 1.69$, 95% CI[1.21 – 2.18]), and it was the highest in the suggestion condition ($M = 2.33$, 95% CI[1.91 – 2.76]). Similarly to the pilot study, these results imply that participants' response to the suggestion might not have been *post*-hypnotic as they reported to be stronger than slightly hypnotised during the Stroop task. Comparing the conditions revealed strong evidence supporting that all three differ from one another. No suggestion and suggestion ($t(35) = 4.70$, $p < .001$, $M_{diff} = 1.40$, $d_z = 0.78$, $B_{H(0, 0.86)} = 1.34*10^3$, RR[0.11, 407]), no suggestion and volition ($t(35) = 2.64$, $p = .012$, $M_{diff} = 0.76$, $d_z = 0.44$, $B_{H(0, 0.86)} = 11.82$, RR[0.16, 5.04]), and suggestion and volition conditions ($t(35) = 2.77$, $p = .009$, $M_{diff} = 0.64$, $d_z = 0.46$, $B_{H(0, 0.86)} = 14.36$, RR[0.12, 5.47]).

Exploration 2: Is there any relationship between the subjective responses of the participants and the extent to which they reduced Stroop interference? Finally, we explored whether or not the subjective ratings of meaninglessness and objective meaninglessness (reduction in Stroop interference) are related. We conducted the analysis of regression slopes where the predictors were the difference scores of subjective ratings (either the difference of no suggestion and suggestion, or the difference of no suggestion

and volition) and the outcome variable was the reduction in Stroop interference wither by the suggestion or by the volitional request. We ran regression analyses for all four subjective meaninglessness items, and modelled the predictions of H1 (asserting a positive relationship) with a half-normal distribution with a mode of zero and SD of 0.6. This latter value was obtained by dividing the maximum expected word blindness effect (60ms) with the maximum expected subjective meaninglessness (100%). The analysis of regression slopes revealed good enough evidence for H1 in all cases but one (Bs > 3.17, *b*s > 0.70), supporting the idea that there is a positive relationship between subjective experience of meaninglessness and the extent to which participants reduced Stroop interference. Table 5 present the results of the regression analyses.

Table 3

Summary Table about the Means of the RTs and Self-report Measures in the three Experimental Conditions

		Experimental condition		
Category	Item (scale)	No Suggestion	Suggestion	Volition
Response times (RTs)	Incongruent (ms)	871 (178)	805 (128)	813 (178)
	Neutral (ms)	772 (113)	738 (109)	738 (137)
	Congruent (ms)	717 (111)	699 (118)	704 (141)
Expectations	Expecting the words to be meaningless (0-5)	0.64 (0.76)	2.22 (1.17)	2.33 (0.99)
	Expecting to name the color easily (0-5)	2.56 (1.23)	2.69 (1.21)	2.83 (1.21)
Experienced Control	Control over meaningfulness (0-3)	2.36 (0.72)	1.50 (0.81)	2.11 (0.67)
	Experienced nature of meaningfulness (0-3)	-	1.39 (0.90)	1.78 (1.15)
Depth of hypnosis	Depth of hypnosis during the task (0-5)	0.93 (1.12)	2.33 (1.26)	1.69 (1.45)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Table 4

Summary Table of the four Items Measuring the Subjective Experience of Meaninglessness

Item	Experimental condition		
	No suggestion	Suggestion	Volition
Q1: „Was the meaning of the words on the screen completely clear to you"	85.1% (17.3)	57.1% (28.4)	55.0% (27.1)
Q2: „Were you aware of only an unclear meaning of the words on the screen"	14.6% (19.0)	39.4% (27.9)	42.8% (26.1)
Q3: „Were you just aware of the color and had no idea of what script of the words were written in"	26.0% (30.6)	45.7% (30.3)	46.4% (28.4)
Q4: „Were the words on the screen written in a clear yet meaningless script"	17.4% (25.2)	44.6% (31.8)	42.5% (30.5)

Note. The Standard Deviations (SD) of the means are shown within the brackets.

Table 5.

Results of the Regression Models Predicting Response Times Based on Subjects` Experience of Word Meaninglessness

		Statistics					
Item	Comparison	*b* (ms/%)	β	*t*	*p*	$B_{H(0, 0.6)}$	RR
Q1	Suggestion	0.94	0.38	2.38	.023	4.02	0.51, $6.31*10^3$
	Volition	1.52	0.50	3.32	.002	3.17	0.58, $7.22*10^3$
Q2	Suggestion	0.70	0.28	1.69	.099	4.03	0.51, $6.29*10^3$
	Volition	1.35	0.40	2.53	.016	2.86	0, 0.61
Q3	Suggestion	0.71	0.32	1.97	.057	5.22	0.45, $5.60*10^3$
	Volition	1.32	0.46	3.02	.004	4.11	0.5, $6.21*10^3$
Q4	Suggestion	0.76	0.35	2.15	.039	4.40	0.49, $6.05*10^3$
	Volition	1.49	052	3.59	.001	3.97	0.49, $6.05*10^3$

Note. The raw slopes are indicated by *b*, whereas the standardised effect sizes are indicated by β.

Discussion

The purpose of this pre-registered experiment was to test a crucial prediction of the cold control theory, namely, whether highly suggestible people can produce comparable word-

blindness effect when they voluntarily imagine that words are meaningless and when they respond to a post-hypnotic suggestion to see words as meaningless characters. Importantly, all of the outcome neutral tests of the experimental manipulation were fulfilled. We replicated the classical word-blindness effect, and there was strong evidence supporting the notion that the participants experienced their response less voluntary in the suggestion than in the volition and no suggestion conditions. Moreover, the participants reported that their experience of meaninglessness felt more like perception in the suggestion condition and more like imagination in the volition condition. This latter finding is in accordance with the claim of cold control theory that engaging in imagination without possessing an accurate HOT of the intention to do so should lead to the conscious experience of hallucination. Finally, we did not define the comparisons of the level of expectations as outcome neutral tests since there is a way to partial out their effect if we were to find a difference between the suggestion and volition conditions. Nonetheless, the tests revealed moderate evidence supporting no difference between suggestion and volition for both of the expectancy measures allowing us to rule out the response expectancy theory (Kirsch, 1985; Kirsch & Lynn, 1997) as an alternative explanation if we were to find any difference between these conditions in terms of the magnitude of the word blindness effect.

Cold control theory asserts that hypnosis influences the HOTs of intentions, inducing one to believe that their action is not voluntary while leaving the first order intention itself unaffected. Hence, as long as the first-order intention is sufficient to produce a certain response (e.g., word blindness effect), one should be able to respond equally to a (post)hypnotic suggestion and to a volitional request. The result of the crucial test is in line with this prediction, however, the Bayes factor did not exceed the level of good enough evidence. Currently, there is anecdotal evidence ($B = 0.68$) supporting the model predicting no difference between the suggestion and the volition conditions in the size of the word blindness effect over the model that predicts stronger word blindness effect in the suggestion than in the volition condition. Therefore, we need to suspend scientific judgment and we cannot draw a conclusion whether or not hypnosis is merely a simple change in the metacognition of HOTs of intending. To come to a conclusion, we will need to continue data collection until the Bayes factor exceeds either 1/3 or 3, which are the conventional cut-off of good enough evidence. Nevertheless, we provide the

interpretation of the results as if we found good enough evidence for H0 and then as if we found good enough evidence for H1.

Testing whether there is an association between the extent to which highs reduce the Stroop interference in the suggestion and in the volition conditions can be a severe test of another core assumption of the cold control theory. This test examines whether the underlying mechanism of the response, or in other words the strategy used by the subjects, is the same when responding hypnotically and voluntarily. Although this analysis was defined as an exploratory one in the pre-registration with the aim to help the sample size estimation of a future study, the results of the analysis are in line with the existence of a strong relationship and it deems implausible that the relationship would be negligible or weak. We can be 97.5% confident that the population effect size is not smaller than .48 (the result of the pilot study is essentially identical with the one of the pre-registered study, r = .74, 95% CI [.52, .87]). This finding further corroborates the idea that the subjects were indeed engaged in the same strategy in the suggestion and the volition conditions, while they responded hypnotically in the former and voluntarily in the latter.

Obtaining good enough evidence for the absence of a difference between suggestion and volition conditions. This finding would imply that the simplest form of cold control theory can account for the hypnotic phenomenon. That is, the sole difference between a hypnotic and a non-hypnotic (or voluntary) response is the form of the accompanying HOT of the intention. Hence, unconscious intentions are no more efficient than conscious ones, or more generally speaking, HOTs have limited or no causal power over first-order states (Rosenthal, 2008). Nonetheless, this finding still begs the question of what is the strategy that highs are using to reduce the Stroop interference effect. For instance, it has been demonstrated that highs can still reduce the Stroop effect when their ability to blur their vision was disrupted and they could not simply look away from the screen to blunt their visual input (Raz et al., 2003). In a recent study of ours, we also found evidence supporting the notion that none of these strategies can be the underlying mechanisms of the word blindness suggestion (Palfi, & Dienes, 2019). We demonstrated in two experiments conducted outside of the hypnotic context that the subjects could diminish the extent of the Stroop effect via looking-away or blurring their vision, however, the RT pattern of the strategies did not follow the RT pattern of the suggestion. A defining characteristic of the word blindness suggestion is that it decreases the interference effect as well as the RTs of incongruent trials, yet, these strategies failed to

reduce the RTs of the incongruent trials. Moreover, we tested two additional candidate strategies in the first experiment: internal rehearsal of the target of the task ("displayed color") and focusing only on a single letter of the presented word (for comparable strategies used in and out of hypnosis see Sheehan, Donovan, & Macload, 1988). The former strategy may improve task performance through enhancing goal-maintenance (De Jong, Berendsen, & Cools, 1999), whereas the latter one may facilitate the resolution of response conflict analogously to the single-letter coloring design, in which coloring solely the last letter of the Stroop word induces a smaller Stroop interference effect than coloring one of the letters in the middle of the word (Besner, Stolz, & Boutilier, 1997; for a review see Flaudias & Llorca, 2014). The participants could not benefit from using any of these strategies and did not reduce either the Stroop interference effect or the RTs of the incongruent trials rendering these strategies unlikely to be the underlying mechanism of the word blindness effect. Highs might follow the instructions of the suggestion and volitional request tightly and the strategy that they use may be imagination itself. Perhaps, imagining the words as being meaningless is enough by itself to reduce Stroop interference. For example, an imaginative strategy like this would only require subjects to entertain a counterfactual model of the world in which words are meaningless. This model might reset top-down cognitive control and facilitate the resolution of conflict. The results of Exploration 2, in which we found good enough evidence for a positive relationship between the subjective experience of meaninglessness and reduction of Stroop interference, are in line with this hypothesis. The more frequently subjects experienced meaninglessness when they responded to the suggestion or to the volitional request, the more efficiently they alleviated Stroop interference. Nonetheless, further research is needed to provide an experimental test of this idea, and it would need to be demonstrated that lows and mediums can reduce Stroop interference by creating and entertaining a counterfactual model of reality.

A seminal theory of hypnotic responding, the response expectancy theory, is also in accordance with the main findings of the pre-registered experiment. Interestingly, this theory is not part of the metacognitive class of theories as it assumes that the creation of a hypnotic response does not require the involvement of intentional cognitive control processes. The mere existence of the expectancies are enough to produce the responses and these responses are accompanied by the feeling of involuntariness by virtue of not being intentional. Importantly, cold control theory and the response expectancy theories

are not mutually exclusive, hypnotic responses might be generated in multiple ways, and the current experiment did not aim to distinguish between the two theories. Finally, the special process theories of hypnosis that presume the existence of processes other than the relinquished metacognition over one`s intentions, expectations and beliefs are not in accordance with these findings. For instance, the integrative cognitive theory of hypnosis (Brown & Oakley, 2004) and many of the dissociation theories, such as the early version of the neodissociation theory (Hilgard, 1977), the dissociated control theory (Bowers, 1992; Woody & Bowers, 1994) and the integrative model of dissociation theories (Woody & Sadler, 2008) surmise differential outcomes for responses to suggestions and volitional requests. In a nutshell, the results corroborate the cold control class of theories of hypnosis and disconfirm the non-metacognitive, special process theories that surmise that hypnosis can give rise to objective responses that cannot be induced via non-hypnotic means.

Retaining the simplest form of cold control theory has clear and profound implications for the clinical application of hypnosis. Strictly speaking, applying hypnosis on its own or using it in tandem with therapeutic approaches, such as cognitive behavioral therapy (Kirsch, Capafons, Cardeña-Buelna, & Amigó, 1999) or suggestions (e.g., indirect suggestions of Erickson and his colleagues [1976]) should produce the same outcome as an appropriate control technique. According to cold control theory, a technique can be deemed appropriate control when the sole difference between the hypnosis and the control technique lies in the nature of the accompanying HOT of the behaviors of the clients engaged in the therapy. Meta-analyses concluding that the application of hypnosis has beneficial effects usually compare hypnosis to standard care groups, which is a mixture of various groups such as no treatment, standard treatment as well as waiting-list controls. However, the advancements in the outcome measures tend to disappear once hypnosis is compared to a therapeutic control group, which is a more adequate control than no treatment or standard treatment (e.g., in irritable bowel syndrome [Schaeafert, 2014] and chronic pain studies [Adachi, 2012]). Interestingly, there is evidence that hypnosis can improve the efficacy of some cognitive-behavioral treatments when the sole difference between the instructions of the hypnotic and non-hypnotic groups was the usage of the term hypnosis in the former (Kirsch, Montgomery, & Sapirstein, 1995). Nonetheless, as argued earlier in this paper, the inclusion of the word hypnosis in the instructions can give rise to many confounding factors through elevating

expectations and motivation.[26] One needs to carefully phrase the instructions used in the control group to make sure that people believe that their voluntary response can be just as efficient as their hypnotic one. Once we have ensured this, we can expect that the clients will benefit just as much from the non-hypnotic than from the hypnotic treatment.

Obtaining good enough evidence for the model predicting larger word blindness effect in the suggestion than in the volition condition. Replicating the findings of the pilot study would imply that cold control theory needs to be revisited as the theory`s prediction that a response to a volitional request should be just as efficient as a response to a (post)hypnotic suggestion would be disconfirmed. Retaining the core assumption of the theory that hypnosis targets the HOTs of intentions, the next simplest model is a one, in which hypnosis can facilitate first-order intentions via altering the awareness over one`s first-order intention. Preserving this assumption is not unfounded as the result of the correlational analysis is in line with the notion that highs were engaged in the same strategy in the volition and suggestion conditions. Some unexpected implications follow from the revised version of cold control theory and the findings of this study: it enables HOTs to have a causal role and influence first-order states (cf. Rosenthal, 2008), and it demonstrates an example in which unconscious control is more efficient than the conscious one (cf. Cleeremans, 2006). These implications raise the question of how does the HOT of the intention influence the execution of the first-order intention. According to cold control theory, highs possessed the first order intention to create the experience of the script as being meaningless in both of the conditions. In the volition condition, as we have found it, it leads to the experience of being involved in imagination as highs entertained multiple models of the world: one in which meaning is accessible and a counterfactual one in which it is not (Perner, 1991). In the suggestion condition, however, highs were not aware of their intention to imagine the words as meaningless and so they perceived the words as meaningless, in other words, they had a hallucination like experience. Perhaps, responding hypnotically makes one sufficiently unaware of the real model of the world easing it to focus only on the counterfactual model. Whereas responding voluntarily requires one to entertain two conflicting models that might hinder the efficiency with which one can focus on the counterfactual one.

[26] A similar point is made by Lynn et al. (2019) and Terhune et al. (2017) in their recent papers in which they reviewed the strength of the empirical evidence underlying the efficiency of hypnotherapies.

Incidentally, these findings challenge the simplest form of the response expectancy theory (Kirsch, 1985) that asserts that responses to hypnotic suggestions can be fully accounted for expectations and beliefs. Observing smaller Stroop effect in the suggestion than in the volition condition indicates that responding to the word blindness suggestion must have involved processes other than the formation of expectations. According to cold control theory, this additional step is the engagement in intentional cognitive control processes and it appears that under specific circumstances unconscious control processes can be more efficient than conscious ones. Frankly, the modified version of cold control theory is not the only one that is consistent with these findings. Several other theories of hypnosis, which are not part of the metacognitive class of theories, are in line with the current results showing the superiority of the suggestion condition over the volition one. For instance, the integrative cognitive theory of hypnosis (Brown & Oakley, 2004) and some of the dissociation theories, such as the early version of the neodissociation theory (Hilgard, 1977) and the integrative model of dissociation theories (Woody & Sadler, 2008). We argue, however, that accepting the explanation of the modified version of cold control theory is the next logical step as it is the most parsimonious explanation provided by the theories that are in accordance with the results.

The findings of this study have far-reaching implications for the application of hypnosis in the clinical context. For instance, the revised version of the cold control theory has clear predictions on when should we expect that adding hypnosis to a therapeutic technique is advantageous compared to the non-hypnotic technique. Assuming that highs can focus better on counterfactual models of reality when they respond hypnotically compared to when they respond voluntarily implies that therapies that involve imagination can be improved (at least for highs) when they are applied in tandem with hypnosis. Imagination is an essential part of many of the cognitive-behavioral treatments that generally consists of cognitive strategies, imagery or systematic desensitization (Kirsch, Montgomery, & Sapirstein, 1995; Spinhoven, 1987). A meta-analysis of 18 studies with various problems and cognitive-behavioral treatments to resolve them revealed evidence favoring the adjunct role of hypnosis (Kirsch, Montgomery, & Sapirstein, 1995). Nonetheless, these studies did not control for potential confounds such as the effect of lifted expectations and motivation in the hypnotic compared to the non-hypnotic group, hence, further research is needed to test the prediction of cold control theory about the superiority of treatments used in tandem with

hypnosis over treatments without hypnosis. For instance, hypnosis might facilitate treatments of chronic pain that involve the imagery of a pleasant feeling (e.g., warmth) to distract one from the unpleasant feeling, or treatments aiming to alleviate phobias using systematic desensitization. Imagining the subject of one`s phobia may feel more real when it is experienced as perceived during hypnosis, which might lead to a more successful desensitization and adaptation. The application of hypnosis could even improve treatments mitigating addictive behavior, such as smoking. Being immersed in the picture how one`s lung might look like after life-long smoking could create stronger motivation to stop smoking when one is doing it hypnotically compared to doing it non-hypnotically. Nevertheless, many other therapeutic exercises do not require subjects to engage in a counterfactual model of the world, and according to the revised version of cold control theory, these techniques should not be catalysed by hypnosis. For example, physical anchors (e.g., touching first and second finger together) are posthypnotic suggestions that can help subjects implement in their everyday life what they have learned during the therapeutic session (Lynn & Kirsch, 2006). An anchor can, for instance, reduce craving for smoking by facilitating access to one's reasons for becoming a non-smoker. When an anchor is used hypnotically, one can recall one`s reasons in a way that is experienced involuntary, however, according to cold control theory, one should be able to recall these reasons just as well in a non-hypnotic way by using a non-hypnotic anchor.

Experience of hypnotic depth. Although it was not related to the main point of the current paper, we aimed to explore whether highly suggestible people report the experience of hypnotic depth while they respond to a post-hypnotic suggestion. In principle, post-hypnotic suggestions are preferred over hypnotic ones in experimental settings as they can control for the influence of the hypnotic induction (e.g., Terhune, Luke & Cohen Kadosh, 2017). However, participants might interpret the situational needs otherwise and they still create the experience of being in a hypnotic state. Our findings are partly in line with this latter assumption. We observed that the participants report stronger hypnotic depth experience in the suggestion than in the no suggestion condition and that the plausible effect size lies within the range of 1.91 – 2.76 on a scale from 0 to 5 (where 1 means slightly hypnotized, and 5 indicates deeply hypnotized). Unfortunately, we did not gauge the level of hypnotic depth during the induction and after the de-induction, which could have provided us with baselines on the scale (this was pre-registered, however, we failed to include it in the procedure). Nonetheless, we know from

another project of ours in which we also used the word-blindness suggestion that the level of hypnotic depth is substantially smaller in the suggestion condition than during the induction, and more importantly, we did not find evidence that hypnotic depth is stronger during the execution of the suggestion than right after the de-induction (Palfi, Parris, Seth & Dienes, 2018). Taken together, these findings make it plausible that at least some of the subjects interpret the context as they need to create the experience of being in a hypnotic state while responding to the posthypnotic suggestion. However, it is safe to assume that even if they do so they do not achieve the same level of hypnotic depth as they would during the hypnotic induction.

Conclusion

The cold control theory, which is perhaps the simplest theory of hypnotic responding, asserts that hypnosis is a metacognitive phenomenon and the sole difference between a response to a hypnotic suggestion and a volitional request lies in the nature of the accompanying HOT of the first-order intention. In a pre-registered multisite experiment, we tested a key prediction of the theory that objective responses (i.e., the word blindness effect) to suggestions and volitional request should be identical (after controlling for the effect of expectations). The strength of the evidence is not good enough to draw conclusions from the critical test of the theory. However, anecdotal evidence suggest that highs reduce the Stroop interference to the same extent when they respond to the word blindness suggestion and when they respond to its volitional request counterpart. These results settle the issue about the influence of hypnosis on first-order states in favour of the cold control theory. Several implications follow from this conclusion: the notion that HOTs have limited or zero causal role is corroborated (Rosenthal, 2008). The clinical application of hypnosis as an adjunct to therapies should be revisited: according to cold control theory, hypnosis cannot increase the successfulness of a therapeutic treatment beyond the effect of elevated expectations and motivation. Nevertheless, to draw strong conclusions the continuation of data collection is imperative.

Chapter V: Can hypnotic suggestibility be measured online?

Introduction

Hypnosis and hypnotic suggestions have been shown to be useful experimental tools to test theories of cognitive neuroscience (Oakley & Halligan, 2013; Raz, 2011), especially theories related to consciousness (Cardeña, 2014; Terhune, Cleeremans, Raz, & Lynn, 2017). For instance, hypnotic suggestions can evoke changes in the feeling of voluntariness (Weitzenhoffer, 1974, 1980) or modify one`s sense of agency (Haggard, Cartledge, Dafydd, & Oakley, 2004; Lush et al., 2017; Polito, Barnier, & Woody, 2013). Responses to suggestions frequently involve alterations in perception, such as the experience of positive and negative hallucinations or delusions (Kihlstrom, 1985; Oakley & Halligan, 2009). Moreover, hypnotic suggestions can be employed to simulate some properties of neurological and psychiatric conditions in healthy subjects (Barnier & McConkey, 2003; Oakley, 2006). Finally, correlations between hypnotisability and measures employed by consciousness researchers (e.g., the rubber hand illusion; the vicarious pain questionnaire; mirror touch synaesthesia) have recently been found (Lush et al., 2019). These correlations suggest that measures common in the consciousness literature are driven by hypnotic suggestibility. There is therefore an increasing need for an expansion of hypnosis research. Unfortunately, the successful application of hypnotic suggestions demands plenty of resources, making it impractical for researchers to run large-scale hypnosis related studies. In order to conduct experiments involving hypnosis, researchers generally need to recruit from a specific subsample of people based on their tendency to respond to hypnotic suggestions. To achieve this, researchers run hypnosis screening sessions before recruitment, so that, for example, they can identify the participants at the lowest and highest end of the scale (low and highly hypnotisable people, respectively). High and low hypnotisability are usually defined as the top and bottom 10%-15% of screening scores (Barnier & McConkey, 2004; Anlló, Becchio & Sackur, 2017). Therefore, screening procedures are time-consuming; to identify a single highly suggestible participant for an experiment, one has to find, on average, ten people who are willing to undertake a screening that can last from 40 up to 90 minutes depending on the applied method.

The hypnosis screening procedure has moved through a long developmental process in which it has become more and more user friendly. Initially, the screening

consisted of two steps, a preliminary group session applying the Harvard Group Scale of Hypnotic Susceptibility Form A (HGSHS:A; Shor & Orne, 1963) and an individual session using the Stanford Hypnotic Susceptibility Scale Form C (SHSS:C; Weitzenhoffer & Hilgard, 1962) conducted with only those scoring very high or low in the first session. The later development of a reliable group screening method, the Waterloo-Stanford Group Scale of Hypnotic Susceptibility (WSGC; Bowers, 1993), has drastically mitigated the time required for screening as it allows researcher to screen up to a dozen people in about 90 minutes (although it was originally intended to act as a second screen after an HGSH:A, a single screen with the WSGC is quite reliable enough to select subjects capable of later having compelling subjective responses to difficult suggestions, e.g. digit-colour synesthesia, Anderson, Seth, Dienes, & Ward, 2014, or compelling objective reductions in Stroop interference to alexia (word blindness) suggestions, e.g. Parris, Dienes, Bate, & Gothard, 2014). Recently, the Sussex Waterloo Scale of Hypnotizability (SWASH; Lush, Moga, McLatchie & Dienes, 2018) was introduced, which is a modified version of the WSGC. The SWASH includes new items to measure the subjective experiences of the participants (compare also the Carleton University Responsiveness to Suggestion Scale [CURSS, Spanos, Radtke, Hodgins, Stam, & Bertrand, 1983], the Creative Imagination Scale [CIS, Wilson & Barber, 1978], and the Experiential Scale for the WSGC [Kirsch, Milling, & Burgess, 1988]). The length of the procedure was reduced to 40 minutes and it can be run with larger groups than the WSGC (Lush et al., 2018). Moreover, the dream and age regression suggestions were not included in the SWASH. These highly personalized items of the WSGC can be risky by virtue of possibly triggering unpleasant memories or emotions (Cardeña & Terhune, 2009; Hilgard, 1974).

Nonetheless, the application of the least demanding methods (such as the SWASH, the CURSS or the CIS), still requires potential participants to attend a group session, which makes the screening procedure relatively time consuming and limits the subject pools to psychology students who are the easiest to incentivise to participate in a group screening on campuses. These two barriers of large-scale hypnosis studies could be overcome by employing fully automatized, online hypnosis screening procedures. In the last two decades, psychological science has witnessed growth in the application of online data collection for experimental purposes, paving the way for researchers to collect large samples in a short period of time (Reips, 2000; though it can come with its own

problems, e.g. Dennis, Goodson & Pearson, 2018). In order to adapt the hypnosis screening procedure online, one needs to ensure that the non “live” version can induce similar objective and subjective hypnotic responses as with a “live” hypnotist. Indeed, suggestibility scores of participants are comparable when the hypnotic induction and suggestions are delivered by a pre-recorded audiotape and when they are delivered by an experimenter (Barber & Calverley, 1964; Fassler, Lynn, & Knox, 2008; Lush, Scott, Moga, & Dienes, 2018). These findings underpin the idea that the participants could easily undergo a hypnosis screening procedure in their own rooms by listening to a pre-recorded script and filling out the booklets online. Nevertheless, online data collection has its own perils, namely, the data acquired by online questionnaires might not be as reliable and the results might not be consistent with the ones of the traditional data collection procedures (Krantz & Dalal, 2000). Therefore, the reliability of new online questionnaires, such as the online version of a hypnosis screening procedure, needs to be tested even if there is evidence that the quality of the data and the findings of online based studies can be similar to those obtained by traditional methods (Gosling, Vazire, Srivastava, & John, 2004; Buhrmester, Kwang, & Gosling, 2011).

In this project, our purpose is to explore the extent to which an online hypnotic screening procedure is reliable and consistent with an offline procedure. To this aim, we measured people`s hypnotic suggestibility with the SWASH on two separate occasions and in two different environments. Henceforth, we call every type of data collection carried out in a controlled environment with the experimenter present an offline screening, whereas undertaking a hypnotic screening alone in one's own room under one’s own control will be called online screening. In addition, we are interested in the extent to which the length of the delay between first and second screen can influence the reliability and the scores of hypnotic suggestibility. The question about the stability of hypnotic suggestibility over periods of few days or even decades have inspired various research projects (e.g., Fassler, Lynn, & Knox, 2008; Lynn, Weekes, Matyi, & Neufeld, 1988; Piccione, Hilgard, & Zimbardo, 1989). To assess the stability of hypnotic suggestibility, we recruited half of the sample from the subject pool of the year of 2016 and the other half from the year of 2017, both of whom have already received offline screening. Therefore, for some of the participants, the delay between the two screenings is not more than 6 months (short delay group), whereas for the others, it is at least one and a half years (long delay group). For practical reasons, the first screening was organised offline, in

groups of 20-40 for all the participants, whereas the second screening was either an online screening or another offline one. By this method, we are able to estimate how strongly the type of the screening and the length of the delay can influence the suggestibility scores of the people; we can also assess their influence on the test-retest reliability and the validity of the screening. Taken together, this project strives to explore whether a well-established offline screening procedure could be replaced for practical purposes by an online version, which could help consciousness researchers run more and larger hypnosis studies by drastically cutting the recruitment related costs.

While responding to hypnotic suggestions, people tend to experience as of being in some form of trance or altered state (Kihlstrom, 2005; Kirsch, 2011). This experience is usually measured by subjective reports of depth of hypnosis (e.g., Hilgard & Tart, 1966), which is, interestingly, strongly associated with people`s ability to respond to hypnotic suggestions (Wagstaff, Cole, & Brunas-Wagstaff, 2008). We investigate this link by assessing the strength of relationship between hypnotic suggestibility scores and depth of hypnosis reports, and the extent to which the mentioned experimental manipulations can influence this relationship. We also aim to evaluate the extent to which depth of hypnosis is influenced by the type of data collection and the length of the delay between screens to ensure that people experience comparable level of hypnotic depth during online and offline screens.

In our analyses, we solely employed estimation procedures instead of testing the existence of differences with an inferential statistical tool such as the null-hypothesis significance test (Fisher, 1925; Neyman & Pearson, 1933) or the Bayes factor (e.g., Dienes, 2011; Rouder et al., 2009). Estimation is recommended over inferential statistics when the existence of a difference is established or it is not relevant (Jeffreys 1961; Wagenmakers et al., 2018). The second point proves to be decisive for our case, since it is not necessary to test the existence of any investigated effect to answer our research questions. For instance, the core aim of the current project was to conclude regarding the applicability of online hypnosis screening by comparing the SWASH scores, the reliability and the validity of online and offline hypnosis screening. Imagine a scenario in which an inferential statistical tool demonstrates evidence for the difference between the offline and online groups in favour of the offline group in all aspects that assess the quality of the measurement. Importantly, this outcome per se cannot give a definite answer to our central question as the mere fact that offline screening is significantly better than online

screening neglects the question of magnitude of the difference. To reject or accept the idea that online screening is viable, we need to know the extent to which the quality of offline and online screening differs so that we can decide whether the benefits of the online screen outbalance its costs. Further, the fact that the two types of screening will correlate cannot be in doubt; the question is simply the strength of the relationship between them.

To explore the range of plausible effect sizes, estimation methods, either from the Bayesian (Kruschke, 2010, 2013; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Wagenmakers, Morey, & Lee, 2016) or from the frequentist school (Cumming, 2014) can be used. Here, we applied a Bayesian tool, estimation by calculating the 95% Bayesian Credibility Intervals, as this is the method that is appropriate to answer our research question; namely, how confident can we be that the true effect size lies within a specific interval (Wagenmakers et al., 2018). Only Credibility Intervals allow us to make claims such as that the true value of the effect size is probably not larger or smaller than a particular value.

Methods

Participants

Psychology students at the University of Sussex participated in an offline hypnosis screening as part one of their modules during the first semester of their studies. We recruited psychology students who had started their BSc studies in the year of 2016 or 2017 and who had provided their contact information in an offline hypnosis screening session. Both subject pools consisted of around 300 students and we randomly assigned half of them to each experimental group (experimental groups described below). Thus, we invited around 150 people for each group. We continued data collection until the end of the spring semester of 2018. In the second session, 73 students participated. However, we could not trace back the data of two students to their first session results and so we needed to exclude them from all of the analyses, leaving us with 71 participants in total. Twenty-six students attended the offline session (23 females, M_{age} = 19.7, SD_{age} = 1.8) and 45 students completed the screening online (41 females, M_{age} = 21.0, SD_{age} = 5.3).

We informed each participant about the nature of the study and only those students were be able to attend who agreed to the terms and conditions of the study. After finishing

the experiment, the participants were debriefed and received a payment of £5 or course credit. The study has been approved by the Ethical Committee of the University of Sussex (Sciences & Technology C-REC).

Materials

One of the authors produced the audio recording of the hypnosis procedure (induction and the suggestions); the length of this recording was 28 minutes. The questionnaire applied in the first session for data collection was created in MatLab (MathWorks, 2016), whereas the questionnaire that was used in the second session was a PHP based website. The PHP script, the materials and the documentation on how to install the software can be accessed at https://osf.io/6twdp/.

Measures

The measures introduced below were utilised in the first occasion of the data collection. The second occasion only included the assessment of the hypnotic suggestibility measured by the SWASH regardless of the type of the session (offline or online). Note that, although several questionnaires were registered along with the first screening, we only used the suggestibility scores of the participants in this project (see our research questions in the last paragraph of the Introduction).

SWASH. The hypnotisability of the students was measured by the SWASH. This scale is a modified version of the WSGC (Bowers, 1993) which contains 10 suggestions and corresponding items measuring objective suggestibility and the subjective experiences of the participants about each suggestions.

Data collection in 2016. As part of the first session in 2016 the following four questionnaires were registered: (a) Barratt Impulsiveness Scale (BIS-11), which consists of 30 items and measures people`s tendency to behave impulsively (Patton & Stanford, 1995); (b) Free Will Inventory (FWI), which includes 29 items measuring people`s beliefs about free will and their relationships with these beliefs (Nadelhoffer, Shepard, Nahmias, Sripada, & Ross, 2014); (c) Short Form of the Five Facet Mindfulness Questionnaire (FFMQ-SF), which is a 24-item scale assessing the mindfulness skills of individuals via self-report (Bohlmeijer, ten Klooster, Fledderus, Veehof, & Baer, 2011); (d) Dissociative Experiences Scale-II (DES-II), which is a 28-item self-report questionnaire developed by Bernstein and Putman (1986).

Data collection in 2017. In 2017, we administered the following four questionnaires in the first data collection session of: (a) a 15 minutes long breath counting exercises based on Study 2 of Levinson, Stoll, Kindy, Merry & Davidson (2014); (b) the Mindful Attention Awareness Scale (MAAS) consisting 15 Likert scale items (Brown & Ryan, 2003); (c) the Schizotypal Personality Questionnaire-Brief (SPQ-B), which consists of dichotomous questions (Raine & Benishay, 1995); (d) the DES-II that was used in 2016.

Design

We employed a 2*2*2 mixed design. The within subject variable is the date of the data collection (first session vs. second session). The between subjects independent variables are the form of the second hypnosis screening session (offline vs. online) and the length of the delay between the first and the second sessions (short delay [few months] vs. long delay [more than a year]).

Procedure

There were three forms of data collection: 1) group sessions at the university with the experimenter present (first, offline screen) 2) individual sessions in a small experimental room at the university with the experimenter present (second, offline screen) 3) individual sessions at home (second, online screen,). All of the participants engaged in the first, offline, screen and later they were invited to attend in a second screen that was either offline or online. The procedure of the screening was identical in each case and followed the steps below.

After providing informed consent, the participants had the opportunity to provide contact details for a database in case they were willing to participate in hypnosis related research in the future. Next, they were asked to adjust the volume of their headphones until it was moderately loud by listening to a test tune. Before starting the hypnotic induction procedure, they were notified that the whole procedure would last about 45 minutes and that they should not take a break. By pressing the start button, participants ran the hypnotic induction and suggestions. After the de-induction, participants were asked to fill out the SWASH response booklet, rating their response to each suggestion. Finally, the participants were thanked for attending and debriefed.

Data analysis

Data transformation. We computed the Objective and Subjective suggestibility scores of the participants as described in the SWASH manual (Lush et al., 2018) and then we doubled all subjective scores so that both of the objective and subjective scores fell between 0 and 10. By taking the weighted average of these derived scores, we calculated the composite SWASH score of each participant, which was used in the majority of the analyses. For more details on the calculation of the SWASH scores, see Lush et al. (2018; manual available at https://osf.io/wujk8/). Given that the distributions of the objective, subjective and composite SWASH scores of the first screen were all fairly normal (see Figure 1 in Preregistration), we assumed that the dataset of the second screen was also normally distributed. Therefore, we planned to use parametric methods to estimate the strength of correlation between continuous variables (Pearson`s *r*).

Bayesian estimation. In this project, we estimated the population effect sizes and did not test hypotheses. Thus, here, we report the estimates (e.g., mean or correlation) and the 95% Bayesian credibility intervals (CI) applying a uniform prior distribution. Note that, although, the bounds of the CIs are numerically equal to the bounds of the confidence intervals (assuming a uniform prior), their interpretation is different (e.g., Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

Implementation of the preregistration

The design and research questions of this study were preregistered at osf.io/3abje. In order to ensure the reproducibility of the analysis and decrease analytic flexibility, we preregistered an analysis script, written in R (R Core Team, 2016), *a prior* to data collection. The script includes all of the steps defined in the preregistration and an additional data simulation, which helped us test and debug the script. In this paper, we present the results of analyses that were preregistered in the above-mentioned R script and results of two additional, non-preregistered analyses: 1) test-retest reliability of SWASH scores; 2) correlation between SWASH and depth of hypnosis scores. We deviated from the analysis script in one aspect. The calculation of the 95% CIs of the differences between two correlations was incorrect in the original script due to an issue with back-transformation of Fisher`s *z* values of difference scores to Pearson`s *r* (e.g., Meng, Rosenthal & Rubin, 1992; Olkin & Finn, 1995). Therefore, we used the cocor R

package (Diedenhofen & Musch, 2015), which is based on the approximation method of Zou (2007), to estimate the 95% CIs of the differences between correlations.

Results

SWASH scores

The mean of the composite SWASH scores in the offline group (M = 3.44) was only slightly larger than the mean of the online group (M = 3.13) rendering their difference negligible (M_{diff} = 0.31, 95% CI [-0.59, 1.22]). Crucially, the difference between the groups is unlikely to be larger than 1.22. The difference between the offline and online groups is likely to be negligible or small for both of the objective (M_{diff} = 0.39, 95% CI [-0.58, 1.36]) and subjective subscales (M_{diff} = 0.24, 95% CI [-0.72, 1.19]). Panel A of Figure 1 demonstrates that the distribution of the composite SWASH scores of the offline group is akin to the online group. The density of the data is similar between the groups even around the right tail (top) of the distribution indicating that similar proportion of the participants scored high on the SWASH in the offline and online groups. The mean of the composite SWASH scores was comparable in the short (M = 3.32) and long delay groups (M = 3.10), and the plausible values of their differences vary around zero with a maximum difference of 1.10 (M_{diff} = 0.21, 95% CI [-0.67, 1.10]). Table 1 presents the means and 95% CIs of all groups and comparisons with the composite, objective and subjective scores separately.

Table 1.

The Mean Composite, Objective and Subjective SWASH Scores with 95% CIs Broken Down by the Type of the Second Screen and the Length of the Delay

	Measure		
Group	Composite	Objective	Subjective
Offline	3.44 [2.73, 4.16]	3.92 [3.17, 4.67]	2.96 [2.19, 3.73]
Online	3.13 [2.54, 3.71]	3.53 [2.89, 4.17]	2.72 [2.13, 3.32]
Difference	0.31 [-0.59, 1.22]	0.39 [-0.58, 1.36]	0.24 [-0.72, 1.19]
Short delay	3.32 [2.72, 3.91]	3.89 [3.26, 4.52]	2.74 [2.14, 3.35]
Long delay	3.10 [2.42, 3.78]	3.28 [2.54, 4.02]	2.93 [2.19, 3.67]
Difference	0.21 [-0.67, 1.10]	0.61 [-0.34, 1.57]	-0.19 [-1.12, 0.75]

Note. Values within the squared brackets represent the 95% Confidence Intervals. Data presented in this table are based solely on the second screen.

Validity

The correlation between the objective and subjective subscales of the SWASH was strong for the offline screen (r = .78, 95% CI [.56, .89]) as well as for the online screen (r = .79, 95% CI [.65, .88]) indicating appropriate validity in this respect. The difference between the offline and online screen in terms of the strength of the correlation between the objective and subjective scales was close to zero (r = - .02, 95% CI [-.25, .17]).

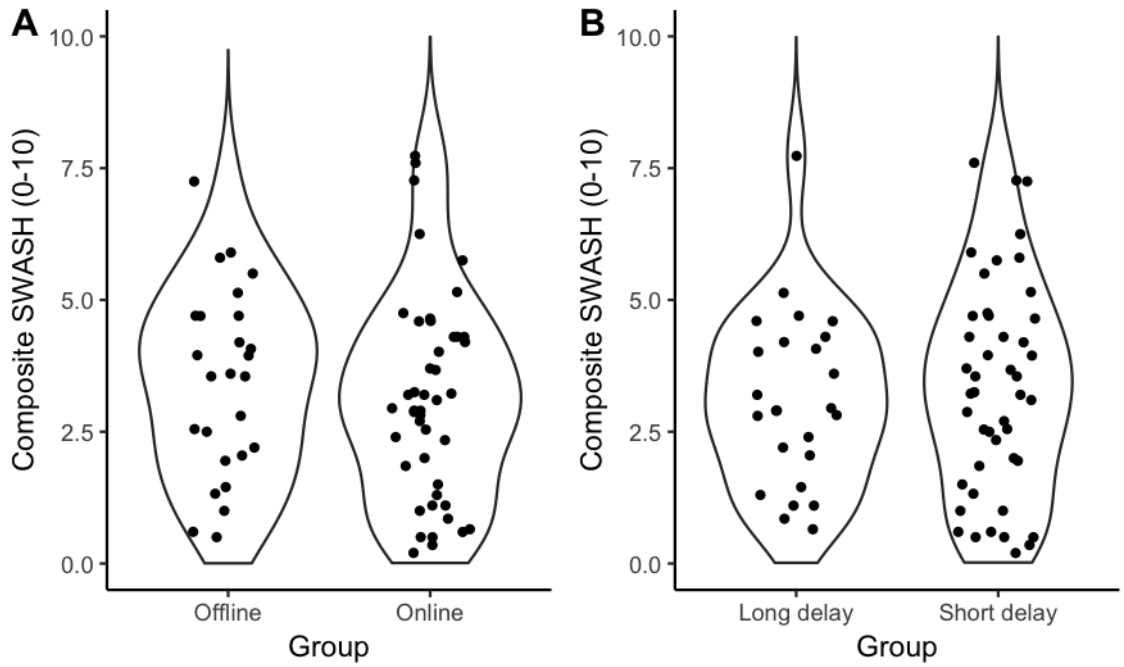

Figure 1. Violin plots depicting the distribution of composite SWASH scores of the second screens broken down either by the type of the screen (offline vs. online, panel A) or by the length of the delay (long vs. short delay, panel B). Each black dot indicates a composite SWASH score of a participant.

Test-retest reliability (non-preregistered)

Correlation between the first and the second screen scores was strong for the subjective subscale but only moderate for the objective subscale irrespective of the type of the screen. For the composite scores, the correlation was strong for the online and offline group as well indicating a good enough test-retest reliability of the SWASH. Interestingly, the test-retest reliability of the online group was possibly higher than that of the offline group, although, only to a small extent (See Table 2 for *r*s and their 95% CIs). The correlation between the first and second screen scores was strong in the short delay group for the subscales as well as for the composite scores. However, the correlation was only

moderate in the long delay group implying that the test-retest reliability of the SWASH is influenced by the length of the delay between the screens from a weak to a moderate extent. Table 2 presents the exact correlation values and their 95% CIs separately for the experimental groups.

Table 2.

Test-retest Reliability of SWASH Scores Broken Down by Type of Screen and Length of Delay

	Measure		
Group	Composite	Objective	Subjective
Offline	.62 [.31, .81]	.43 [.05, .70]	.69 [.42, .85]
Online	.74 [.57, .85]	.59 [.35, .75]	.77 [.61, .87]
Difference	-.12 [-.45, .14]	-.16 [-.57, .20]	-.07 [-.37, .15]
Short delay	.79 [.65, .88]	.65 [.44, .79]	.81 [.68, .89]
Long delay	.55 [.20, .78]	.37 [-.03, .67]	.56 [.21, .78]
Difference	.24 [-.02, .61]	.28 [-.08, .70]	.25 [-.01, .61]

Note. The correlation values are all Pearson`s *r*s and the 95% CIs are reported within the squared brackets.

Depth of hypnosis

Difference between the groups. The participants reported somewhat higher depth of hypnosis scores in the offline (M = 2.15, 95% CI [1.61, 2.70]) than in the online (M = 1.73, 95% CI [1.31, 2.16]) group. Nonetheless, the difference between the groups is not substantial and the maximum plausible value of this difference is 1.10 (M = 0.42, 95% CI [-0.26, 1.10]). The mean of the depth of hypnosis scores in the short delay (M = 1.80, 95% CI [1.35 – 2.26]) compared to the long delay group (M = 2.04, 95% CI [1.59, 2.49]) differed only to a small extent (M = -0.24, 95% CI [-0.87, 0.40]). Figure 2 portrays the distribution of the depth of hypnosis scores broken down by the type of the second screen (panel A) and the length of the delay between the first and second screen (panel B). The depth of hypnosis scores are similarly distributed in the offline and online groups.

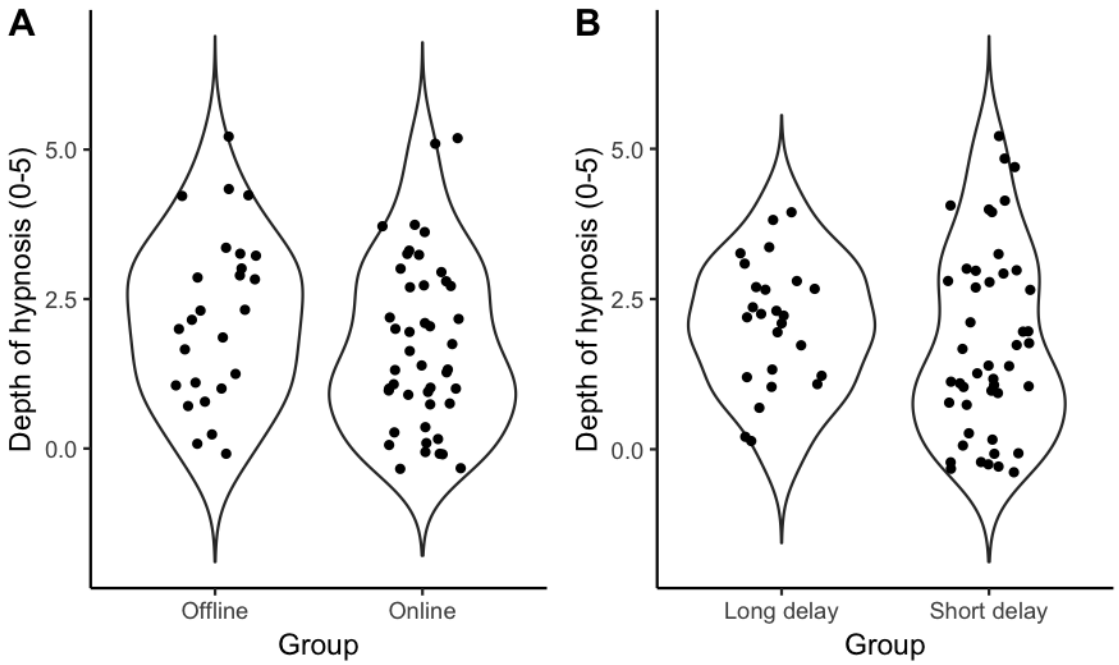

Figure 2. Violin plots representing the distribution of depth of hypnosis scores separately for the offline and online screens (panel A), and for the short and long delay groups (panel B).

Correlation between SWASH and depth of hypnosis scores (non-preregistered). The correlation between the SWASH and depth of hypnosis scores was strong for all but one measure in the online and for all in the offline screen group (all $r >$ .54). The strength of the correlation is unlikely to be larger than .21 in the offline group than in the online group rendering the difference between the two groups minimal. There was strong correlation between the depth of hypnosis scores and all measures in the short delay group (all $r >$ .70), and the correlations were moderate to strong in the long delay group (all $r >$ .31). The difference between the two groups for the strength of the correlations was weak to moderate, and it was the highest for the objective scores. Table 3 shows all of the correlation values and their 95% CIs separately for the experimental groups and for all of the measures.

Table 3.

Correlation Between SWASH and Depth of Hypnosis Scores Broken Down by the Type of Screen and the Length of Delay

	Measure		
Group	Composite	Objective	Subjective
Offline	.76 [.53, .89]	.66 [.37, .84]	.77 [.54, .89]
Online	.70 [.52, .83]	.54 [.29, .72]	.81 [.67, .89]
Difference	.06 [-.21, .28]	.12 [-.22, .43]	-.04 [-.28, .15]
Short delay	.79 [.65, .88]	.70 [.51, .82]	.83 [.71, .90]
Long delay	.55 [.20, .78]	.31 [-.10, .63]	.70 [.41, .86]
Difference	.24 [-.03, .60]	.38 [.02, .81]	.13 [-.07, .42]

Note. The correlation values are all Pearson`s *r*s and the 95% CIs are reported within the squared brackets.

Discussion

The purpose of the present study was to explore whether online hypnosis screening is feasible as the adaptation of this method could ease the recruitment related costs of hypnosis research. To this aim, we estimated the extent to which offline and online hypnosis screening scores, measured by the SWASH, are comparable. The results revealed that the difference between offline and online groups was small to negligible in all aspects and, importantly, applying online rather than offline screening is unlikely to reduce the composite screening score by more than 1.22 and the objective score by more than 1.36 out of ten. To put these effect sizes in perspective, for instance, a recent meta-analysis of four studies investigating the influence of standard induction procedures on suggestibility found that, on average, people score 1.46 higher (out of ten) on scales assessing objective responses to suggestions if they had received a priori induction compared to no induction (Martin & Dienes, 2018). Moreover, the average SWASH score in the online group was comparable to the result of an earlier screen conducted in group sessions at the same university (Lush et al., 2018). Finally, it is not only the average scores in the online group that can be deemed acceptable, the distribution of SWASH scores were also akin in the offline and online groups even at the positive end of the scale. This implies that some people can successfully respond to many suggestions when they undertake an online screening (See Figure 1). None of this was obvious before the data were collected.

The correlation between objective and subjective scores was strong for both of the offline and online groups; crucially, the correlation in the online group can only be as small as .65. This indicates that the validity of the SWASH remained acceptable even with online data collection. Moreover, the strength of the correlation between the subjective and objective components of the SWASH found by Lush et al (2018) was .70, which is consistent with our results. The strength of the correlation between SWASH scores of the first and second screens was medium in the offline and strong in the online group. The lower bound of the 95% CI in the online group was .57 implying that the test re-test reliability of the online measurement is adequate. These values are also appropriate in relative terms. For instance, Fassler et al. (2008) employed the CURSS which has an objective and a subjective subscale such as the SWASH, in two occasions and the test re-test correlations were .59 and .77 for the objective and subjective components, respectively. These results are in line with the correlations found by us in the online group. Overall, the psychometric properties of online screening were excellent; the quality of data collected online has shown to be consistent with the quality of offline data gathered within this study and as part of earlier studies with the SWASH and other hypnosis screening tools.

Modern theories of hypnosis advocate the notion that all hypnosis is self-hypnosis, since the hypnotic subject is the one who actively responds to the suggestions and creates the requested experience (Kihlstrom, 2008; Raz, 2011). This does not mean, however, that the experimenter has no influence on the responsiveness of the subject. For instance, the presence of an experimenter can be helpful in building up a rapport and facilitating responsiveness of the participants (e.g., Gfeller, Lynn, & Pribble, 1987). Nonetheless, the experimenter can also bias the responses of the subjects (e.g., Barber & Calverley, 1966; Troffer & Tart, 1966), and importantly, this level of bias can strongly vary across participants as it is almost impossible to deliver the induction and suggestions in an identical way multiple times. Therefore, the application of fully automatized screenings, such as the online version, can subserve the standardization of the assessment of hypnotic suggestibility.

Introducing online hypnosis screening would markedly decrease the amount of time experimenters need to invest to find participants for their studies. However, to complete a screening procedure, the participants still need to spend 45-60 minutes without taking a break; otherwise, the data would be not usable for recruitment purposes. A

substantial part of the screening is assigned to the standard hypnotic induction, which consists of various suggestions mostly to relax; however, the responses to these suggestions are not assessed directly during the screening (e.g., Shor & Orne, 1963; Weitzenhoffer & Hilgard, 1962). Would it be feasible to exclude the standard induction from the screening procedure to save time for the participants? Cognitive theories of hypnosis, such as the cold control theory (Barnier, Dienes, & Mitchell, 2008; Dienes & Perner, 2007), emphasise the role of the feeling of involuntariness in differentiating hypnotic from non-hypnotic responses. This feeling is also known as the "classical suggestion effect" (Weitzenhoffer, 1974, 1980). Therefore, according to cold control theory, not the practice of induction, but the feeling of involuntariness is the demarcation criterion, and it is important to ensure with self-report measures that the participants experienced a reduction in the level of control over their own behaviour (e.g., Palfi, Parris, McLatchie, Kekecs, & Dienes, 2018)[27]. From a practical perspective, it is important to bear in mind that the presence of a standard induction can increase responsiveness to the suggestions in the screening,oin average, by 1.46 (Martin & Dienes, 2018) compared to the absence of the induction; and that the strength of the effect of an induction fluctuates across suggestions (Terhune & Cardeña, 2016). Nonetheless, as argued earlier in this paper, a general reduction of responsiveness does not qualify as decisive argument for retaining the induction procedure. As long as the absence of the induction does not produce a floor-effect or alters markedly the ranking of the suggestibility scores, the screening can be perfectly adequate for screening people for individual differences in response. Indeed, there are existing attempts to assess responsiveness to suggestions without exposing the participants to an induction, such as the Barber Suggestibility Scale (Barber & Glass, 1962) and the CIS (Wilson & Barber, 1978). These scales can be easily administered in a context presented as a test of imagination while applying motivational instructions to replace the induction or simply leaving out the induction. The existing evidence suggest that employing motivational instructions creates similar level of responsiveness as the application of the induction; however, the absence of the induction significantly dwindles the level of responsiveness to suggestions (Barber & Wilson, 1978). Future research could explore the extent to which the exclusion or replacement of

[27] An operational definition of hypnosis necessitates to usage of induction to render suggestions hypnotic, and labels all suggestions without a priori induction imaginative suggestion (Braffman & Kirsch, 1999; Kirsch, 1997; Kirsch & Braffman, 2001). This line of thinking would preclude us from omitting the induction in case we want to measure hypnotic suggestibility.

the induction from the SWASH would be feasible and assess whether it would be beneficial.

A secondary interest of the current study was to assess the extent to which the length of the delay between the first and second screening affects the outcome of the screen and the psychometric properties of the measurement tool. Repeated assessment of suggestibility can negatively affect the suggestibility scores, for instance, if the delay amid the two occasions takes only a few days or weeks (Barber & Calverley, 1966; Fassler et al., 2008; Lynn et al., 1988). This reduction in suggestibility may be caused by boredom; the participants can become disengaged with the procedure by virtue of finding it repetitive (Barber & Calverley, 1966; Fassler et al., 2008). In our case, the short delay was a minimum of 5 months and we found no indication of substantial differences between the short and long delay groups among the SWASH subscales. For instance, Fassler et al. (2008) found a difference of 0.77 on the objective scores between the first and second session[28], but according to our data, the largest plausible difference is only .34. Nonetheless, the effect of boredom on the subjective scores observed by Fassler et al. (2018) was 1.05[29], which is compatible with our results as the lower bound of the difference in that aspect was 1.12. Taken together, our data imply that the negative effect of boredom might wear off or becomes negligible after 5 months; however, more research is needed to settle this matter and identify the ideal amount of delay that can prevent boredom effects in repeated designs.

We note that our sample was restricted to university students, which might preclude the generalization of our findings, crucially, the applicability of online hypnosis screening, to a wider population. Nonetheless, the problem of generalizability represents a universal issue in experimental hypnosis research. For instance, a meta-analysis on 27 studies investigating hypnotically induced analgesia found that from the studies with non-clinical samples (N = 19), only one was run with people recruited from the local community whereas all the other studies were run with students (Montgomery, Duhamel, & Redd, 2000). Recruiting from a wider population would not only increase generalisability of the findings but it would further facilitate researchers to run large-scale

[28] The reported raw difference was 0.54; however, we adjusted this value from a scale of 0-7 to the scale of the SWASH, which is 0-10.
[29] The reported raw difference was 2.2; we adjusted this value as well from a scale of 0-21 to the scale of the SWASH in which values can vary between 0 and 10.

hypnosis studies strengthening the replicability of the findings. Future research is needed to explore the extent to which online hypnosis research can be applied to screen and recruit people from local communities.

Finally, the vast majority of our participants were females; hence, the gender imbalance in our sample might be another factor hindering the generalizability of our findings. Research on the link between gender and hypnotic suggestibility has provided ambiguous results with some studies finding virtually no effect (Cooper & London, 1966; Dienes, Brown, Hutton, Kirsch, Mazzoni, & Wright, 2009; McConkey, Barnier, Maccallum, & Bishop, 1996) and some studies demonstrating a small effect size (Green, 2004; Green & Lynn, 2010; Morgan & Hilgard, 1973; Page & Green, 2007; Rudski, Marra, & Graham, 2004). Studies showing a small effect size of gender consistently found that women score higher than men, which might be caused by a divergence in a personality trait that partly underlies suggestibility or difference between women and men in how they assess the difficulty of the suggestions (Rudski, Marra, & Graham, 2004). Nonetheless, these explanations are conjectures that have yet to be tested. With only seven men in the current data set, we can only speculate how much gender might moderate the difference the online compared to the offline measurement of hypnotic suggestibility.

Conclusion

Altogether, the online assessment of hypnotic suggestibility appears to be feasible and the benefits far outweigh the downsides involved with its application. Although, online screening might be less engaging than the traditional, offline measurement of suggestibility and so it can result in slightly lower suggestibility scores, our study suggests that the effect size of this negative impact lies within acceptable boundaries. Crucially, the application of online hypnosis screening can subserve the execution of large-scale data collection with heterogeneous samples consisting of student and non-student participants as well. Furthering our knowledge based on small sample studies comes with many risks (e.g. Loken & Gelman, 2017), but the relative high cost of hypnosis screening procedures hinders the researchers of the field from running well-powered studies. Therefore, we argue that the adaptation of online hypnosis screening is salutary and it helps experimental hypnosis research to realise its full potentials.

Chapter VI: Why good enough Bayesian "evidence for H1" in one condition and good enough "evidence for H0" in another does not mean good enough evidence for a difference between conditions

Introduction

"The manipulation in condition A was statistically significant and by contrast, we found no statistically significant effect in condition B". Many believe that these findings are sufficient to support the claim that there is a difference between conditions A and B in the effect of the manipulation. However, such an inference does not follow from these results, as it requires the test of the difference between the conditions in the effect of the manipulation, or in other words, the test of the interaction of condition by manipulation (Abelson, 1995, p. 111; Gelman & Stern, 2006). This inferential mistake is common in neuroscience (Nieuwenhuis et al., 2011) and one can safely assume that psychologists are also not immune from committing it. While it is perhaps an old saw now when it comes to null hypothesis significance testing (NHST), how does this relate to the use of Bayes factors? As soon as conventional cut-offs are used for Bayes factors (see Box 1 for a brief introduction on the interpretation of the Bayes factor via the conventional cut-offs), there may be conditions where the inferential mistake is even more likely than with frequentist statistics. When there is good enough Bayesian evidence for H1 in one condition and for H0 in another, surely one can conclude that the effect is bigger in the first condition than the second! The reader can explore the extent to which they are attracted towards this inappropriate conclusion, which asserts an interaction without directly testing it, by considering the hypothetical study in Box 1. They can also test whether they would be more inclined to accept this inappropriate conclusion when it is based on frequentist or on Bayesian statistics. We discuss this scenario in detail in Example 2.

The Bayes factor is a continuous measure of the strength of relative evidence for H1 over H0 based on the ability of these hypotheses to predict the data at hand (Dienes, 2016; Kruschke & Liddell, 2018; Rouder et al., 2009). A Bayes factor of 1 means that the two hypotheses under comparison predicted the data equally well. The convention we follow (but it is not universal) is that the larger the Bayes factor the better H1 fits the data compared to H0, and the smaller it is, H0 is more in line with the data compared to H1. To aid decision making about the hypotheses, Jeffreys (1961) suggested $B > 3$ to be the cut-off of substantial evidence for H1 over H0. Note that this value was chosen with the intention that the Bayes factor should lead to similar judgment as NHST, when one is about to reject H0 (a statistical test resulting in $p = .05$ will usually provide a Bayes factor around 3, so long as the obtained effect size is about that predicted). By symmetry, we interpret $B < 1/3$ as substantial evidence for H0 over H1. However, it does not indicate that this cut-off should be automatically accepted as the level of good enough evidence. Indeed, it is a rough guideline and it remains a matter of scientific debate (e.g., currently, the level of good enough evidence for H1 is defined as $B > 6$ at *Cortex* and as $B > 10$ at *Nature Human Behavior* for Registered Reports). Nonetheless, in this tutorial, we apply the cut-off of substantial evidence as good enough evidence.

Box 1. The interpretation of the Bayes factor

The central goal of this tutorial is to substantiate the statistical intuition in the reader that to claim the existence of a difference between two conditions or groups, one always needs to test the interaction, and this principle is as true for Bayesian as frequentist statistics. In this tutorial, we present the scenarios that the reader can stumble upon when they calculate Bayes factors for the evidence of the presence of an effect in an experimental and a control condition (or group) that had a significant and a non-significant statistical test, respectively. By this approach, we aim to demonstrate that there are cases when using the Bayes factor instead of frequentist statistics could make it more likely to commit the inferential mistake, and there are cases when it may be the other way around. We use a hypothetical study as a case study and by increasing the sample size or reversing the effect size, we cover all the scenarios.

At a Golf Club in Sussex, a coach stumbled upon a sport psychology paper concluding that mental training (e.g., imagining to hit the ball with the club) can help golfers improve their skills when it is used combined with real training. Before implementing the mental training in all of their groups, they decided to test whether players can benefit from it. Therefore, they asked their students in one of their groups to engage in mental training twice every week for the next 3 months. They also had a control group in which the students underwent a real training but they were not told to do mental exercise, and the students had roughly identical skills to those in the mental training group. They assessed the performance of the students at baseline and after 3 months of training. The evaluation was performed on an interval scale from 0 to 10, and based on past studies with other sports they expected that after 3 months of training performance could improve by about 2 units.

To draw conclusions from their analyses, they used Null Hypothesis Significance Testing (NHST), and they set the level of alpha at the traditional .05. They reported the results of two statistical tests and a conclusion, which was based on these tests. Evaluate the appropriateness of this conclusion on a scale that ranges from 0 to 10 where 0 means that you feel that the conclusion is completely inappropriate and 10 means that you feel that the conclusion is completely appropriate based on the information at your disposal.

Comparing baseline and post-training performance in the control training group yielded a non-significant result ($t(19) = 0.29$, $M_{diff} = 0.11$, $p = .776$). However, when they analysed the data of the group of golfers who engaged in the mental training, they found a significant difference between baseline and post-training conditions, with better performance after 3 months of training ($t(19) = 2.61$, $Mdiff = 0.81$, $p = .017$). Based on these, they concluded that traditional training is more efficient when it is combined with mental training than when it is not combined with it.

Let us consider that the coach used the Bayes factor to draw conclusions instead of NHST. The Bayes factor is a continuous measure of relative evidence, it can tell us the extent to which our data supports one model (H1) over another (H0), which they reported as B (As we explain later, a model of H1 is needed. Given the researchers had reasons for expecting an effect of about 2 units, we used a half-normal with SD = 2.). By convention, Bayes factors larger than 3 indicate good enough evidence for H1 (i.e., we can conclude that an intervention works) and Bayes factors smaller than 1/3 indicate good enough evidence for H0. Assess the appropriateness of their conclusion that was, this time, based on the Bayes factors by choosing a value from the same scale of appropriateness. Zero indicates that you feel that their conclusion is completely inappropriate and 10 means that you feel that their conclusion is completely appropriate.

Comparing the baseline and the post-training conditions of the control group yielded good enough evidence for H0 ($t(19) = 0.29$, $M_{diff} = 0.11$, $p = .776$, $B = 0.24$). However, when they analysed the data of the mental training group, they found good enough evidence supporting H1 (i.e., difference between the baseline and the post-training conditions), with better performance after 3 months of training than at baseline ($t(19) = 2.61$, $Mdiff = 0.81$, $p = .017$, $B = 6.12$). Based on these, they concluded that traditional training is more efficient when it is combined with mental training than when it is not combined with it.

Box 2. Test your intuitions

The case study

Consider the hypothetical study from Box 2, in which a golf coach is trying to test whether or not adding mental training to traditional training can improve golf performance. To investigate this question, they randomly assigned students into traditional training (henceforth control group) and traditional plus mental training (henceforth mental training group) groups. They assessed golf performance at baseline and after 3 month of training. Therefore, they had a 2x2 mixed design. Hence, the crucial test of the idea that one can benefit more from golf training if it is combined with mental boils down to a 2x2 interaction of time of assessment (baseline vs post-training) and type of the training (traditional vs traditional plus mental). For the sake of simplicity, imagine that golf performance was measured on a scale from 0 to 10, and the coach expected that the mental training should improve performance with about 2 units.

Justifying the model of H1 and the model of the data

To compute a Bayes factor, one needs to specify the parameters of the models under comparison and of the model of the data (also referred to as likelihood; see Box 3 for more information on the essential parts of the Bayes factor). The model of H0 assumes no difference in the population. To model the prediction of H1 we employed a half-normal distribution with a mode of zero. The properties of the normal distribution align with the scientific intuition that small effect sizes are more probable than large ones (Dienes & Mclatchie, 2018), and we opted for a half-normal as it is in line with the directional prediction of H1. To specify the standard deviation of the distribution, we applied the expectation of the coach from the hypothetical study who assumed that performance should improve by about 2 units. When we have an effect size estimate based on earlier studies for instance, we can use it as the standard deviation of the distribution modelling the predictions of H1 (Dienes, 2014). Nonetheless, one might think that the effect size of an earlier study is not a plausible representation of the alternative, or there are no relevant studies. In these cases, there are several heuristics (e.g., Dienes, 2019) that one can follow to specify the predictions of H1. We used the t-distribution as the likelihood function, which is recommended over the normal distribution when the variance of the data is estimated as it is unknown (Dienes & Mclatchie, 2018). Finally, we will notate all of the Bayes factors as $B_{H(0, 2)}$ following the convention introduced by Dienes (2014). This notation includes all the necessary information about the model of H1: H indicates that it is a half-normal distribution; zero refers to the mode and 2 to the standard deviation of

the distribution. All Bayes factors reported in this paper represent evidence for H1 over H0.

Specifying all of these parameters requires the researcher to make many decisions, which has the side effect of increasing analytic flexibility and so the opportunity to cherry pick the results supporting the researcher`s pet theory. The most crucial step when one can introduce bias is perhaps the model specification of H1 that, in some cases, can have drastic effect on our conclusions. One way to reduce bias is by constraining analytic flexibility through pre-registering the exact parameters of the model of H1 or the strategy with which one will acquire those parameters (Chambers, 2013; Munafò et al., 2017). One can also consider reporting a "Robustness Region" (RR) that indicates the range of parameters (e.g., SDs of the half-normal distribution modelling H1) that would lead to the same conclusion (i.e., good enough support for H1 over H0, insensitive evidence, good enough support for H0 over H1) as the chosen model specification (Dienes, 2019). RRs can diminish bias by increasing transparency over the analytic choices in a similar manner as multiverse analyses (Steegen, Tuerlinckx, Gelman, Vanpaemel, 2016). RRs have the additional benefit that in cases when model specifications can be motivated in different ways, we can ascertain the robustness of our conclusion to the model specification by simply checking whether all plausible parameters lie within the RR. In this tutorial, we report the RR for every B in the format of "$RR_{conclusion}$[min, max]" where min indicates the smallest and max indicates the largest SD of the model of H1 that brings us to the same conclusion. We will indicate the original conclusion in the subscript of RR by reporting one of the following: "$B < 3$", "$1/3 < B < 3$" or "$B > 3$".

In order to assess the predictive ability of the hypotheses, we need to create models that represent their predictions. Modelling the prediction of H0 as no difference is the straightforward part of the process. However, specifying the predictions of H1 requires scientifically informed decisions in every case and so it can be a subject of debate. For instance, one needs to define the shape and parameters of the distribution representing the predictions of H1 on the possible population effect sizes. This brings up questions, such as the distribution should be uniform, t or normal; one-tailed or two-tailed; centred on zero or on a non-zero population value; what should be the level of variance in the model. The discussion of these decisions is beyond the scope of the current tutorial and so we refer the reader to Dienes (2015, 2019), Dienes and Mclatchie (2018). Nonetheless, in the case study of the current paper, we justify all of the choices about the model specifications. Finally, one needs to define the likelihood function, which is modelling the probability of the data along different population effect sizes.

Box 3. The anatomy of the Bayes factor

Disclosure

All the materials of this tutorial are available on the Open Science Framework page of the project at: https://osf.io/xrctq/. The page includes the R script of the analyses introduced here and the script of the Bayes factor function. It also contains the R script of a Bayes factor Shiny app that is a simple and interactive web application that can calculate the Bayes factors of 2x2 between groups and within participants designs. Box 4 demonstrates an example of the usage of the Bayes factor R script (namely, the test of the interaction in Example 1) and Figure 1 portrays how the Shiny app can be applied to compute all three Bayes factors of Example 1. The Bayes factor Shiny app can be accessed at https://bencepalfi.shinyapps.io/Bayesian_Interaction_App/

To calculate the Bayes factor in R (or RStudio), one needs to obtain the summary statistics of the data (mean, standard error and degrees of freedom) and decide on the parameters of the model of H1.

The following R script reproduces the results of the test of the interaction of Example 1 (all the text preceded by the # symbol are comments helping the reader and will be ignored by R when the script is run):

```
#Loads the Bayes factor function
#Note that the current R file and the file containing the function should be placed in the same folder
source("BayesFactor_normalH1_tlikelihood.R")

#Calculates the Bayes factor
Bf(sd = 0.491, obtained = 0.7, dfdata = 38, meanoftheory = 0, sdtheory = 2,  tail = 1)
```

The first three arguments of the function specify the parameters of the likelihood: the standard error, the estimate (i.e., raw effect size) and the degrees of freedom of the distribution, respectively. The last three arguments define the parameters of the model of H1: the centre (or mode if it is a one-tailed distribution) and the SE of the distribution, and whereas it is one- or two-tailed. When all parameters are provided, the function returns a vector containing the value of the Bayes factor (evidence for H1 over H0).

Box 4. Calculating the Bayes factor in R

Example 1: When the Bayes factor helps us avoid committing the inferential mistake

Suppose that the researchers found that the test comparing baseline and post-training performance was significant in the mental training group (t (19) = 3.58, M_{diff} = 1.11, p = .002), and it was not-significant in the control group (t (19) = 1.08, M_{diff} = 0.41, p = .295). Based on this, they concluded that the effect of the training is only expressed (at least after 3 months of training) if it is combined with mental training. This conclusion is premature for two reasons. First, one cannot claim the absence of an effect based on a

non-significant test (Cohen, 1994; Dienes, 2014, Rouder et al, 2007) and so it is false to imply that we have evidence against the effectiveness of the training in the beginner group. We don`t have evidence for the contrary either, simply, we need to refrain from decision-making. Second, a more relevant point for the purpose of the current tutorial, the dissimilarity of two categorical statements (i.e. significant vs not-significant) does not grant a meaningful categorical statement about their difference (i.e., the difference between the two is not necessarily significant in itself; Abelson, 1995, p. 111). From the second point, it follows that one needs to test the difference directly to make any meaningful claim on the interaction of the groups. The test of the interaction, however, yields a non-significant result (t (38) = 1.43, M_{diff} = 0.70, p = .162) meaning that one needs to suspend judgment regarding the influence of the mental training on traditional golf training.

The question arises: how would we decide in this scenario if we were to rely on a Bayes factor to form conclusions about the hypotheses? These data translate into substantial evidence for the effect of the training in the mental training group ($B_{H(0, 2)}$ = 46.36, $RR_{B > 3}$[0.2, 36.3]) and leaves us with insensitive evidence for the effect of training in the control group ($B_{H(0, 2)}$ = 0.57, $RR_{1/3 < B < 3}$[0, 3.5]) as well as for the interaction directly comparing the effects of the groups ($B_{H(0, 2)}$ = 1.14. $RR_{1/3 < B < 3}$[0, 7.4]). Clearly, one cannot easily claim that good enough evidence for the effect in one group and insensitive evidence in the other group is good enough evidence in itself for the difference between the groups. Apparently, using the Bayes factor may help us avoid the inferential mistake regarding the interaction, even if we were to ignore the results of the direct test of the interaction. Hence, using Bayes factors may increase the chance that one would conclude that the available data are simply not enough to make a decision about the hypotheses.

The only way to come to a conclusion regarding whether or not mental training combined with traditional training is superior to traditional training is to collect more data until we obtain evidence in one direction or the other. Optional stopping is not a problem for Bayesian statistics, the Bayes factor will retain its meaning regardless of the stopping rule applied (Dienes, 2016; Rouder, 2014)[30]. Thus, we can check the Bayes factor every time we recruit a new participant and stop once the Bayes factor reaches a good enough

[30] Recently, this claim has been called into question under some conditions by Heide and Grünwald (2017). For replies to the concerns see Rouder (2019), and Wagenmakers, Gronau and Vandekerckhove (2019).

level of evidence. For example, in this scenario, assuming that the raw effect sizes and their variances remain constant, we would need to recruit 94 participants in total (47 per group) to have substantial evidence for the interaction ($B_{H(0, 2)} = 3.09$, $RR_{B > 3}[0.3, 2.0]$). In this scenario, the evidence for the efficacy of the training in the control group would still be insensitive with a $B_{H(0, 2)} = 0.89$, $RR_{1/3 < B < 3}[0.8, 5.4]$. Thus, evidence for an effect in one group, coupled with no evidence one way or the other in the other group, could still be evidence for a difference in effects between the two groups.

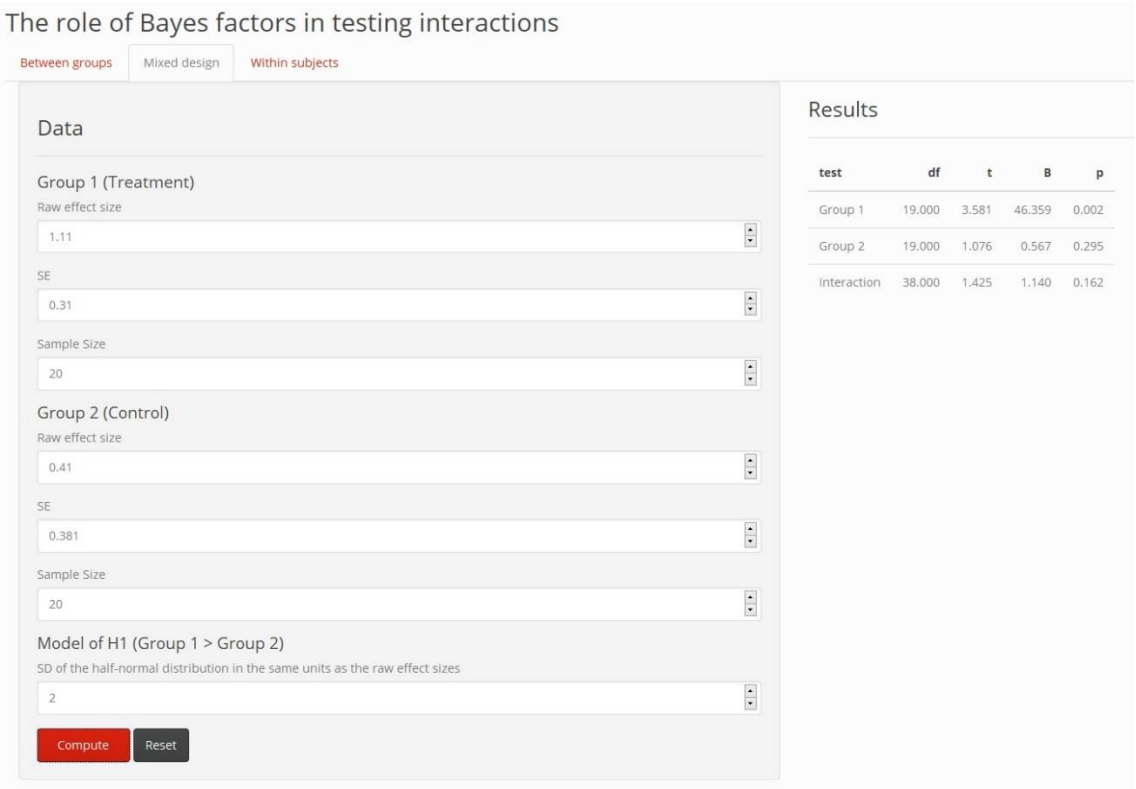

Figure 1. Print screen of the Shiny app that calculates the Bayes factor separately for the two groups and for the interaction based on the following statistical parameters: raw effect sizes and their SEs for the two groups and their interaction, the sample size, and the SD of the half-normal distribution that models the predictions of H1. In the top left corner, one can change between the "between groups", "mixed design" and "within subjects" options. The between groups and mixed design options are identical in that they run an independent t-test to test the interaction, and they request the same input parameters. For the within subjects design, one needs to provide the difference of the conditions and their standard deviation separately. The results appear on the right side of the screen, once the calculate button is pressed. The Shiny app reports the degrees of freedoms, the t-values, the Bayes factor and the p-values for the groups (or conditions) and for their interaction.

Example 2: When the Bayes factor might exacerbate the problem and seemingly creates an inferential paradox

Now let us consider the scenario from text Box 2 that only differs from Example 1 in that the raw effect sizes of differences between baseline and post-training conditions were reduced by 0.3 units in both of the groups. All other parameters (e.g., the standard deviations and the difference between the control and mental training groups) were kept constant. In this scenario, the results of significance tests probing the efficacy of the training, separately in the control and mental training groups, are identical to those of Example 1, being non-significant and significant, respectively. However, if we calculate the Bayes factors, it reveals that this scenario is different from Example 1 as we gain good enough evidence for the presence of the effect in the mental training group (t (19) = 2.61, M_{diff} = 0.81, p = .017, $B_{H(0, 2)}$ = 6.12, $RR_{B > 3}$[0.2, 4.3]) and good enough evidence for the absence of the effect in the control group (t (19) = 0.29, M_{diff} = 0.11, p = .776, $B_{H(0, 2)}$ = 0.24, $RR_{B < 1/3}$ [1.5, Inf]). It might seem intuitive to conclude that the evidence for the difference of the two must be substantial in itself as well (c.f., your feeling of appropriateness about the conclusion in Box 2). However, that is an unwarranted conclusion as the rule that "a meaningful categorical statement does not follow from the difference between the two categorical statements" (Abelson, 1995, p. 111) applies to Bayesian just as much as it applies to frequentist statistics. Hence, regardless of how tempting it feels to claim that the group with substantial evidence for H1 must be different from the group with substantial evidence for H0, we need to directly compare these two conditions to unravel whether there is an interaction.

Compared to Example 1, it appears that in this case relying on the Bayes factor rather than on the p-value would not help us avoid making the inferential mistake of ignoring the test of the interaction. On the contrary, using the Bayes factor may even amplify the problem as having good enough evidence for H1 in group A and for H0 in group B can easily create the false impression that there is no need for further statistical analyses and the two must be different. However, neglecting the test of the interaction is an inferential mistake, moreover, it would lead us to an incorrect conclusion, since the test of the interaction must yield the same result as in Example 1 as we kept the difference between the groups and their standard deviation fixed. It means that the test of the interaction is non-significant (t (38) = 1.43, M_{diff} = 0.70, p = .162) and the Bayes factor is insensitive ($B_{H(0, 2)}$ = 1.14. $RR_{1/3 < B < 3}$[0, 7.4]).

Seemingly, we got ourselves into a paradox in which we can claim that an effect exists in group A and it does not exist in group B, however, we cannot state that the effect is stronger in group A than in group B. These conclusions are inconsistent with one another, but the Bayes factor should not take the blame for it. The cause of the existence of this paradox is that we introduced cut-offs to interpret the Bayes factor and so we reduced its continuous nature to a categorical one. That is, the Bayes factors underlying the claims that there is an effect in group A and that the interaction is insensitive point in the same direction. Hence, we created the inconsistency by imposing a cut-off and labelling the first as good enough evidence for H1 and the second as insensitive evidence. Nevertheless, applying a cut-off to discern good enough from insensitive evidence is useful for scientific practice as it allows us to draw conclusions. And we often need to draw conclusions in order to move on with an experiment: Have we equated two conditions in terms of a possibly confounding variable, such as expectancy; have we established that manipulation does what it says; have we ruled out a nuisance alternative theory, and so on (Only a statistician and not a scientist would recommend not ever drawing any conclusions from statistics!). In other words, we gain a clear rule telling us when we have good enough evidence to make such a decision. On the other hand, if we rely on this decision rule, we need to accept that it can lead us to paradoxical situations.

Fortunately, there is a way to escape this paradox. There is no need to consider the evidence at our disposal as fixed. Therefore, the remedy to this problem is to collect more data until the Bayes factor of the crucial test exceeds one of the cut-off values (as mentioned earlier, optional stopping does not invalidate conclusions based on the Bayes factor). Hence, unless our resources are limited, we should always be able to resolve this paradox by simply continuing the data collection process and accumulating more evidence. For instance, assuming that the raw effect sizes and their variances stay constant while we collect data for this study, we would need to recruit the same number of participants (68 per group) as we needed in Example 1 to obtain evidence for the difference between the groups.

Discussion

In this tutorial, we aimed to illustrate how the application of the Bayes factor with cut-offs relates to the old problem of the tendency to compare the statistical significance of the tests of two groups rather than the groups themselves. We introduced two scenarios

in which group A had a significant effect whereas group B had a non-significant effect. In Example 1, employing the Bayes factor instead of the NHST may help us avoid the inferential mistake as the test of the non-significant group turned out to be insensitive and it is unlikely that one would assume that the difference of good enough evidence for H1 in group A and insensitive evidence in group B indicates a clear difference between the two. In Example 2, however, the Bayes factor in group B provided good enough evidence for H0 and in such a scenario applying the Bayes factor instead of the NHST may increase the probability of committing the inferential mistake because the conclusions from the simple effects and the interaction contrast are literally inconsistent.

We observed that drawing a conclusion from the Bayes factor could sometimes lead to a paradox (i.e., good enough Bayesian evidence for H1 in group A, good enough Bayesian evidence for H0 in group B and the lack of good enough Bayesian evidence for their interaction). The reason of the paradox is that we use cut-offs that exchange the continuous measure of evidence to a categorical and ultimate claim about the state of the world in order to guide our decisions about the hypotheses. This situation bears a strong resemblance to Arrow`s theorem (1951) that demonstrates that there is no consistent way to explore the preference of a group ("will of the people"), and any ranked voting system (i.e. a system that turns strengths of opinion into a categorical outcome) will lead to paradoxes, perhaps undermining our faith in representative government and democracy itself. However, as Deutsch (2011) pointed out, Arrow`s theorem considers only a particular stage of decision making, as if preferences and options were fixed in social decision making. As long as preferences can be altered through open discussion and reasoning, and it is possible to modify or replace the options, democracy can be used consistently in selecting good policies. This conclusion is just as true for science, which seeks good explanations rather than policies or governments. When it comes to science, it would be a mistake to assume that evidence or the list of options (tested hypotheses) are fixed. Hence, even if we stumble upon a paradox, it remains a transient state and by accumulating more evidence, or by modifying the hypotheses (e.g., replacing a one-sided H1 with a one-sided H2 pointing into the other direction) we can dissolve the inconsistency. That is, the issue we raise about cut-offs leading to paradox is a very general one, not unique to science, let alone Bayes factors. The solution is just as general.

Continuing data collection until we obtain good enough evidence for one of the theories can be challenging in some cases as we do not have unlimited resources. Thus,

estimating the sample size we might need to find good enough evidence for a hypothesis over another one should play an essential role in the planning phase of an experiment. To this aim, we can compute the rough estimate of the sample size we need to probably obtain a Bayes factor that is equal to or larger than a specific value (i.e., the cut-off of good enough evidence defined by us). For instance, to have a long-term relative frequency of 50% to obtain a Bayes factor of 3 (or 1/3), we should simply replicate the steps of the sample size elevation of Example 1 and 2. That is, we can take the raw effect size and its standard deviation of a pilot study and assume that these parameters remain constant while we raise the sample size (see Dienes [2015] for a detailed tutorial). If we want larger probability, then we need to continue increasing the sample size until we reach a Bayes factor of 20 (or 1/20). This procedure estimates the sample size that has a long-term relative frequency of 80% to obtain a Bayes factor of 3 (or 1/3). See Table 1 for the Bayes factors one needs to reach with this sample size estimation procedure to ensure that their design is set to find good enough evidence with a given long-term relative frequency. Also, see the Supplementary Materials for more information on how these thresholds were determined, and on how the reader can calculate the threshold for any chosen combination of a cut-off of good enough evidence and long-term relative frequency of finding good enough evidence. For an alternative view on how to plan the design of a future experiment to achieve good enough evidence, see Schönbrodt and Wagenmakers (2018). Finally, it is important to bear in mind that the sample size estimation is useful for planning, such as roughly estimating how long data collection will take, it has no influence on the inference made once the data are in. The final Bayes factor obtained is the measure of evidence for H1 over H0. The meaning of the Bayes factor is independent of the sample size estimation procedure (Dienes, 2016).

Table 1

Table of Bayes factors with which one can approximate the Sample Size of a Design that will find Good Enough Evidence for a Given Cut-Off with a Specific Long-Term Relative Frequency

	Long-term relative frequency of good enough evidence			
Cut-off of good enough evidence	50%	80%	90%	95%
B > 3 or B < 1/3	3	20	70	220
B > 6 or B < 1/6	6	40	150	520
B > 10 or B < 1/10	10	85	350	1370

Note. Using optional stopping with a relatively low cut-off of good enough evidence (i.e., 3) may deteriorate the long-term relative frequency of the design to find good enough evidence for H1 when in fact H1 is true. Given that the null is true, designs using these thresholds can deliver good enough evidence for H0 with higher probability than their corresponding expected long-term relative frequencies. See the Supplementary Materials for the demonstration of these claims via simulations.

In conclusion, it is evident that the Bayes factor is not a panacea for the inferential mistake discussed in this tutorial, since we demonstrated that the reliance on the Bayes factor mitigated the issue in Example 1, but it exacerbated it in Example 2. By depicting these two examples, we intended to raise awareness that any claim about the moderating effect of an independent variable should be supported by a sensitive test of the interaction regardless whether one uses frequentist or Bayesian statistics. Irrespective of how paradoxical it seems, good enough Bayesian evidence for H1 in group A and good enough Bayesian evidence for H0 in group B does not necessarily mean good enough Bayesian evidence for the difference of the two.

Chapter VII: General Discussion

Summary

The studies presented in this thesis investigated the relationship between metacognition of intentions and the implementation of intentional acts. This idea was examined by using hypnosis as an experimental tool to create alterations of reality and the feeling of involuntariness in highly suggestible people. According to the cold control theory (Dienes & Perner, 2007), hypnosis creates the illusion of involuntariness while leaving first-order intentions unaffected, hence it should not provide highs with abilities they do not possess outside of hypnosis. This prediction is challenged by a phenomenon (i.e., word blindness effect) providing a great opportunity to understand the boundaries of the theory. Chapters II-IV focused on unravelling the mechanisms by which highs respond to the word blindness suggestion and manage to reduce Stroop interference in order to understand whether unconscious intentions can be more effective than conscious intentions, and whether cold control theory needs to be revised. Chapter V presented an empirical evaluation of the online screening of hypnosis by the Sussex-Waterloo Scale of Hypnotisability. Chapter VI is a statistics tutorial illustrating why Bayesian evidence for H0 in one condition and Bayesian evidence for H1 in another condition does not mean Bayesian evidence between the two conditions.

Hypnosis as a purely metacognitive phenomenon

Hypnosis is a powerful tool in altering perceptions of reality and creating the feeling of involuntariness (Weitzenhoffer, 1974, 1980; Terhune, Cleeremans, Raz & Lynn, 2017). A simple explanation of this intriguing phenomenon is provided by the cold control theory that postulates that hypnosis is a purely metacognitive phenomenon. This hypothesis follows from the two core assumptions of the cold control theory. First, according to cold control theory, responses to hypnotic suggestions are intentional acts, which are implemented by regular cognitive control processes. Hence, subjects need to form first-order intentions to respond to hypnotic suggestions (cf. Norman & Shallice, 1986). Second, hypnosis targets the HOTs of intending rather than the first-order intentions themselves. A hypnotic suggestion requests subjects to replace accurate HOTs of intending with inaccurate ones so that one`s intentional act can be experienced as if it is involuntary (cf. classical suggestion effect of Weitzenhoffer, 1974, 1980). Based on these

two assumptions, cold control theory predicts that responses to hypnotic suggestions cannot be more effective by objective standards than responses to volitional requests, if subjects create identical first-order intentions in these cases (cf. limited casual function of HOTs in general, Rosenthal, 2008). In other words, the theory proposes that hypnosis is a purely metacognitive phenomenon.

The current thesis did not test directly the two underlying assumptions of the cold control theory. However, some of the Chapters present indirect evidence supporting these ideas. For instance, as highlighted in Chapter I, experimental evidence from many studies demonstrated that responses to hypnotic suggestions are goal-directed and appropriate to the context in which they were created (Spanos & Barber, 1974; Comey & Kirsch, 1999; Hilgard, 1977; Spanos, 1986). As argued earlier, this finding implies that hypnotic actions are implemented by intentional control processes. If hypnosis could directly induce behavior (e.g., by activating a schemata [Norman & Shallice, 1986]), it would be possible that the action of the subject is not in line with her goals. Chapter II presented evidence that the word blindness suggestion effect is modulated by the proportion of incongruent trials of the experimental blocks. High compared to low proportion of incongruent trials resulted in stronger reduction in Stroop interference by the suggestion. This finding coincides with the notion of hypnotic responses being strategic and goal-directed, and so it presents indirect evidence for the idea that hypnotic responses are driven by first-order intentions.

On many occasions, the first-order intentions that hypnotic subjects form to comply with suggestions are strategies in which they engage without being aware of doing so (e.g., imagining the buzz of a mosquito to create the experience of a mosquito being present). If this core assumption of cold control theory is true then subjects should be able to engage in the same strategy in hypnotic and non-hypnotic ways. Indeed, the correlation results of the Pilot and the Registered studies of Chapter IV present evidence supporting this notion. The studies demonstrated that when highs respond to the word blindness suggestion and to the appropriate volitional request, then there is a correlation between the two conditions in extent to which highs managed to reduce Stroop interference. Interestingly, this correlation was found to be rather strong in both of the studies (we can be 97.5 confident that the correlation is stronger than .48). A strong correlation is what one would expect if one were to assume that highs respond via identical cognitive processes in the two conditions.

The second assumption of cod control theory, which claims that hypnosis changes metacognition of intentions, remained untested as well in this thesis. Nonetheless, Chapter IV presents evidence that is in line with this notion. If hypnosis changes metacognition over one's intention to act upon a suggestion then one should experience and report reduced level of control over one's response to a suggestion than to a volitional request. Indeed, subjects reported reduced level of control over the meaningfulness of the Stroop words in the Pilot as well as in the Registered study of Chapter IV. Moreover, subjects` phenomenology was altered by hypnosis in the predicted way: they reported to experience meaninglessness as if they perceived it in the suggestion condition and as if they imagined it in the volition condition. This corroborates the prediction of cold control theory that expects that subjects have hallucination like experiences by relinquishing their metacognition over their intentions to create the experience of meaninglessness. Future research could explore the domain specificity or generality of the role of metacognition in hypnosis. Neuroimaging studies (Morales, Lau & Flemming, 2018) and studies with patients with lesions to anterior prefrontal cortex (Fleming, Ryu, Golfinos, & Blackmon, 2014) showed that metacognition of memory and perception have overlapping as well as independent neural substrates indicating that metacognition can be domain-specific, however, this question remained unexplored for the domain of intentions. Although cold control theory has no clear prediction on this question, unraveling whether or not suggestibility is solely related to metacognitive abilities of the domain of intentions can further our understanding of HOT theories of consciousness (Lau & Rosenthal, 2011).

The central aim of this thesis was to test the hypothesis of cold control about hypnosis being a purely metacognitive phenomenon and having no effect on first-order intentions. As argued in Chapters I, III and IV, the vast majority of studies testing the superiority of hypnotic over non-hypnotic responses demonstrated that non-hypnotic responses can be just as effective as hypnotic ones (e.g., Barber, 1966; Erdelyi, 1994; Levitt & Brady, 1964; Milling, Kirsch, Meunier & Levine, 2002; Nogrady, McConkey, & Perry, 1985; Spanos, 1986). However, a single phenomenon in which non-hypnotic responses are unable to achieve what hypnotic responses can is sufficient to challenge this prediction. The word blindness suggestion appears to be a phenomenon exactly like that. Therefore, understanding what highs do to reduce Stroop interference, and demonstrating that they can do the same in a voluntary manner is indispensable to retain the current version of cold control theory. Chapter III demonstrated compelling evidence

supporting that the effect of word blindness suggestion cannot be reproduced by simple volitional strategies such as looking-away from the target, focus on a single-letter of the Stroop word, blurred vision or goal-maintenance by internal rehearsal of the goal of the task. These results are unfortunate for the cold control theory and for the idea that hypnosis is purely metacognitive. Sadly, Chapter IV did not settle the issue either. The comparison of the volition condition and the word blindness suggestion in respect of the reduction of the Stroop interference effect did not yield good enough evidence for H0 or H1 (i.e., superiority of the suggestion). Although there was anecdotal evidence supporting that a volitional request to create the experience of meaninglessness is as effective as the word blindness suggestion in reducing Stroop interference, for now, the cold control theory hangs in the balance.

Imagination but not mental imagery may be the key to word blindness

Good enough evidence for equal reduction in Stroop interference by volitional and hypnotic responses would be the favorable outcome for the cold control theory. However, this outcome would still not shed light on the underlying mechanism of the word blindness suggestion itself. Identifying the strategy by which highs alleviate Stroop interference is crucial to test whether or not lows and mediums could use it to reduce Stroop interference outside of the hypnotic context. In addition, the exploration of a successful strategy would open new avenues for research on the Stroop effect and on cognitive control (e.g., how does the strategy fit computational models of cognitive control?). Chapter III presented a hypothesis that the underlying strategy of the word blindness suggestion may be imagination itself. Perhaps, highs take the instructions to see the words as a meaningless script literally, and they imagine the words as meaningless, which may result in the reduction of Stroop interference. Here, the term imagination should be understood in its broadest possible form, namely, the formulation and engagement in a counterfactual model of reality (Perner, 1993). This general definition of imagination is in line with contemporary concepts of imagination in cognitive sciences (e.g., Currie and Ravenscroft, 2002) and in the field of hypnosis (e.g., imaginative suggestions by Braffman & Kirsch, 1999).

Crucially, imagining something, or in other words, considering the world in a counterfactual way (*think of*) is not identical to holding a belief (*think that*) about what the world is like (Perner, 1993). Therefore, the underlying mechanism and the effects of imagination are discernable from those of beliefs (cf. response expectancies by Kirsch,

1985). The imagination hypothesis proposes that entertaining a model of the world in which words are meaningless might be enough to help highs decrease Stroop interference via, for instance, resetting top-down cognitive control processes. If being engaged in this counterfactual model is necessary to alleviate interference then subjects should experience more words as meaningless when they respond to the suggestion than when they do the Stroop task without trying to see the words as being meaningless. Indeed, this pattern of results was observed in Chapter II as well as in Chapter IV. Moreover, the imagination hypothesis predicts that the percentage of trials on which highs experienced the words as meaningless is positively related to the extent to which the suggestion reduced Stroop interference. All corresponding tests of this proposal remained insensitive in Chapter II (Exploration 3), however, Chapter IV presented good enough evidence for a positive relationship for the suggestion as well as for the volitional request.

Importantly, the proposed definition of imagination includes but does not equal to visual or mental imagery (Currie and Ravenscroft, 2002). Limiting imagination to imagery would imply that highs have a vivid, visual experience of meaningless words that are not present when they respond to the word blindness suggestion. This is not necessarily the case. Highs may experience meaninglessness as a conceptual or propositional hallucination, they may see the words crisply, in the form that they are presented, and still experience that the words have no meaning as if they are written in characters that are unknown to the reader. This would not be surprising as they were instructed to experience meaninglessness precisely this way by the suggestion as well as by the volitional request. The items of the subjective measures applied in Chapter II and IV were designed to assess whether subjects followed the instructions and experienced what the suggestion or volitional request asked them to experience: for instance, “the words on the screen were written in a clear yet completely meaningless script”. Since the subjects were asked to create a conceptual rather than a visual hallucination, the findings of Chapter IV about the positive relationship between subjective reports of meaninglessness and reduction in Stroop interference corroborate the imagination as counterfactual model rather than imagination as visual imagery idea. Nonetheless, vividness of mental imagery was not measured in any of the studies, so more research is needed to establish whether or not vividness of mental imagery is related to the alleviation of Stroop interference. An interesting experimental test of this question would be to give the word blindness suggestion to people with aphantasia, who are unable to create vivid

mental images (Zeman, et al., 2010, 2015). This inability to form vivid mental images is linked to an impaired performance on working memory tasks that require the engagement of visual imagery, however, it does not affect performance on other cognitive tasks such as the test of spatial memory (Jacobs, Schwarzkop, & Silvanto, 2018; cf. Baddeley, 2001). This finding suggests that aphantasia should not prevent many forms of imagination. Hence, people with aphantasia should be able to respond to hypnotic suggestions that involve imagination, including the word blindness suggestion, unless the suggestion specifically requests the formation of a vivid mental image.

Finally, provided that there is good enough evidence that highs can decrease Stroop interference to the same extent when they respond to the volitional request used in Chapter IV and to the word blindness suggestion, the imagination hypothesis would gain empirical support. An obvious follow up study would test whether or not lows and mediums can reduce Stroop interference when they are asked to voluntarily imagine the Stroop words as meaningless. According to the cold control theory, highs are different from non-highs in their ability to influence the metacognition over their own intentions, or in other words, to create the feeling of involuntariness while they respond intentionally. Therefore, if imagination is the key to the reduction of Stroop interference (when highs respond to the word blindness suggestion) then the cold control theory expects lows and mediums to be able to use this strategy just as well as highs do.

The cornerstone of a good scientific explanation (or theory) is that it is hard to vary it without altering what the explanation is meant to account for (Deutsch, 2011). The reach of a theory is its ability to provide explanations for phenomena for which the theory was not created to account for (note that good rather than bad theories possess reach, as it is easier to determine the reach of an explanation that is hard to vary). Importantly, formulating good theories and identifying the reach of them are driving factors of the accumulation of knowledge and scientific progress. The imagination hypothesis is an explanation with a straightforward reach. Any phenomenon with the following two features can potentially be explained by the imagination hypothesis: 1) people entertain a counterfactual model of the world in which words are meaningless; 2) people manage to reduce Stroop interference. The word blindness suggestion is not the only phenomenon that meets these criteria. Another relevant phenomenon is the social priming of dyslexia. It was demonstrated by two independent labs that when people are asked to imagine what is it like to have dyslexia, they perform better on the Stroop task (indexed by the Stroop

interference effect) than after imagining a neutral scenario that does not involve experiencing words as meaningless (Augustinova & Ferrand, 2014b; Goldfarb, Aisenberg, & Henik, 2011). As argued in Chapter III, the word blindness suggestion effect and the dyslexia prime effect produce similar patterns of results. Importantly, they both reduce RTs of incongruent trials and they only affect the response conflict component of Stroop interference. These similarities between these two phenomena suggest that the imagination hypothesis has a reach beyond the word blindness suggestion. Nonetheless, the imagination hypothesis is still a conjecture. Hence, future research should test whether dyslexia priming and word blindness happen via the same cognitive mechanisms and whether this underlying mechanism is imagination.

Implications of cold control theory

The perspective of hypnosis in therapies.

The core prediction of the theory that was investigated in this thesis is that hypnotic and non-hypnotic responses differ solely in the form of their accompanying HOT, and so hypnosis cannot provide people with abilities they do not possess outside of hypnosis (Chapters III and IV). This hypothesis has clear and substantial implications for the application of hypnosis in the clinical and therapeutic context. It implies that clients of therapies cannot benefit from responding in a hypnotic rather than in a non-hypnotic manner. Indeed, evidence supports that the effect of hypnosis on enhancing therapies is negligible when hypnotic and control groups receive identical treatment apart from the delivery of hypnosis (e.g., Adachi, 2012; Schaeafert, 2014). Nonetheless, a hypnotic compared to a non-hypnotic therapy can differ in other aspects than the form of the accompanying HOT of the responses of the subjects. Beliefs about hypnosis are so deeply ingrained in our culture that the ritual of hypnosis became an effective tool to influence expectations and the motivation of the subjects (Kirsch, 1985; Spanos, 1986; see also in Chapters II and IV). People with positive attitudes towards hypnosis (mostly highs and mediums) experience an elevation in expectations and motivation in the hypnotic context. Therefore, hypnosis may facilitate the responses of highs and mediums to hypnotic suggestions via, for instance, response expectancies (Kirsch, 1985). In line with this assumption, a meta-analysis of studies investigating cognitive therapies used in tandem with hypnosis demonstrated that the inclusion of the term hypnosis in the instructions can have a positive effect on the outcome of some therapies (Kirsch, Montgomery, &

Sapirstein, 1995). This finding subserves the idea that hypnosis can be beneficial in the clinical context even if the hypothesis about hypnosis being a purely metacognitive phenomenon is confirmed. Nonetheless, the evidence relating to this hypothesis that was presented in Chapter IV is not conclusive, hence further research is needed to understand the boundaries of cold control and its implications for clinical hypnosis.

The reach of cold control.

The cold control theory was created as a theory of hypnotic responding. It provides a simple, mechanistic explanation of how hypnotic subjects are able to respond intentionally and yet experience their response as if it is involuntary. However, cold control is not limited to the context of hypnosis, the theory does not assume that the unique ability of highs with which they can relinquish metacognition over their own intentions is restricted to hypnosis in any way. Hence, cold control may be the underlying mechanism of actions and experiences that are deliberately formed but sensed as if they happened automatically (Dienes et al., 2019). This is particularly important for studies of conscious experiences highs (and mediums in many cases) can easily transfer demand characteristics (Orne, 1962) into conscious experiences without being aware of doing so. For instance, a study demonstrated strong correlation between hypnotic suggestibility and several measures of conscious experience such as the rubber hand illusion, the vicarious pain questionnaire and the mirror touch synaesthesia (Lush et al., 2019). These findings imply that cold control may partially or completely account for these phenomena. Therefore, unraveling the limits of cold control, for instance, whether it can enhance the efficiency of first order intentions (Chapter III and IV) is indispensable to understand its implications beyond the field of hypnosis.

Advancing the measurement of individual differences in hypnotic suggestibility

The results of Chapter V demonstrated evidence that the online administration of SWASH and hypnosis screening, in general, are viable. Crucially, comparing the data of offline and online screen revealed that the two methods are comparable in the following aspects. The distribution of the scores was akin in the two samples and there was no sign of floor effect in the online sample; the internal validity (i.e., correlation of objective and subjective scores) was good for both of the offline and online measures; finally, there was a strong correlation between offline and online scores. The application of online hypnosis screening paves the way for large-scale hypnosis studies with heterogeneous samples.

Low power is a widespread problem in psychological science (e.g., Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1992) and an obvious way to treat it is to increase the sample size of the studies. This issue applies to hypnosis research as well. Many interesting questions of the field of hypnosis can only be sensitively tested by large-scale studies: for instance, some predictions can be tested by correlational designs (cf., Chapter II & III) or in many cases the effect size predicted by H1 is either small or it is predicted that there is no effect (note that, generally, one needs more data to reach good enough evidence for H0 than for H1). To enhance the sensitivity of a study, one could also increase the number of trials, however, as we have seen it in Chapter II-IV, studies involving hypnosis procedures are already long and demanding for the participants. Hence, in order to ensure that the collected data have good enough quality (e.g., participants do not get tired during the study), increasing the sample size rather than the number of trials seems to be the felicitous choice.

Online screening may also contribute to the improvement of the generalisability of hypnosis research by increasing the diversity of the samples. Most of the hypnosis studies, as well as cognitive psychology studies, are conducted on student samples (e.g., Montgomery, Duhamel, & Redd, 2000). This is an unfortunate fact that challenges the external-validity of the findings of the field (cf., the vision statement of the newly appointed editor of APS in which she encourages the submission of studies with more diverse samples; “New Psychological Science Editor”, 2019). The application of online screening could help hypnosis researchers to increase the generalisability of their findings by, for instance, screening and recruiting subjects from local communities.

The application of the online SWASH is a substantial leap in the right direction, however, there are remaining issues regarding the measurement of hypnotic suggestibility that can be improved. For instance, the current version of the SWASH still takes about an hour to finish. This may disincentivize people to undertake the test or some participants may lose interest in the screening procedure during participation. The SWASH and many other screening procedures consist of three parts: induction, delivery of suggestion and a questionnaire in which participants report their experiences. The second and third part can only be shortened by reducing the precision of measurement. For instance, taking out some of the suggestions could speed up the screen, however, with fewer items, SWASH would be less accurate in assessing the suggestibility of individuals. The first part, the induction, consists of many suggestions, for instance, to relax or to focus attention.

However, responses to these suggestions are not evaluated as part of the screen (e.g., Shor & Orne, 1963; Weitzenhoffer & Hilgard, 1962). Therefore, leaving out the induction could substantially decrease the time of the screen. This could be problematic as some definitions of hypnosis require the usage of inductions and deem suggestions without a-priori inductions imaginative suggestions (Braffman & Kirsch, 1999; Kirsch, 1997; Kirsch & Braffman, 2001). Nonetheless, the cold-control theory argues that it is not the a-priori induction that defines whether a response is hypnotic or not. According to cold control theory, if an intentional response is sensed as involuntary then the response is hypnotic regardless of the presence of an induction. This criterion can be empirically tested by including items in the screen that measure the feeling of involuntariness (e.g., Chapter IV; Polito, Barnier, & Woody, 2013). Future research could explore the extent to which the exclusion of the induction from the SWASH would be feasible and assess whether responses to the suggestions with and without a-priori induction are felt as involuntary.

The Bayes factor and hypothesis testing

This thesis applied the Bayes factor to draw conclusions about competing hypotheses, and this decision was only partly motivated by the practical advantages of the Bayes factor over the NHST that were mentioned in Chapter I (i.e., testing the null, and optional stopping). NHST consists of a clear set of rules with which one can make decisions (i.e., rejecting H0) with controlled long-term error rates, but this procedure will never tell us how strongly the evidence supports our pet theory, and this is the type of answer many researchers, including me, are looking for (Dienes, 2011). If a researcher is interested in whether the evidence at her disposal is supporting a baseline model, such as the null, or an alternative model, then the Bayes factor is the ideal statistical tool for her. To assess the evidence in favour of the competing hypotheses, one needs to clarify the predictions of these hypotheses. This can be done by considering the predictions in terms of standardised effect sizes (Rouder et al., 2009; Wagenmakers, 2007) or in terms of raw units of measurement (Dienes, 2008). In the former case, our model representing the predictions of the hypotheses incorporates information about signal as well as about noise. However, including information about measurement error makes it difficult for researchers to consider what exactly their theory predicts, as theories usually remain silent on that matter. The signal part, on the other hand, is easy to grasp and interpret with regard

to the theory (Baguley, 2009; Dienes, 2019). Therefore, all the Bayes factors reported in this thesis followed the approach in which predictions of H1 are considered in raw units of measurement.

Optional stopping, one of the practical benefits of the Bayes factor, came handy for Chapters II-IV. I conducted sequential analyses in all of these Chapters and stopped data collection either when the strength of evidence of the outcome neutral and crucial tests exceeded the cut-off of good enough evidence (Chapters II-III) or when I ran out of subjects (Chapter IV). In the latter case, a well supported conclusion cannot be drawn, however, data collection can be continued any time and stopped once we have good enough evidence for H0 or H1 (which we are in the process of doing). Recently, the claim that optional stopping is not a problem for Bayesians (Dienes, 2016; Rouder, 2014) has been challenged on the grounds that the Bayes factor is not well calibrated under certain conditions (Heide & Grünwald, 2017). Replies argue that the Bayes factor is well calibrated as long as the model representing the predictions of H1 is in line with our beliefs about H1 (Rouder, 2019; Wagenkakers, Gronau & Vandekerckhove, 2019). However, there are scenarios in which these two things are not aligned. For instance, one can use the heuristics applied in this thesis (see Dienes, 2019 for other heuristics) to model the predictions of H1 on a pilot study (e.g., Chapter IV) or on the results of earlier relevant studies (e.g., Chapters II-III). However, in some cases, one needs to resort to using the very same dataset to inspire the model of H1 and to test it.[31] Applying, for example, the room-to-move heuristic on our sample while we use optional stopping results in a dynamic estimate of the expected effect size. Every time we calculate the Bayes factor, the model of H1 will be slightly different as the room-to-move changes. This begs the question whether or not applying heuristics on our sample in tandem with optional stopping can lead us to miscalibrated conclusions. A future study should investigate this question and explore the extent to which dynamic models of H1 can miscalibrate the Bayes factor, and unravel how the size of the sample relates to this. Note that there is a sense in which the Bayes factor was never miscalibrated at any point in time: So long as at that point in time the model of H1 represents the best information about the plausibility of different effect sizes, the Bayes factor at that time is well calibrated with respect to the best model we have.

[31] Note that as long as one is using a different aspect of the data to model and test H1, this procedure is not invalidating the Bayes factor (Dienes, 2019).

The ability of the Bayes factor to distinguish between insensitive evidence and good enough evidence for H0 was exploited in Chapters II-IV. For instance, to conclude that a strategy cannot decrease Stroop interference, one needs a tool with which one can differentiate between data insensitivity and good enough evidence for the model predicting no Stroop reduction (Chapter III). Without such a tool, one could only conclude that the test failed to show evidence for the effect of a certain strategy in alleviating interference (cf. NHST). An even more interesting scenario arose in Chapter II that investigated the idea whether or not there is a minimum amount of cognitive conflict (indexed by interference on incongruent trials) that highs need to register to respond to the word blindness suggestion. Altering the proportion of incongruent trials allowed us to create blocks in which cognitive conflict occurred frequently and a block in which it rarely happened due to a low proportion of incongruent trials. If the amount of experienced conflict plays a role in the operation of the word blindness suggestion then we should find a difference between these blocks concerning the extent to which highs reduced Stroop interference. Moreover, if the minimum amount of necessary conflict idea is correct then highs should not be able to reduce Stroop interference in the low proportion incongruence block. This leaves us with a clear and consistent set of predictions: 1) no word blindness effect in the low proportion incongruence block; 2) word blindness effect in the high proportion incongruence block; 3) interaction of the block and the word blindness effect. However, what happens if we obtain good enough evidence supporting the first two claims but the test of the third claim remains insensitive? This is the issue that was discussed in Chapter VI. The tutorial highlights that we cannot conclude that the conditions differ unless the direct test, the test of the interaction, provides good enough evidence for it. This might strike us as being counterintuitive as it means that we cannot claim that the condition with no effect and the condition with an effect are different. However, as argued in Chapter VI, it is merely a temporary state that can be altered by continuing data collection until all tests are sensitive. Tutorials like this can illustrate Bayesian principles in a simple way, helping researchers use the Bayes factor to draw conclusions from their results. For instance, a recent study of Aczel et al. (2018) demonstrated that several papers published in high-tier journals in 2015 contained misrepresentations of non-significant results. Many claims about null findings were subserved with non-significant statistical tests, and only 14 from 137 papers (10%) reported Bayes factors to assess the strength of the evidence supporting H0. This later finding underlies that the Bayes factors has not yet reached its full potential and that there

is a need for clear and user friendly papers advocating the usage of the Bayes factor (e.g., Wagenmakers et al., 2018b).

A crucial weakness of the current tools with which researchers can calculate the Bayes factor is that they do not offer robust versions (cf. robust versions of NHST are recommended in many cases or even to be used as default tools by Wilcox, 2017; Field & Wilcox, 2017). To calculate the Bayes factor, we need to operate with many assumptions that may or may not meet reality. For instance, usually we model the data generating process with a normal distribution (or with a t distribution, which approximates the normal when we have a large sample size), and so we do not know the extent to which a Bayes factor is miscalibrated if we sample from populations that deviate from this. A future project should explore the extent to which Bayes factors are miscalibrated if skewness or kurtosis is introduced in the distribution representing the population effect sizes, while the Bayes factor is still calculated with a normal likelihood function. This project could potentially provide us with a rule of thumb helping researchers to assess whether or not their sample introduces bias (or large enough bias) in their results. A straightforward follow up study would be to test robust methods of the Bayes factor. For instance, trimming and winsorizing could be applied on our sample before we calculate the Bayes factor, and it could be tested whether these procedures can alleviate the bias introduced by violations of the assumptions.

Concluding remarks

Evidence presented in this thesis is in line with the prediction of cold control theory about hypnosis being a purely metacognitive phenomenon. The word blindness suggestion seemingly challenges this simple model of hypnosis, by providing an example in which highs can objectively perform better in a hypnotic than in a non-hypnotic way, therefore, understanding the underlying mechanisms of this suggestion is essential for theory testing. The word blindness suggestion operates via the reduction of response conflict rather than semantic conflict (Chapter II), and it is unlikely that highs look-away, focus on a single letter, blur their vision or internally rehearse the goal of the task to reduce Stroop interference when they respond to the suggestion (Chapter III). The presented evidence suggests that highs do not gain special abilities due to the word blindness suggestion, as they can reduce Stroop interference via imagining the words as being meaningless with and without being aware of doing so (Chapter IV). However, the

strength of the relevant evidence is not good enough, therefore, future research should settle this matter (research that is actually now ongoing). The online evaluation of hypnotic suggestibility is viable, as the online compared to offline measurement reduces responsiveness only to a small extent (Chapter V). Adapting the online hypnosis screening can help experimental hypnosis research to realise its full potential by, for instance, facilitating the conduction of large-scale hypnosis studies with more heterogeneous samples. Finally, when it comes to hypothesis testing with Bayes factors, one should bear in mind that the principle that any assertion about the existence of an interaction necessitates the comparison of the conditions is as true for Bayesian as it is for frequentist statistics (Chapter VI).

References

Abelson, R. A. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Acunzo, D., & Terhune, D. (2019). A critical review of standardized measures of hypnotic suggestibility. Retrieved from: https://psyarxiv.com/qge27/

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ... & Wagenmakers, E. J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science*, 2515245918773742.

Anderson, H. P., Seth, A. K., Dienes, Z., & Ward, J. (2014). Can grapheme-color synesthesia be induced by hypnosis?. *Frontiers in human neuroscience*, *8*, 220.

Anlló, H., Becchio, J., & Sackur, J. (2017). French norms for the Harvard Group Scale of hypnotic susceptibility, form A. *International Journal of Clinical and Experimental Hypnosis*, *65*(2), 241-255.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*(2), 235-244.

Arrow, K. J. (1951). *Social choice and individual values*. John Wiley & Sons, Inc.

Augustinova, M., & Ferrand, L. (2012). Suggestion does not de-automatize word reading: Evidence from the semantically based Stroop task. *Psychonomic Bulletin & Review*, *19*(3), 521–527.

Augustinova, M., & Ferrand, L. (2012b). The influence of mere social presence on Stroop interference: New evidence from the semantically-based Stroop task. *Journal of Experimental Social Psychology*, *48*(5), 1213-1216.

Augustinova, M., & Ferrand, L. (2014). Automaticity of word reading: Evidence from the semantic Stroop paradigm. *Current Directions in Psychological Science*, *23*(5), 343-348.

Augustinova, M., & Ferrand, L. (2014b). Social priming of dyslexia and reduction of the Stroop effect: What component of the Stroop effect is actually reduced?. *Cognition*, *130*(3), 442-454.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Baddeley, A. (2001). The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1345-1350.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 603–617.

Barber, T. X. (1966). The effects of 'hypnosis' and motivational suggestions on strength and endurance: a critical review of research studies. *British Journal of Clinical Psychology*, *5*(1), 42-50.

Barber, T. X., & Calverley, D. S. (1964). Comparative effects on" hypnotic-like" suggestibility of recorded and spoken suggestions. *Journal of Consulting Psychology*, *28*(4), 384.

Barber, T. X., & Calverley, D. S. (1966). Toward a theory of hypnotic behavior: experimental evaluation of Hull's postulate that hypnotic susceptibility is a habit phenomenon 1. *Journal of Personality*, *34*(3), 416-433.

Barber, T. X., & Glass, L. B. (1962). Significant factors in hypnotic behavior. *The Journal of Abnormal and Social Psychology*, *64*(3), 222.

Barber, T. X., & Wilson, S. C. (1978). The Barber suggestibility scale and the creative imagination scale: Experimental and clinical applications. *American Journal of Clinical Hypnosis*, *21*(2-3), 84-108.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, *71*(2), 230.

Barnier, A. J., Dienes, Z., & Mitchell, C. J. (2008). How hypnosis happens: New cognitive theories of hypnotic responding. In M. Heap., R. J. Brown & D. A. Oakley (Eds.), *The Oxford handbook of hypnosis: Theory, research, and practice* (pp. 141-177). London: Routledge.

Barnier, A. J., & McConkey, K. M. (1999). Absorption, hypnotizability and context: non-hypnotic contexts are not all the same. *Contemporary Hypnosis*, *16*(1), 1-8.

Barnier, A. J., & McConkey, K. M. (2003). Hypnosis, human nature, and complexity: Integrating neuroscience approaches into hypnosis research. *International Journal of Clinical and Experimental Hypnosis*, *51*(3), 282-308.

Barnier, A. J., & McConkey, K. M. (2004). Defining and identifying the highly hypnotizable person. *The highly hypnotizable person: Theoretical, experimental and clinical issues*, 30-60.

Benham, G., Woody, E. Z., Wilson, K. S., & Nash, M. R. (2006). Expect the unexpected: Ability, attitude, and responsiveness to hypnosis. *Journal of Personality and Social Psychology*, *91*(2), 342.

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of Nervous and Mental Disease*, *174*(12), 727–735.

Besner, D., Stolz, J. A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic bulletin & review*, *4*(2), 221-225.

Bohlmeijer, E., ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment*, *18*(3), 308–320.

Bor, D., Schwartzman, D. J., Barrett, A. B., & Seth, A. K. (2017). Theta-burst transcranial magnetic stimulation to the prefrontal or parietal cortex does not impair metacognitive visual awareness. *PloS one*, *12*(*2*), e0171793.

Bor, D., Barrett, A. B., Schwartzman, D. J., & Seth, A. K. (2018). Response to Ruby et al: On a ‘failed’attempt to manipulate conscious perception with transcranial magnetic stimulation to prefrontal cortex. *Consciousness and cognition*.

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

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

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Bowers, K. S. (1990). Unconscious influences and hypnosis. In J. L. Singer (Ed.), *Repression and dissociation: Implications for personality theory, psychopathology, and health* (pp. 143-178). Chicago: University of Chicago Press.

Bowers, K. S. (1992). Imagination and dissociation in hypnotic responding. *International Journal of Clinical and Experimental Hypnosis*, *40*(4), 253-275.

Bowers, K. S. (1993). The Waterloo-Stanford Group C (WSGC) scale of hypnotic susceptibility: Normative and comparative data. *International Journal of Clinical and Experimental Hypnosis*, *41*(1), 35–46.

Braffman, W., & Kirsch, I. (1999). Imaginative suggestibility and hypnotizability: an empirical analysis. *Journal of personality and social psychology*, *77*(3), 578.

Brown, T. L., Gore, C. L., & Carr, T. H. (2002). Visual attention and word recognition in Stroop color naming: Is word recognition" automatic?". *Journal of Experimental Psychology: General*, *131*(2), 220.

Brown, R. J., & Oakley, D. A. (2004). An integrative cognitive theory of hypnosis and high hypnotizability. In M. Heap, R. J. Brown, & D. A. Oakley (Eds.), *The highly hypnotizable person: Theoretical, experimental and clinical issues* (pp. 152–186). London: Brunner-Routledge.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, *84*(4), 822.

Bugg, J. M. (2012). Dissociating levels of cognitive control: The case of Stroop interference. *Current Directions in Psychological Science*, *21*(5), 302–309.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365.

Cardeña, E. (2014). Hypnos and psyche: How hypnosis has contributed to the study of consciousness. *Psychology of Consciousness: Theory, Research, and Practice*, *1*(2), 123.

Cardeña, E., & Terhune, D. B. (2009). A note of caution on the Waterloo-Stanford Group Scale of Hypnotic Susceptibility: A brief communication. *Intl. Journal of Clinical and Experimental Hypnosis*, *57*(2), 222-226.

Casiglia, E., Schiff, S., Facco, E., Gabbana, A., Tikhonoff, V., Schiavon, L., ... & Nasto, H. H. (2010). Neurophysiological correlates of post-hypnotic alexia: A controlled study with Stroop test. *American Journal of Clinical Hypnosis*, *52*(*3*), 219-233.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609-610.

Chambers, C. (2017). The seven deadly sins of psychology a manifesto for reforming the culture of scientific practice. Princeton, NJ: Princeton University Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective. In Q. Jing, M. Rosenzweig, G. d'Ydewalle, H. Zhang, H.-C. Chen, & K. Zhang (Eds.), *Progress in psychological science around the world: Vol. 1. Neural, cognitive, and developmental issues* (pp. 401–418). Hove, England: Psychology Press.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332.

Coltheart, M., Cox., R., Sowman, P., Morgan, H., Barnier, A., Langdon, R., Connaughton, E. , Teichmann, L., & Williams, N. (2018). Belief, delusion, hypnosis, and the right dorsolateral prefrontal cortex: a transcranial magnetic stimulation study. Registered Reports. *Cortex*

Cooper, L. M., & London, P. (1966). Sex and hypnotic susceptibility in children. *International Journal of Clinical and Experimental Hypnosis*, *14*(1), 55-60.

Comey, G., & Kirsch, I. (1999). Intentional and spontaneous imagery in hypnosis: The phenomenology of hypnotic responding. *International Journal of Clinical and Experimental Hypnosis*, *47*(1), 65-85.

Crick, F., & Koch, C. (1998). Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature*, *391*(6664), 245.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Davoli, C. C., & Abrams, R. A. (2009). Reaching out with the imagination. *Psychological Science*, *20*(3), 293-295.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it?. *Science*, *358*(6362), 486-492.

de Heide, R., & Grünwald, P. D. (2017). Why optional stopping is a problem for Bayesians. *arXiv preprint arXiv:1708.08278*.

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: Dissociable causes of interference effects in conflict situations. *Acta psychologica*, *101*(2-3), 379-394.

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: a dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 731.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). Mturk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note.

Derbyshire, S. W. (2000). Exploring the pain "neuromatrix". *Current review of pain*, *4*(6), 467-477.

Derbyshire, S. W., Whalley, M. G., & Oakley, D. A. (2009). Fibromyalgia pain and its modulation by hypnotic and non-hypnotic suggestion: An fMRI analysis. *European Journal of Pain*, *13*(5), 542-550.

Derbyshire, S. W., Whalley, M. G., Stenger, V. A., & Oakley, D. A. (2004). Cerebral activation during hypnotically induced and imagined pain. *Neuroimage*, *23*(1), 392-401.

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. UK: Allen Lane.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, *6*(3), 274-290.

Dienes, Z. (2012). Is hypnotic responding the strategic relinquishment of metacognition?. In Beran, M., Brandl, J. L., Perner, J., & Proust, J. (Eds.), *Foundations of metacognition* (pp. 267-277). Oxford University Press.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.

Dienes, Z (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), Behavioural Methods in Consciousness Research. Oxford: 760 Oxford University Press , pp 199-220.

Dienes, Z. [Zoltan Dienes]. (2015, April 23). *How many participants might I need?* [Video file]. Retrieved from https://www.youtube.com/watch?v=10Lsm_o_GRg

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78-89.

Dienes, Z. (2019). How do I know what my theory predicts?. (Preprint: https://psyarxiv.com/yqaj4/)

Dienes, Z., Brown, E., Hutton, S., Kirsch, I., Mazzoni, G., & Wright, D. B. (2009). Hypnotic suggestibility, cognitive inhibition, and dissociation. *Consciousness and cognition*, *18*(4), 837-847.

Dienes, Z., & Hutton, S. (2013). Understanding hypnosis metacognitively: rTMS applied to left DLPFC increases hypnotic suggestibility. *Cortex*, *49*(2), 386-392.

Dienes, Z., Lush, P., Palfi, B., Rooseboom, W., Scottt, R., Parris, B., Seth, A., & Lovell, M. (2019). Phenomenological control as cold control. Manuscript submitted for publication.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207-218.

Dienes, Z., & Perner, J. (2007). Executive control without conscious awareness: the cold control theory of hypnosis. In Jamieson, G. (Ed.), *Hypnosis and conscious states: The cognitive neuroscience perspective*, (pp. 293-314). Oxford University Press.

Dijksterhuis, A., & Van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of personality and social psychology*, *74*(4), 865.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PloS one*, *7*(1), e29081.

Duthoo, W., Abrahamse, E. L., Braem, S., Boehler, C. N., & Notebaert, W. (2014). The heterogeneous world of congruency sequence effects: an update. *Frontiers in Psychology*, *5*, 1001.

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 380–390.

Egner, T., Ely, S., & Grinband, J. (2010). Going, going, gone: characterizing the time-course of congruency sequence effects. *Frontiers in psychology*, *1*, 154.

Egner, T., & Raz, A. (2007). Cognitive control processes and hypnosis. *Hypnosis and conscious states: The cognitive neuroscience perspective*, 29-50.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of learning and motivation*, *44*, 145-200.

Erdelyi, M. H. (1994). Hypnotic hypermnesia: The empty set of hypermnesia. *International Journal of Clinical and Experimental Hypnosis*, *42*(4), 379-390.

Erickson, M. H., Rossi, E.L., & Rossi, S.I. (1976). *Hypnotic Realities: The Induction of Hypnosis and Forms of Indirect Suggestions*. New York, NY: Irvington.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143-149.

Fassler, O., Lynn, S. J., & Knox, J. (2008). Is hypnotic suggestibility a stable trait? *Consciousness and Cognition*, *17*(1), 240–253.

Feynman, R. P., & Leighton, R. (1992). *" Surely you're joking, Mr. Feynman!": adventures of a curious character*. Random House.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy*, *98*, 19-38.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1935). The design of experiments. Tweeddale: Oliver and Boyd.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906-911.

Flaudias, V., & Llorca, P. M. (2014). A brief review of three manipulations of the Stroop task focusing on the automaticity of semantic access. *Psychologica Belgica*, *54*(2).

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811-2822.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Gfeller, J. D., Lynn, S. J., & Pribble, W. E. (1987). Enhancing hypnotic susceptibility: Interpersonal and rapport factors. *Journal of Personality and Social Psychology*, *52*(3), 586.

Goldfarb, L., Aisenberg, D., & Henik, A. (2011). Think the thought, walk the walk–Social priming reduces the Stroop effect. *Cognition*, *118*(2), 193-200.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*(4), 480.

Gratton, G., Cooper, P., Fabiani, M., Carter, C. S., & Karayanidis, F. (2018). Dynamics of cognitive control: Theoretical bases, paradigms, and a view for the future. *Psychophysiology*, *55*(3), e13016.

Green, J. P. (2004). The five factor model of personality and hypnotizability: little variance in common. *Contemporary Hypnosis*, *21*(4), 161-168.

Green, J. P., & Lynn, S. J. (2010). Hypnotic responsiveness: Expectancy, attitudes, fantasy proneness, absorption, and gender. *International Journal of Clinical and Experimental Hypnosis*, *59*(1), 103-121.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337-350.

Haggard, P., Cartledge, P., Dafydd, M., & Oakley, D. A. (2004). Anomalous control: when 'free-will'is not conscious. *Consciousness and cognition*, *13*(3), 646-654.

Hilgard, E. R. (1965). *Hypnotic susceptibility*. New York, NY: Harcourt, Brace & World, Inc.

Hilgard, J. R. (1970). *Personality and hypnosis: A study of imaginative involvement*. Chicago: University of Chicago Press.

Hilgard, J. R. (1974). Sequelae to hypnosis. *International Journal of Clinical and Experimental Hypnosis*, *22*(4), 281-298.

Hilgard, E. R. (1977). The problem of divided consciousness: A neodissociation interpretation. *Annals of the New York Academy of Sciences*, *296*(1), 48-59.

Hilgard, E. R. (1991). A neodissociation interpretation of hypnosis. In S. J. Lynn & J. W. Rhue (Eds.), *The Guilford clinical and experimental hypnosis series. Theories of hypnosis: Current models and perspectives* (pp. 83-104). New York: Guilford Press.

Hilgard, E. R., & Tart, C. T. (1966). Responsiveness to suggestions following waking and imagination instructions and following induction of hypnosis. *Journal of Abnormal Psychology*, *71*(3), 196.

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037.

Huguet, P., Dumas, F., & Monteil, J-M. (2004). Competing for a desired reward in the Stroop task: When attentional control is unconscious but effective versus conscious but ineffective. *Canadian Journal of Experimental Psychology*, *58*(3), 153-167.

Iani, C., Ricci, F., Gherri, E., & Rubichi, S. (2006). Hypnotic suggestion modulates cognitive conflict: the case of the flanker compatibility effect. *Psychological Science*, *17*(8), 721-727.

Iani, C., Ricci, F., Baroni, G., & Rubichi, S. (2009). Attention control and susceptibility to hypnosis. *Consciousness and Cognition*, *18*(4), 856-863.

Jack, A. I., & Shallice, T. (2001). Introspective physicalism as an approach to the science of consciousness. *Cognition*, *79*(1-2), 161-196.

Jacobs, C., Schwarzkopf, D. S., & Silvanto, J. (2018). Visual working memory performance in aphantasia. *Cortex, 105*, 61-73.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of memory and language*, *30*(5), 513-541.

Jamieson, G. A., & Sheehan, P. W. (2002). A critical evaluation of the relationship between sustained attentional abilities and hypnotic susceptibility. *Contemporary Hypnosis*, *19*(2), 62-74.

Jamieson, G. A., & Sheehan, P. W. (2004). An empirical test of Woody and Bowers's dissociated-control theory of hypnosis. *Int J of Clinical & Exp Hypnosis*, *52*(3), 232-249.

Jamieson, G. A., & Woody, E. (2007). Dissociated control as a paradigm for cognitive neuroscience research and theorizing in hypnosis. *Hypnosis and conscious states: The cognitive neuroscience perspective*, 111-129.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control: A computational theory of the Stroop task. *Psychological review*, *125*(1), 59.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General*, *132*(1), 47.

Kihlstrom, J. F. (1985). Hypnosis. *Annual review of psychology*, *36*(1), 385-418.

Kihlstrom, J. F. (1998). Dissociations and dissociation theory in hypnosis: Comment on Kirsch and Lynn (1998). *Psychological Bulletin*, *123*(2), 186-191.

Kihlstrom, J. F. (2001). Hypnosis and the psychological unconscious. In *Assessment and Therapy* (pp. 215-225). Academic Press.

Kihlstrom, J. F. (2005). Is hypnosis an altered state of consciousness or what?. *Contemporary Hypnosis*, *22*(1), 34-38.

Kihlstrom, J. F. (2008). The domain of hypnosis, revisited. *The Oxford handbook of hypnosis: Theory, research, and practice*, 21-52.

Kinnunen, T., Zamansky, H. S., & Block, M. L. (1994). Is the hypnotized subject lying?. *Journal of Abnormal Psychology*, *103*(2), 184.

Kinoshita, S., Mills, L., & Norris, D. (2018). The semantic Stroop effect is controlled by endogenous attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Kirsch, I. (1985). Response expectancy as a determinant of experience and behavior. *American Psychologist*, *40*(11), 1189.

Kirsch, I. (1997). Response expectancy theory and application: A decennial review. *Applied and preventive Psychology*, *6*(2), 69-79.

Kirsch, I. (1999). Hypnosis and placebos: response expectancy as a mediator of suggestion effects. *Anales de Psicología/Annals of Psychology*, *15*(1), 99-110.

Kirsch, I. (2011). The altered state issue: Dead or alive?. *International Journal of Clinical and Experimental Hypnosis*, *59*(3), 350-362.

Kirsch, I., & Braffman, W. (2001). Imaginative suggestibility and hypnotizability. *Current directions in psychological science*, *10*(2), 57-61.

Kirsch, I. E., Capafons, A. E., Cardeña-Buelna, E. E., & Amigó, S. E. (1999). *Clinical hypnosis and self-regulation: Cognitive-behavioral perspectives*. American Psychological Association.

Kirsch, I., & Lynn, S. J. (1997). Hypnotic involuntariness and the automaticity of everyday life. *American Journal of Clinical Hypnosis*, *40*(1), 329-348.

Kirsch, I., Mazzoni, G., Roberts, K., Dienes, Z., Hallquist, M. N., Williams, J., & Lynn, S. J. (2008). Slipping into trance. *Contemporary Hypnosis*, *25*(3-4), 202-209.

Kirsch, I., Milling, L. S., & Burgess, C. (1998). Experiential scoring for the Waterloo-Stanford Group C scale. *International Journal of Clinical and Experimental Hypnosis*, *46*(3), 269-279.

Kirsch, I., Montgomery, G., & Sapirstein, G. (1995). Hypnosis as an adjunct to cognitive-behavioral psychotherapy: A meta-analysis. *Journal of consulting and clinical psychology*, *63*(2), 214.

Kirsch, I., Silva, C. E., Carone, J. E., Johnston, J. D., & Simon, B. (1989). The surreptitious observation design: An experimental paradigm for distinguishing artifact from essence in hypnosis. *Journal of Abnormal Psychology*, *98*(2), 132.

Kosslyn, S. M., Thompson, W. L., Costantini-Ferrando, M. F., Alpert, N. M., & Spiegel, D. (2000). Hypnotic visual illusion alters color processing in the brain. *American Journal of Psychiatry*, *157*(8), 1279-1284.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.

Krebs, R. M., Boehler, C. N. & Woldorff, M. G. (2010). The influence of reward associations on conflict processing in the Stroop task. *Cognition, 117*, 341-347.

Kriegel, U. (2007). A cross-order integration hypothesis for the neural correlate of consciousness. *Consciousness and Cognition*, *16*(4), 897-912.

Kruschke, J. K. (2010). Doing Bayesian data analysis: A tutorial with R and BUGS. Burlington, MA: Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178-206.

Kunde, W., Reuss, H., & Kiesel, A. (2012). Consciousness and cognitive control. *Advances in cognitive psychology*, *8*(1), 9.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, gby065.

Lau, H. (2011). Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 1-7.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763-18768.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, *15*(8), 365-373.

Lau, H., & Rosenthal, D. (2011b). The higher-order view does not require consciously self-directed introspection: response to Malach. *Trends in cognitive sciences*, *15*(11), 508-509.

Laurence, J.-R., Beaulieu-Prévost, D., & du Chéné, T. (2008). Measuring and understanding individual differences in hypnotizability. In M. R. Nash & A. Barnier (Eds.), *The Oxford handbook of hypnosis* (pp. 225-253). Oxford: Oxford University Press.

Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge university press.

Levinson, D. B., Stoll, E. L., Kindy, S. D., Merry, H. L., & Davidson, R. J. (2014). A mind you can count on: validating breath counting as a behavioral measure of mindfulness. *Frontiers in psychology*, *5*.

Levitt, E. E., & Brady, J. P. (1964). Muscular endurance under hypnosis and in the motivated waking state. *International Journal of Clinical and Experimental Hypnosis*, *12*(1), 21-27.

Lewis, I. M. (2003). Trance, possession, shamanism and sex. *Anthropology of Consciousness*, *14*(1), 20-39.

Lifshitz, M., Bonn, N. A., Fischer, A., Kashem, I. F., & Raz, A. (2013). Using suggestion to modulate automatic processes: From Stroop to McGurk and beyond. *Cortex*, *49*(2), 463-473.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, *39*, 367–386.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & cognition*, *7*(3), 166-174.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

Lush, P., Botan, V., Scott, R. B., Seth, A., Ward, J., & Dienes, Z. (2019). Phenomenological control: response to imaginative suggestion predicts measures of mirror touch synaesthesia, vicarious pain and the rubber hand illusion. Manuscript submitted for publication.

Lush, P., Caspar, E. A., Cleeremans, A., Haggard, P., Magalhães De Saldanha da Gama, P. A., & Dienes, Z. (2017). The power of suggestion: posthypnotically induced changes in the temporal binding of intentional action outcomes. *Psychological science*, *28*(5), 661-669.

Lush, P., Moga, G., McLatchie, N., & Dienes, Z. (2018). The Sussex-Waterloo Scale of Hypnotizability (SWASH): measuring capacity for altering conscious experience. *Neuroscience of Consciousness*, *2018*(1), niy006.

Lush, P., Scott, R. B., Anil, S., & Dienes, Z. (2019). *Is the rubber hand illusion a suggestion effect?*. Manuscript in preparation.

Lush, P., Scott, R. B., Moga, G., & Dienes, Z. (2018). *Norms for a computerized version of the SWASH*. Manuscript in preparation.

Lynn, S. J. (1997). Automaticity and hypnosis: A sociocognitive account. *International Journal of Clinical and Experimental Hypnosis*, *45*(3), 239-250.

Lynn, S. J., Green, J. P., Polizzi, C., Ellenberg, S., Gautam, A. & Aksen, D. (in press). Hypnosis, Hypnotic Phenomena, and Hypnotic Responsiveness: Clinical and Research Foundations--A 40 Year Perspective. *International Journal of Clinical & Experimental Hypnosis*.

Lynn, S. J., Weekes, J. R., Matyi, C. L., & Neufeld, V. (1988). Direct versus indirect suggestions, archaic involvement, and hypnotic experience. *Journal of Abnormal Psychology*, *97*(3), 296.

Macleod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. *Memory & Cognition*, *26*(2), 201-211.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163-203.

MacLeod, C. M. (1992). The Stroop task: The" gold standard" of attentional measures. *Journal of Experimental Psychology: General*, *121*(1), 12.

MacLeod, C. M. (2011). Hypnosis and the control of attention: Where to from here?. *Consciousness and Cognition*, *20*(2), 321-324.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126.

Magalhães De Saldanha da Gama, P. A., Slama, H., Caspar, E. A., Gevers, W., & Cleeremans, A. (2013). Placebo-suggestion modulates conflict resolution in the Stroop task. *PLoS ONE*, *8*(10), e75701.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive psychology*, *15*(2), 197-237.

Martin, J. R., & Dienes, Z. (2018). *Bayes to the rescue: Does the type of hypnotic induction matter?*. Manuscript submitted for publication.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science*, *14*(2), 186-188.

McConkey, K., Barnier, A. J., Maccallum, F. L., & Bishop, K. (1996). A normative and structural analysis of the HGSHS: A with a large Australian sample. *Australian Journal of Clinical & Experimental Hypnosis*.

McConkey, K. M., & Sheehan, P. W. (1995). *Hypnosis, memory, and behavior in criminal investigation*. New York: Guilford Press.

McConkey, K. M., Szeps, A., & Barnier, A. J. (2001). Indexing the experience of sex change in hypnosis and imagination. *International Journal of Clinical and Experimental Hypnosis*, *49*(*2*), 123-138.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175.

Miller, M. E., & Bowers, K. S. (1993). Hypnotic analgesia: Dissociated experience or dissociated control?. *Journal of Abnormal Psychology*, *102*(1), 29.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Milling, L. S., Kirsch, I., Meunier, S. A., & Levine, M. R. (2002). Hypnotic analgesia and stress inoculation training: Individual and combined effects in analog treatment of experimental pain. *Cognitive Therapy and Research*, *26*(3), 355-371.

Montgomery, G. H., Duhamel, K. N., & Redd, W. H. (2000). A meta-analysis of hypnotically induced analgesia: How effective is hypnosis?. *International Journal of Clinical and Experimental Hypnosis*, *48*(2), 138-153.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, *132*(2), 297.

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *Journal of Neuroscience*, *38*(14), 3534-3546.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103-123.

Morgan, A. H., & Hilgard, E. R. (1973). Age differences in susceptibility to hypnosis. *International journal of clinical and experimental Hypnosis*, *21*(2), 78-85.

Munafò, Marcus R., Brian A. Nosek, Dorothy VM Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John PA Ioannidis. "A manifesto for reproducible science." *Nature human behaviour* 1, no. 1 (2017): 0021.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition, 25*, 27–41.

Neely, J. H., & Kahan, T. A. (2001). Is semantic activation automatic? A critical re-evaluation. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), Science conference series. The nature of remembering: Essays in honor of Robert G. Crowder (pp. 69-93). Washington, DC, US: American Psychological Association.

Nelson, T. O. & Narens, L. (1990) Metamemory: A theoretical framework and new findings. In G. Bower. (Eds.), *The psychology of learning and motivation* (pp. 1-45). New York: Academic.

New Psychological Science Editor Plans to Further Expand the Journal's Reach. (2019, August 22). Retrieved from https://www.psychologicalscience.org/observer/new-psychological-science-editor-plans-to-further-expand-the-journals-reach

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1-19.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, *231*(694-706), 289-337.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*(9), 1105.

Nogrady, H., McConkey, K. M., & Perry, C. (1985). Enhancing visual memory: Trying hypnosis, trying imagination, and trying again. *Journal of Abnormal Psychology, 94*(2), 195-204.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1–17). New York, NY: Plenum Press.

Oakley, D. A. (2006). Hypnosis as a tool in research: experimental psychopathology. *Contemporary Hypnosis*, *23*(1), 3-14.

Oakley, D. A., & Halligan, P. W. (2009). Hypnotic suggestion and cognitive neuroscience. *Trends in cognitive sciences*, *13*(6), 264-270.

Oakley, D. A., & Halligan, P. W. (2013). Hypnotic suggestion: opportunities for cognitive neuroscience. *Nature Reviews Neuroscience*, *14*(8), 565.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155–164.

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *The Journal of abnormal and social Psychology*, *58*(3), 277.

Orne, M. T., Sheehan, P. W., & Evans, F. J. (1968). Occurrence of posthypnotic behavior outside the experimental setting. *Journal of Personality and Social Psychology*, *9*(2p1), 189.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Page, R. A., & Green, J. P. (2007). An update on age, hypnotic suggestibility, and gender: a brief report. *American Journal of Clinical Hypnosis*, *49*(4), 283-287.

Palfi, B., Parris, B. A., & Dienes, Z. (2019). *Strategies that reduce Stroop interference.* (Manuscript in preparation)

Palfi, B., Parris, B. A., McLatchie, N., Kekecs, Z., & Dienes, Z. (2018). Can unconscious intentions be more effective than conscious intentions? Test of the role of metacognition in hypnotic response. *Cortex*, (Stage 1 Registered Report).

Palfi, B., Parris, B. A., Seth, A. K., & Dienes, Z. (2018). *Does unconscious control depend on conflict?*. (Manuscript in preparation. Preprint: https://psyarxiv.com/a68js/).

Parris, B. A., Bate, S., Brown, S. D., & Hodgson, T. L. (2012). Facilitating goal-oriented behaviour in the Stroop task: when executive control is influenced by automatic processing. *PloS one*, *7*(10), e46994.

Parris, B. A., Dienes, Z., Bate, S., & Gothard, S. (2014). Oxytocin impedes the effect of the word blindness post-hypnotic suggestion on Stroop task performance. *Social cognitive and affective neuroscience*, *9*(7), 895-899.

Parris, B. A., Dienes, Z., & Hodgson, T. L. (2012). Temporal constraints of the word blindness posthypnotic suggestion on Stroop task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 833.

Parris, B. A., & Dienes, Z. (2013). Hypnotic suggestibility predicts the magnitude of the imaginative word blindness suggestion effect in a non-hypnotic context. *Consciousness and cognition*, *22*(3), 868-874.

Parris, B. A., Dienes, Z., & Hodgson, T. L. (2013). Application of the ex-Gaussian function to the effect of the word blindness suggestion on Stroop task performance suggests no word blindness. *Frontiers in Psychology*, *4*, 647.

Parris, B. A., Sharma, D., & Weekes, B. (2007). An optimal viewing position effect in the Stroop task when only one letter is the color carrier. *Experimental Psychology*, *54*(4), 273-280.

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774.

Perner, J. (1991). *Understanding the representational mind*. MIT Press.

Perry, C., & Laurence, J. R. (1980). Hypnotic depth and hypnotic susceptibility: A replicated finding. *International Journal of Clinical and Experimental Hypnosis*, *28*(3), 272-280.

Perugini, E. M., Kirsch, I., Allen, S. T., Coldwell, E., Meredith, J. M., Montgomery, G. H., & Sheehan, J. (1998). Surreptitious observation of responses to

hypnotically suggested hallucinations: A test of the compliance hypothesis. *International Journal of Clinical and Experimental Hypnosis*, *46*(2), 191-203.

Piccione, C., Hilgard, E. R., & Zimbardo, P. G. (1989). On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality and Social Psychology*, *56*(2), 289.

Ploszay, A. J., Gentner, N. B., Skinner, C. H., & Wrisberg, C. A. (2006). The effects of multisensory imagery in conjunction with physical movement rehearsal on golf putting performance. *Journal of Behavioral Education*, *15*(4), 247-255.

Polito, V., Barnier, A. J., & Woody, E. Z. (2013). Developing the Sense of Agency Rating Scale (SOARS): An empirical measure of agency disruption in hypnosis. *Consciousness and cognition*, *22*(3), 684-696.

Polito, V., Barnier, A. J., Woody, E. Z., & Connors, M. H. (2014). Measuring agency change across the domain of hypnosis. *Psychology of Consciousness: Theory, Research, and Practice*, *1*(1), 3.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), Information processing and cognition: The Loyola symposium (pp. 153–175). Hillsdale, NJ: Erlbaum.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception, & Psychophysics*, *72*(7), 2013-2025.

Puccioni, O., & Vallesi, A. (2012). Sequential congruency effects: disentangling priming and conflict adaptation. *Psychological Research*, *76*(5), 591-600.

Raine, A., & Benishay, D. (1995). The SPQ-B: a brief screening instrument for schizotypal personality disorder. *Journal of personality disorders*, *9*(4), 346-355.

Raz, A. (2011). Hypnosis: a twilight zone of the top-down variety: Few have never heard of hypnosis but most know little about the potential of this mind–body regulation technique for advancing science. *Trends in cognitive sciences*, *15*(12), 555-557.

Raz, A., & Campbell, N. K. (2011). Can suggestion obviate reading? Supplementing primary Stroop evidence with exploratory negative priming analyses. *Consciousness and cognition*, *20*(*2*), 312-320.

Raz, A., Fan, J., & Posner, M. I. (2005). Hypnotic suggestion reduces conflict in the human brain. *Proceedings of the national Academy of Sciences of the United States of America*, *102*(28), 9978-9983.

Raz, A., Landzberg, K. S., Schweizer, H. R., Zephrani, Z. R., Shapiro, T., Fan, J., & Posner, M. I. (2003). Posthypnotic suggestion and the modulation of Stroop interference under cycloplegia. *Consciousness and cognition*, *12*(*3*), 332-346.

Raz, A., Kirsch, I., Pollard, J., & Nitkin-Kaner, Y. (2006). Suggestion reduces the Stroop effect. *Psychological Science*, *17*(2), 91-95.

Raz, A., Landzberg, K. S., Schweizer, H. R., Zephrani, Z. R., Shapiro, T., Fan, J., & Posner, M. I. (2003). Posthypnotic suggestion and the modulation of Stroop interference under cycloplegia. *Consciousness and cognition*, *12*(*3*), 332-346.

Raz, A., Shapiro, T., Fan, J., & Posner, M. I. (2002). Hypnotic suggestion and the modulation of Stroop interference. Archives of General Psychiatry, *59*(12), 1155–1161.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). San Diego, CA: Academic Press.

Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychological review*, *110*(1), 88.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*(3), 329–359.

Rosenthal, D. M. (2002). Consciousness and Higher-Order Thought. *Encyclopedia of Cognitive Science*.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia*, *46*(3), 829-840.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-308.

Rouder, J. (2019). On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald. (Preprint: https://psyarxiv.com/m6dhw/)

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference?. *Topics in Cognitive Science*, *8*(3), 520-547.

Rouder, J., Morey, R., & Wagenmakers, E. J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra: Psychology*, *2*(1).

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195-223.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225-237.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive neuroscience*, *1*(3), 165-175.

Ruby, E., Maniscalco, B., & Peters, M. A. (2018). On a 'failed'attempt to manipulate visual metacognition with transcranial magnetic stimulation to prefrontal cortex. *Consciousness and cognition*, *62*, 34-41. Spanos, N. P. (1986). Hypnotic behavior: A social-psychological interpretation of amnesia, analgesia, and "trance logic". *Behavioral and Brain Sciences*, *9*(3), 449-467.

Rudski, J. M., Marra, L. C., & Graham, K. R. (2004). Sex differences on the HGSHS: A. *International Journal of Clinical and Experimental Hypnosis*, *52*(1), 39-46.

Sarbin, T. R., & Coe, W. C. (1972). *Hypnosis: A social psychological analysis of influence communication*. Holt, Rinehart and Winston.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, *25*(1), 128-142.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*(6), 657-680.

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies?.

Semmens-Wheeler, R., Dienes, Z., & Duka, T. (2013). Alcohol increases hypnotic susceptibility. *Consciousness and cognition*, *22*(3), 1082-1091.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., ... & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PloS one*, *8*(4), e56515.

Sheehan, P. W., Donovan, P., & MacLeod, C. M. (1988). Strategy manipulation and the Stroop effect in hypnosis. *Journal of Abnormal Psychology*, *97*(4), 455.

Shor, R. E., & Orne, E. C. (1963). Norms on the Harvard Group Scale of Hypnotic Susceptibility, Form A. *International Journal of Clinical and Experimental Hypnosis*, *11*(*1*), 39-47.

Shor, R. E., Pistole, D. D., Easton, R. D., & Kihlstrom, J. F. (1984). Relation of predicted to actual hypnotic responsiveness, with special reference to posthypnotic amnesia. *International Journal of Clinical and Experimental Hypnosis*, *32*(4), 376-387.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Simon, J. R., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics*, *6*(1), 99-105.

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2015). *Specification curve: descriptive and inferential statistics on all reasonable specifications* (Working paper). doi:10.2139/ssrn.2694998

Slusher, M. P., & Anderson, C. A. (1987). When reality monitoring fails: The role of imagination in stereotype maintenance. *Journal of Personality and Social Psychology*, *52*(4), 653.

Spanos, N. P. (1986). Hypnotic behavior: A social-psychological interpretation of amnesia, analgesia, and "trance logic". *Behavioral and Brain Sciences*, *9*(3), 449-467.

Spanos, N. P., & Barber, T. X. (1974). Toward a convergence in hypnosis research. *American Psychologist*, *29*(7), 500-511.

Spanos, N. P., Radtke, H. L., Hodgins, D. C., Stam, H. J., & Bertrand, L. D. (1983). The Carleton University Responsiveness to Suggestion Scale: normative data and psychometric properties. *Psychological Reports*, *53*(2), 523-535.

Spinhoven, P. (1987). Hypnosis and behavior therapy: A review. *International journal of clinical and experimental hypnosis*, *35*(1), 8-31.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the Stroop task: evidence from ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1398.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108-112.

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *The Quarterly Journal of Experimental Psychology*, *31*(1), 121-132.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Taylor, S. E., Pham, L. B., Rivkin, I. D., & Armor, D. A. (1998). Harnessing the imagination: Mental simulation, self-regulation, and coping. *American psychologist*, *53*(4), 429.

Terhune, D. B., & Cardeña, E. (2016). Nuances and uncertainties regarding hypnotic inductions: toward a theoretically informed praxis. *American Journal of Clinical Hypnosis*, *59*(2), 155-174.

Terhune, D. B., Cleeremans, A., Raz, A., & Lynn, S. J. (2017). Hypnosis and top-down regulation of consciousness. *Neuroscience & Biobehavioral Reviews*, *81*, 59-74.

Terhune, D. B., Luke, D. P., & Cohen Kadosh, R. C. (2017). The induction of synaesthesia in non-synaesthetes. In Deroy, O. (Ed.), *Sensory Blending: On Synaesthesia and Related Phenomena*, (pp. 215-247). Oxford University Press.

The Stand. (2017, November 4). *The Economist*. Retrieved from https://www.economist.com/news/essays/Luther

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory & cognition*, *20*(6), 727-735.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2016). Bayesian inference for Kendall's rank correlation coefficient. The American Statistician. Advance online publication. doi:10.1080/00031305.2016.1264998

van Gaal, S., De Lange, F. P., & Cohen, M. X. (2012). The role of consciousness in cognitive control and decision making. *Frontiers in human neuroscience*, *6*, 121.

Van Gaal, S., Lamme, V. A., & Ridderinkhof, K. R. (2010). Unconsciously triggered conflict adaptation. *PloS one*, *5*(7), e11508.

Van Gulick, R. (1994). Deficit Studies and the Function of Phenomenal. In G. Graham, L. Stephens (Ed.), Philosophical psychology (pp. 25-50). Cambridge, MA: MIT Press.

Varga, K., Németh, Z., & Szekely, A. (2011). Lack of correlation between hypnotic susceptibility and various components of attention. Consciousness and Cognition, 20, 1872–1881.

Veling, H. & Aarts, H. (2010). Cueing task goals and earning money: Relatively high monetary rewards reduces failures to act on goals in a Stroop task. *Motivation and Emotion, 34*, 184–190.Witt, J. K., & Proffitt, D. R. (2008). Action-specific influences on

distance perception: a role for motor simulation. *Journal of experimental psychology: Human perception and performance*, *34*(6), 1479.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779-804.

Wagenmakers, E. J., Gronau, Q. F., & Vandekerckhove, J. (2019). Five Bayesian Intuitions for the Stopping Rule Principle. (Preprint: https://psyarxiv.com/5ntkd)

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35-57.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Meerhoff, F. (2018b). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review*, *25*(1), 58-76.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169-176.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

Wagstaff, G. F., Cole, J. C., & Brunas-Wagstaff, J. (2008). Measuring hypnotizability: The case for self-report depth scales and normative data for the Long Stanford Scale. *Intl. Journal of Clinical and Experimental Hypnosis*, *56*(2), 119-142.

Weitzenhoffer, A. M. (1949). The production of antisocial acts under hypnosis. *The Journal of Abnormal and Social Psychology*, *44*(3), 420.

Weitzenhoffer, A. M. (1974). When is an "instruction" an "instruction"?. *International Journal of Clinical and Experimental Hypnosis*, *22*(3), 258-269.

Weitzenhoffer, A. M. (1980). Hypnotic susceptibility revisited. *American Journal of Clinical Hypnosis*, *22*(3), 130-146.

Weitzenhoffer, A. M., & Hilgard, E. R. (1962). *Stanford hypnotic susceptibility scale, form C* (Vol. 27). Palo Alto, CA: Consulting Psychologists Press.

White, D., Risko, E. F., & Besner, D. (2016). The semantic Stroop effect: An ex-Gaussian analysis. *Psychonomic Bulletin & Review*, *23*(5), 1576-1581.Zahedi, A., Stuermer, B., Hatami, J., Rostami, R., & Sommer, W. (2017). Eliminating stroop effects with post-hypnotic instructions: Brain mechanisms inferred from EEG. *Neuropsychologia*, *96*, 70-77.

Wilcox, R. R. (2017). Introduction to robust estimation and hypothesis testing (4th ed.), Elsevier, Burlington, MA

Wilson, S. C., & Barber, T. X. (1978). The Creative Imagination Scale as a measure of hypnotic responsiveness: Applications to experimental and clinical hypnosis. *American Journal of Clinical Hypnosis*, *20*(4), 235-249.

Woody, E. Z., Bowers, K. S., Lynn, S. J., & Rhue, J. W. (1994). *A frontal assault on dissociated control* (pp. 52-79). Guilford Press.

Woody, E.Z., & Sadler, P. (2008) Dissociation theories of hypnosis. In Nash, M. R., & Barnier, A. J. (Eds.), *The Oxford handbook of hypnosis: Theory, research, and practice*, (pp. 81 - 110). Oxford University Press.

Zahedi, A., Stuermer, B., Hatami, J., Rostami, R., & Sommer, W. (2017). Eliminating stroop effects with post-hypnotic instructions: Brain mechanisms inferred from EEG. *Neuropsychologia*, *96*, 70-77.

Zamansky, H. S., Scharf, B., & Brightbill, R. (1964). The effect of expectancy for hypnosis on prehypnotic performance. *Journal of Personality*, *32*(2), 236-248.

Zeman, A. Z., Dewar, M., & Della Sala, S. (2015). Lives without imagery-Congenital aphantasia.

Zeman, A. Z., Della Sala, S., Torrens, L. A., Gountouna, V. E., McGonigle, D. J., & Logie, R. H. (2010). Loss of imagery phenomenology with intact visuo-spatial task performance: A case of ‘blind imagination’. *Neuropsychologia*, *48*(1), 145-155.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods*, *12*(4), 399-413.

Supplementary Materials of Chapter II

Supplementary Question

As a secondary question, we aimed to assess whether the participants experience any depth of hypnosis (Hilgard & Tart, 1966) during the suggestion conditions. We applied posthypnotic suggestions and so we expected that the participants should not experience of being in a hypnotic state while they respond to the suggestion (e.g., Terhune, Luke & Cohen Kadosh, 2017). This secondary question is not related to the main research question, so we present the results and their interpretation here, in the Supplementary Materials labelled as Exploration 4.

Supplementary Methods

Data analysis. As an exploratory analysis, we aimed to probe the extent to which the word blindness suggestion might influences the Stroop facilitation effect. Hence, we calculated the Stroop facilitation scores in each condition and block separately (mean of neutral RTs - mean of congruent RTs) as we did for the main analyses with the interference scores. The analyses were in accordance with the main tests we ran on the Stroop interference scores of the merged dataset, namely, all of these tests are direct comparisons of two conditions and so can be conducted by running paired samples t-tests.

Moreover, we conducted the outcome neutral and crucial tests; and the exploration on the facilitation effect while employing ratio scores as a dependent variable instead of difference scores. That is, for instance, instead of calculating the Stroop interference by taking the difference of the mean of incongruent RTs and the mean of neutral RTs, we used the ratio of the two condition means (mean of neutral RTs / mean of incongruent RTs); and then took the ratio of the ratios at every step of the data processing instead of calculating the difference scores. To make sure that the distribution is close to normal, we ran a natural log (ln) transformation on the computed values for each analysis. All statistical analyses are one-sided t-tests in which we tested the conditions against zero (note that the expected effect size for a null effect for ratio scores is 1, however, we used ln transformation on the data and so equal condition means would result in a score of zero as $\ln(1) = 0$).

Bayes factor.

Analyses with difference scores. To determine the SD of the models of H1s, we applied the very same rule as we did for the analysis of the Stroop interference effect and

the effect of the suggestion on Stroop interference. That is, we extracted the mean facilitation effect from the meta-analysis of Parris, Dienes & Hodgson (2013) and used it as the SD of the half-normal distribution (with a mode of zero) modelling the prediction of H1 that assumes the presence of the facilitation effect. To specify the SD of the model testing the suggestion effect, we used the mean suggestion effect (mean facilitation effect in the no suggestion conditions - mean facilitation effect in the suggestion conditions) of the same meta-analysis. This latter test is exploratory, however, we set a directional hypothesis that is identical to the hypothesis regarding the reduction of the interference effect. Thus, we employed a half-normal distribution with a mode of zero as the model of H1, in which it is predicted that the facilitation effect will be smaller in the high incongruence proportion block than in the low incongruence one. The distribution representing the predictions of the model have an SD of 14ms, which is the half of the mean facilitation effects gathered in the meta-analysis of Parris, Dienes & Hodgson (2013).

To test the robustness of our conclusions based on Bs that employed a uniform distribution to model the predictions of H1, we conducted all of the corresponding analyses using a distribution that is in line with the scientific intuition that small effect sizes are more probable than large effect sizes. A half-Cauchy distribution with a mode of zero and a scale that equals to $1/7^{th}$ of the maximum effect size fits this assumption well (Dienes, 2017). Specifically, such a distribution predicts that it is equally likely that the real effect sizes is smaller than $1/7^{th}$ of the maximum than it is larger than $1/7^{th}$ of the maximum effect size. We report the Bs calculated with a half-Cauchy distribution that has a mode of zero and a scale of r as $B_{HC(0, r)}$.

Analyses with ratio scores. We applied half-normal distributions with a mode of zero to model the predictions of all H1s as we did for the difference score analyses. Moreover, we employed the same process to determine the parameters (SDs) of the H1 models for the analyses with ratio scores as the parameters of the corresponding analyses with difference scores. Using the raw condition means (see Supplementary Materials of Palfi, Parris, & Dienes, 2019) rather than the calculated interference and facilitation scores, we were computed ratio scores (e.g., mean of Neutral RTs divided by mean of Incongruent RTs) and directly apply them as the SDs of H1 models. For instance, to acquire the expected effect size of the Stroop effect, we calculated the mean of the congruent conditions and divided it by the mean of the incongruent conditions ((652+644)

/ (783+695) = 0.88). The computed expected effect sizes of the Stroop interference, Stroop facilitation, the effect of the word blindness suggestion on interference and facilitation are 0.91, 0.96, 0.92 and 0.98, respectively. For the test of the proportion congruency effects, we halved the expected effect sizes of the standard Stroop and interference effects. Note that halving a ratio based effect sizes imply that we took the arithmetic mean of the expected effect size and one. Therefore, the expected effect size of the proportion manipulation on the Stroop and interference effects are 0.94 and 0.96, respectively. Finally, to calculate the B, we ln transformed the derived expected effect sizes and used them as the SDs of the H1 half-normal models (e.g., the SD of the half-normal distribution modelling the Stroop effect is ln(0.88) = -1.28).

Supplementary Results

Analyses with difference scores.

Exploration S1: Is there an interaction between the suggestion, the type of the block and the extent of the Stroop facilitation effect? There was a Stroop facilitation effect ($t(22) = 6.24$, $p < .001$, $M_{diff} = 93$ ms, $d_z = 1.30$, $B_{H(0, 28)} = 1.79*10^3$, RR[4, $2.90*10^4$]) as the participants took more time to respond to neutral trials ($M = 807$ ms, $SD = 113$) than to congruent trials ($M = 714$ ms, $SD = 112$). Surprisingly, the difference between the suggestion and no suggestion conditions was larger in the low incongruence proportion block ($M = 50$ ms, $SD = 110$) than in the high incongruence proportion block ($M = -21$ ms, $SD = 154$). Indeed, the Bayes factor was closer to reach the level of good enough evidence for an interaction showing a greater effect of suggestion for the low rather than high incongruence block ($t(22) = -2.13$, $p = .045$, $M_{diff} = -71$ms, $d_z = -0.44$, $B_{H(0, 14)} = 1.91$, RR[13, 544]) than for an interaction predicting the effect the other way around $B_{H(0, 14)} = 0.58$, RR[0, 19]. The data were non-evidential regarding the suggestion effect in both of the low incongruence proportion block ($t(22) = 2.13$, $p = .042$, $M_{diff} = 50$ms, $d_z = 0.45$, $B_{H(0, 14)} = 2.53$, RR[10, 291]) and the high incongruence proportion block ($t(22) = -0.66$, $p = .515$, $M_{diff} = -21$ ms, $d_z = 0.14$, $B_{H(0, 14)} = 0.75$, RR[0, 32]).

Supporting test 1: Does the suggestion, the type of the block and their interaction influence subjects` experiences of word meaninglessness? The effect of suggestion on the reports about meaninglessness is robust over the two distributions modelling the predictions as all Bs show strong evidence for H1. However, the picture about the absence of the effect of the block and the interaction between suggestion and

block is less clear as the Bs indicate insensitive evidence for the items Q1, Q2 and Q3. Table S1 depicts all Bs with the half-Cauchy and uniform distributions.

Supporting test 2: Does the suggestion affect subjects' expectations? The evidence for the effect of the suggestion on the subjects` expectations to experience meaninglessness is strong irrespective of the chosen distribution ($B_{HC(0,\ 14)}$ = 84.60, RR[0.03, 278], $B_{U[0,\ 100]}$ = 7.71 *10^3).

Exploration 1: Is there any relationship between the expectations and the subjective experiences to the suggestion? The conclusion of all Bs based on half-Cauchy distributions are in line with the Bs using uniform distributions as all the difference scores show strong evidence for the positive relationship between expectations and subjective reports of meaninglessness. See Table S2 for the exact Bs.

Exploration 2: Is there any relationship between the objective responses and the expectations of the participants? The analysis revealed data insensitivity for all three cases with both of the alternatives. The evidence is insensitive about the relationship between the objective responses and the expectations in the no suggestion condition ($B_{HC(0,\ 3.54)}$ = 1.44, RR[0, 2.81*10^2], $B_{U[0,\ 24]}$ = 2.07), in the suggestion condition ($B_{HC(0,\ 3.54)}$ = 1.17, RR[0, 16.87], $B_{U[0,\ 24]}$ = 1.35) as well as for the difference between the conditions ($B_{HC(0,\ 3.54)}$ = 0.95, RR[0, 2.17*10^2], $B_{U[0,\ 24]}$ = 0.85).

Exploration 3: Is there any relationship between the subjective and objective responses of the participants? All Bs comparing the evidence for the existence and absence of the effect between the objective and subjective responses are insensitive with the half-Cauchy distribution. These results are in accordance with the ones based on the uniform distributions. See Table S3 for the exact Bs.

Exploration 4: Do subjects experience some depth of hypnosis during the post-hypnotic suggestion? The participants reported experiencing a hypnotic state or being relaxed during the hypnotic induction (*M* = 1.78, 95% CI [1.56 – 2.01]) and they reported being in a normal or relaxed state after the de-induction (*M* = 0.65, 95% CI [0.40 - 0.90]). The contrast of these two time points revealed strong evidence in favour of the difference between them regarding the experienced depth of hypnosis ($t(22) = 7.81$, $p < .001$, M_{diff} = 1.13, d_z = 1.63, $B_{U[0,\ 3]}$ = 5.20*10^5, RR[0.08, 3.20*10^3]). Although, the suggestion was *post*hypnotic, the participants claimed to be relaxed or even hypnotised in the suggestion

condition (M = 0.97, 95% CI [0.65 - 1.28]), whereas they felt that they were in a normal state while engaging in the Stroop task in the no suggestion condition (M = 0.51, 95% CI [0.24 - 0.78]). To further investigate this state that the participants experienced while the posthypnotic suggestion was active, we compared the suggestion condition to the no suggestion one, and to the induction and de-induction time points. There was a difference between the suggestion and no suggestion conditions in the reported depth of hypnosis ($t(22) = 2.64$, $p = .015$, $M_{diff} = 0.45$, $d_z = 0.55$, $B_{U[0, 3]} = 3.46$, RR[0.09, 343]). However, the contrast of the suggestion condition and de-induction did not provide sensitive evidence ($t(22) = 1.62$, $p = .118$, $M_{diff} = 0.31$, $d_z = 0.34$, $B_{U[0, 3]} = 0.57$, RR[0, 5.01]). Moreover, during the induction, the participants experienced deeper hypnosis than in the suggestion condition ($t(22) = 5.94$, $p < .001$, $M_{diff} = 0.81$, $d_z = 1.24$, $B_{U[0, 3]} = 7.00*10^3$, RR[0.08, $2.06*10^3$]).

All effects are robust to the shape of the distribution used to model the predictions of H1s. The participants reported greater hypnotic depth during the induction than right after the de-induction ($B_{HC(0, 0.43)} = 3.25*10^5$, RR[0.02, 333]). There was also a difference between the suggestion and no suggestion conditions ($B_{HC(0, 0.43)} = 8.20$, RR[0.02, 148]) and between the induction and the suggestion condition ($B_{HC(0, 0.43)} = 7.63*10^3$, RR[0.02, 226]). The evidence for the difference between the suggestion condition and de-induction is insensitive ($B_{HC(0, 0.43)} = 1.71$, RR[0, 3.37]).

Analyses with ratio scores.

Outcome neutral tests. We found strong evidence for the Stroop interference ($t(22) = 6.20$, $p < .001$, $M_{diff} = 0.12$, $d_z = 1.29$, $B_{H(0, 0.09)} = 2.04*10^4$, RR[0.01, 40.20])[32] and for the Stroop effects ($t(22) = 14.94$, $p < .001$, $M_{diff} = 0.24$, $d_z = 3.11$, $B_{H(0, 0.13)} = 4.92*10^{10}$, RR[0.01, 82.26]). The type of the block influenced the extent of the Stroop interference ($t(22) = 2.94$, $p = .007$, $M_{diff} = 0.07$, $d_z = 0.61$, $B_{H(0, 0.04)} = 14.59$, RR[0, 0.55]) and Stroop effects ($t(22) = 6.54$, $p < .001$, $M_{diff} = 0.15$, $d_z = 1.36$, $B_{H(0, 0.06)} = 1.14*10^4$, RR[0.01, 35.60]) in the predicted direction showing evidence for the proportion congruency effect. In other words, the extent of the interference and Stroop effects was smaller in the high incongruence than in the low incongruence proportion block.

[32] Of note, for all the computed values and predicted effect sizes, a negative M_{diff} indicates a difference in the non-predicted direction.

Crucial test. There was a three-way interaction between the type of the block, the suggestion and the interference effect ($t(22) = 2.15$, $p = .043$, $M_{diff} = 0.09$, $d_z = 0.45$, $B_{H(0, 0.08)} = 4.90$, RR[0.02, 0.24]). The suggestion decreased the interference in the high incongruence proportion block ($t(22) = 2.25$, $p = .035$, $M_{diff} = 0.06$, $d_z = 0.47$, $B_{H(0, 0.08)} = 5.19$, RR[0.02, 0.37]), whereas the suggestion did not influence the extent of the interference in the low incongruence proportion block ($t(22) = -0.99$, $p = .335$, $M_{diff} = -0.03$, $d_z = -0.21$, $B_{H(0, 0.08)} = 0.22$, RR[0.04, Inf]).

Exploration S1: Is there an interaction between the suggestion, the type of the block and the extent of the Stroop facilitation effect? There is evidence for the Stroop facilitation effect ($t(22) = 6.62$, $p < .001$, $M_{diff} = 0.12$, $d_z = 1.38$, $B_{H(0, 0.04)} = 5.27*10^3$, RR[0.01, 39.41]). The ratio score analysis also showed some evidence for the unexpected finding that the suggestion diminished the facilitation effect to a greater extent in the low incongruence than in the high incongruence block ($t(22) = -2.03$, $p = .055$, $M_{diff} = -0.08$, $d_z = -0.42$, $B_{H(0, 0.02)} = 2.01$, RR[0.02, 0.80]); and it supported the null over the model predicting the contrary ($B_{H(0, 0.02)} = 0.55$, RR[0, 0.03]). In fact, the suggestion modulated the extent of the facilitation effect in the low incongruence block ($t(22) = 2.21$, $p = .038$, $M_{diff} = 0.06$, $d_z = 0.46$, $B_{H(0, 0.02)} = 2.92$, RR[0.02, 0.47]) while the results for the high incongruence block were insensitive ($t(22) = -0.55$, $p = .588$, $M_{diff} = -0.02$, $d_z = 0.11$, $B_{H(0, 0.02)} = 0.73$, RR[0, 0.03]).

Discussion of Supplementary Results

The experience of being in a hypnotic state.

The convention of giving a suggestion posthypnotically is the standard way to distinguish the effects of hypnotic inductions and suggestions, and so control the potential confounds that can be introduced by the induction procedure (Terhune, Luke & Cohen Kadosh, 2017), such as the experience of being in an altered state of consciousness. Interestingly, we found evidence that highs reported experiencing a deeper hypnotic state in the suggestion than in the no suggestion condition. Indeed, hypnotic trance was earlier hypothesised to happen spontaneously during the implementation of a posthypnotic response (Erickson & Erickson, 1941). However, our participants reported being in a substantially deeper hypnotic state during the induction than during the Stroop task. Further, the evidence regarding the difference between the moment when the participants were asked to return to their normal state of wakefulness (after de-induction) and the

suggestion condition is insensitive. In sum, these data indicate that the mere application of posthypnotic suggestions cannot prevent participants from sensing any form of altered state of consciousness while responding to the suggestion opposing the traditional view that posthypnotic suggestions are not accompanied by any feeling of hypnotic depth (e.g., Terhune, Luke & Cohen Kadosh, 2017). Nonetheless, the observed hypnotic depth in the suggestion condition did not even exceed the level of feeling relaxed (M = 0.97, 95% CI [0.65 - 1.28]) rendering this change in the experience of being in an altered state most probably negligible.

Table S1.

Bayes Factors with Half-Cauchy and Uniform Distributions Modelling the Predictions of H1s about the Effect of Suggestion, Block and their Interaction on Subjective responses of meaninglessness

		Statistics		
Item	Predictor	$B_{HC(0, 14)}$	RR	$B_{U[0, 100]}$
Q1	Suggestion	$2.47 * 10^2$	$1, 8.63*10^3$	196
	Block	0.19	0, 29	0.04
	Suggestion*Block	3.26	0, 100	1.13
Q2	Suggestion	$1.86 * 10^2$	$1, 5.86*10^3$	120
	Block	0.36	0, 32	0.08
	Suggestion*Block	1.63	0, 53	0.47
Q3	Suggestion	8.71	$1, 5.64*10^3$	4.15
	Block	0.29	0, 45	0.06
	Suggestion*Block	0.65	0, 25	0.17
Q4	Suggestion	55.87	$1, 5.03*10^3$	33.66
	Block	0.11	6, Inf	0.02
	Suggestion*Block	0.33	12, Inf	0.08

Table S2.

Bayes Factors with Half-Cauchy and Uniform Distributions Modelling the Predictions of H1s Testing the Questions of Exploration 1

		Statistics of Exploration 1		
Item	Condition	$B_{HC(0, 3.57)}$	RR	$B_{U[0, 25]}$
Q1	No Suggestion	$1.00*10^5$	0.48, $9.16*10^3$	$2.99*10^5$
	Suggestion	$3.23*10^2$	0.64, $5.30*10^3$	$1.39*10^3$
	Difference	$1.50*10^4$	0.45, $6.47*10^3$	$9.26*10^4$
Q2	No Suggestion	1.63	2.18, 25.59	1.73
	Suggestion	$1.68*10^2$	0.61, $4.29*10^3$	$8.08*10^2$
	Difference	68.95	0.48, $3.81*10^3$	$2.32*10^2$
Q3	No Suggestion	11.13	0.88, $1.20*10^3$	36.14
	Suggestion	$2.54*10^2$	0.80, $3.96*10^2$	$1.15*10^3$
	Difference	$1.95*10^3$	0.49, $6.09*10^3$	$1.03*10^4$
Q4	No Suggestion	2.62	1.21, 98.23	3.97
	Suggestion	2.03	0, 42.48	2.54
	Difference	67.86	0.67,	271

Table S3.

Bayes Factors with Half-Cauchy and Uniform Distributions Modelling the Predictions of H1s Testing the Questions of Exploration 3

		Statistics of Exploration 3		
Item	Condition	$B_{HC(0, 0.14)}$	RR	$B_{U[0, 1]}$
Q1	No Suggestion	1.06	0, 1.83	1.11
	Suggestion	0.88	0, 0.39	0.65
	Difference	0.93	0, 0.96	0.77
Q2	No Suggestion	1.10	0, 2.36	1.19
	Suggestion	1.02	0, 1.13	0.91
	Difference	1.05	0, 2.20	1.04
Q3	No Suggestion	1.21	0, 7.51	1.45
	Suggestion	0.80	0, 0.57	0.52
	Difference	0.96	0, 2.21	0.80
Q4	No Suggestion	1.09	0, 2.23	1.17
	Suggestion	0.65	0, 0.71	0.33
	Difference	0.85	0, 1.03	0.63

Supplementary Materials of Chapter III

Experiment 1

Supplementary Methods

Bayesian parameter estimation with 95% Credibility intervals. To explore the extent to which hypnotisability and self-reports of strategy usage are related to the objective effectiveness of the strategies (reduction in Stroop interference), we employed parameter estimation rather than hypothesis testing. Hence, for these analyses, we report the strength of the association and its 95% Credibility Intervals (CI). For the first type of analyses, we report the Pearson`s r correlation coefficients, and for the second type of analyses, we computed the Kendall`s τ as the self-report scores followed a negatively skewed distribution. To calculate the 95% CIs of the Kendall`s τ, we applied the *credibleIntervalKendallTau* function of van Doorn, Ly, Marsman, and Wagenmakers (2016). Note that as these measures are standardized effect sizes, they do not indicate how much e.g. a given change in estimated percentage change in strategy use predicts a change in milliseconds of interference.

Supplementary Results

Supporting analysis of interest: What is the strength of the relationship between hypnotisability and the extent to which people reduce the Stroop interference via the strategies? As an additional analysis to the test of the regression slopes, we explored the strength of the association between the SWASH scores and the extent to which the participants reduced the Stroop interference by using the strategies. Virtually all correlation coefficients equal zero and the range of the plausible estimates lies within the boundaries of small to medium effect sizes. The results in descending order of the correlation coefficients are blurring (r = .03, 95% CI = [-.23, .29]), looking-away (r = .01, 95% CI = [-.25, .27]), single-letter (r = .01, 95% CI = [-.25, .27]), and goal-maintenance (r = -.01, 95% CI = [-.27, .25]) strategies.

Exploration: Estimation of the strength of the relationship between self-reports of strategy usage and the reduction in the Stroop interference effect. Finally, we explored the strength of the association between the participants` subjective reports on how successfully they applied the strategies and the objective efficiency of the strategies in reducing the extent of the Stroop interference effect. For the looking-away strategy, data are consistent with a null as well as a small effect size (τ = .16, 95% CI =

[-.02, .32]). We can be 95% confident that the correlations between self-reports and the blurring (τ = - .04, 95% CI = [-.21, .13]) and single-letter focus (τ = .03, 95% CI = [-.14, .20]) strategies are not stronger than .13 and .20, respectively. For the goal-maintenance strategy, the plausible values of the correlation do not exceed the level of a negligible association (τ = -.14, 95% CI = [-.28, .06]).

Experiment 2

Supplementary Results

Exploration: Estimation of the strength of the relationship between self-reports of strategy usage and the reduction in the Stroop interference effect. Finally, we explored the strength of the relationship between the experienced and objective efficiency of the strategies. The results are comparable for the looking-away (τ = .02, 95% CI = [-.20, .24]) and blurring strategies (τ = .03, 95% CI = [-.19, .25]). We can be 95% confident that the correlations lie within the region of weak effect sizes, and the data are consistent with either a negative or positive relationship for both of the strategies.

Supplementary Materials of Chapter IV

Pilot Experiment

Supplementary Results

Supplementary analyses of the Outcome neutral tests 2. We analysed how the suggestion and volition can influence the experience of control over the meaningfulness of the appearing words. The participants reported the highest level of control in the no suggestion condition (M = 2.33, SD = 0.84). They reported slightly lower level of control in the volition condition (M = 2, SD = 0.69) and the lowest level of control was reached in the suggestion condition (M = 1.1, SD = 0.71). The analyses revealed strong evidence for the difference between suggestion and no suggestion on the level of reported control ($t(29) = 6.50$, $p < .001$, $M_{diff} = 1.23$, $B_{H(0,\ 1.5)} = 1.29*10^5$, RR[0.08, $4.02*10^2$]) and insensitive evidence for the difference between volition and no suggestion volition ($t(29) = 2.41$, $p = .023$, $M_{diff} = 0.33$, $B_{H(0,\ 1.5)} = 2.76$, RR[1.37, 12.81]).

To estimate the extent to which the experienced meaninglessness felt voluntary, we also investigated the subjective nature of this phenomenon, by measuring whether people experience the suggestion as perception and the voluntarily produced meaninglessness as imagination. As the corresponding item of the questionnaire had only two levels (either imagination or perception), we can apply it as a dichotomous or a continuous variable. In the main text, we reported the analysis considering the variable as continuous, and here, we introduce the analysis in which the variable is taken as dichotomous. Based on the answers of 26 participants (4 participants did not provide an answer in one of the conditions), we estimated the odds ratio (OR) of the experienced nature of meaninglessness (either perceived or imagined) influenced by the type of the suggestion. The estimate (OR = 4, 95% CI[0.64-25.02]) is in the predicted direction, namely, indicating higher probability for experienced imagination over perception in the volition condition compared to suggestion condition, but the estimation covers a broad range of possible values.

Table S1

The 2x2 Contingency Table of the Experienced Nature of the Meaninglessness

Suggestion	Volition	
	Imagined	Perceived
Imagined	8	2
Perceived	8	8

Note. The table contains the data of 26 participants as 4 people did not provide an answer for either question.

Pre-registered Experiment

Supplementary Methods

Sample size estimation. To have an approximation of how many participants we might need in the registered experiment to find supporting evidence for the null (given that the null were true) in respect of the difference between the suggestion and volition conditions, we conducted a sample size estimation based on Dienes (2015, April 23). The core idea of this sample size estimation procedure is that we assume that the level of noise in measurement, indexed by the standard deviation of the crucial measure (difference in RTs between suggestion and volition conditions), will be identical across the two experiments. This allows us to calculate the B corresponding to any sample size with a chosen sample mean. In this analysis, we surmised that if the null were true then our sample mean would be 0 ms, and if there is a difference between the hypnotic and volitional suggestions then our sample mean would be identical to the one acquired in the pilot study (25 ms). We calculated the Bs with both raw effect sizes for sample sizes varying between 6 and 60 (Figure S1). Given that there is a difference of 25 ms between the two conditions in our sample we need to recruit around 24 people to provide moderate evidence for the alternative hypothesis. Whereas, to find moderate evidence favoring the null hypothesis, one might need to recruit 41 people with a sample mean of 0 ms.

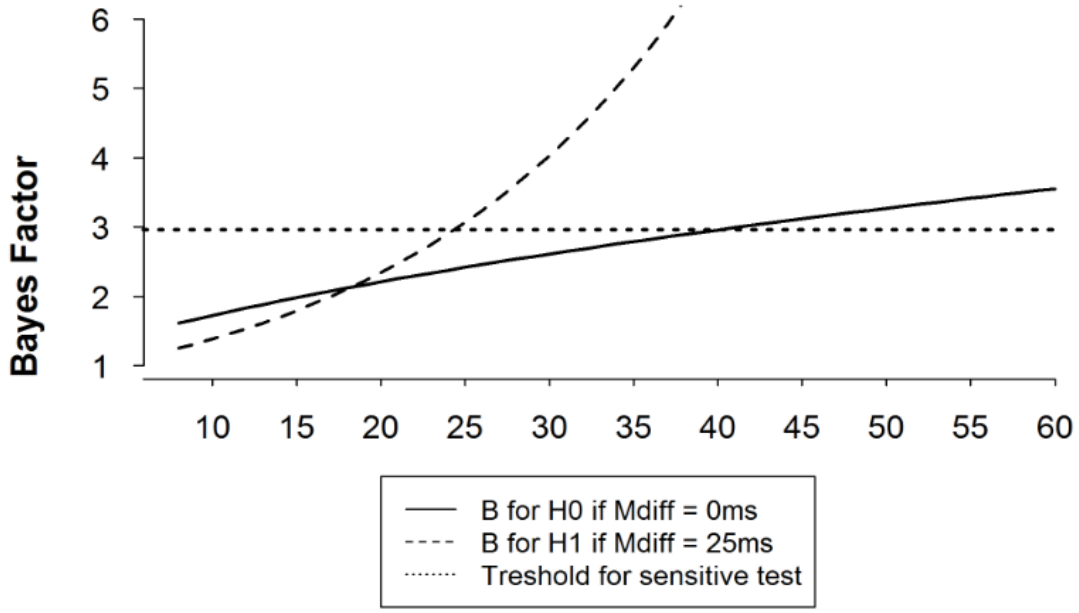

Figure S1. Sample size estimation. The bold line indicates Bs (H0 over H1) calculated with a sample mean of 0 ms and varying sample sizes, whereas the dashed line shows the Bs (H1 over H0) computed with a sample mean of 25 ms. The SD of the sample was taken from the pilot study; thus, the analysis assumes that the level of noise will be the same in a future study. The dotted line shows the threshold of a sensitive B based on Jeffreys (1961).

Supplementary Results

Analysis plan. Here, we will report the results of the analyses comparing the no suggestion and suggestion, and the no suggestion and volition conditions in respect of the experienced level of control. In addition, to demonstrate the variability in noise among the four participating labs, we will report a table with the Ms and SDs of the crucial analysis (difference between suggestion and volition conditions in terms of RTs) broken down by the place of data collection.

Supplementary analyses of the Outcome neutral tests 2. As expected, the reported level of control over how meaningful the words appeared to the participants was the highest in the no suggestion condition (M = 2.36, SD = 0.72) followed by the volition condition (M = 2.11, SD = 0.67) and it was the lowest in the suggestion condition (M = 1.50, SD = 0.81). The contrast of the conditions revealed that we need to suspend judgment regarding the effect of the volitional request on the level of control ($t(35)$ = 1.51, p = .141, M_{diff} = 0.25, $B_{H(0, 1)}$ = 0.92, RR[0, 2.8]). However, we found strong

evidence supporting for the effect of the suggestion in reducing the experienced level of control ($t(35) = 4.49$, $p < .001$, $M_{diff} = 0.86$, $B_{H(0, 1)} = 957$, RR[0.08, 240]).

Comparison of the participating labs in terms of the effect size of the crucial test. The crucial test of the current experiment was the comparison of the suggestion and the volition conditions in terms of the magnitude of the Stroop interference effect. Table S2 demonstrates the means and SDs of the conditions and of their differences separately for US and LaU. The raw effect sizes and SDs were comparable among the places of data collection apart the volition effect for which the level of noise appears to be substantially smaller at LaU than at US. Nonetheless, it is important to bear in mind that the results of LaU are based on only three participants.

Table S2

Summary Table about the Means of the RTs of the Stroop Interference effect in the Suggestion and Volition Conditions and for their Difference Broken Down by the Place of Data Collection

Place of Data Collection	Experimental condition		
	Suggestion	Volition	Difference
University of Sussex (US)	31 (86)	24 (100)	7 (72)
Lancaster University (LaU)	29 (89)	17 (38)	12 (81)

Note. The means are reported in ms and the Standard Deviations (SD) of the means are shown within the brackets.

Supplementary Materials of Chapter VI

Sample size estimation

Calculating the thresholds

Consider running a pilot study from which we take the standard error of the parameter tested and assume that will change only according to N in future studies. If we assume that H1 is true and we choose a raw effect size we expect to find, we can estimate the sample size we might need to find good enough evidence for H1 (e.g., $B > 3$) by simply increasing the sample size one-by-one and calculating the Bayes factor until it reaches the pre-set threshold of good enough evidence (Dienes, 2015, April 23). Note that elevating the sample size increases the evidence by reducing noise, the standard deviation of the sampling distribution (i.e., the standard error), while we assume that signal, the effect size measured in raw units, remains the same. This method has a 50% long-term relative frequency to find good enough evidence for H1 if the real effect size is identical to the effect size we used in the sample size estimation procedure. This statement follows from two principles. First, $B = 3$ roughly corresponds to $p = .05$, when the expected effect size is about the same as the obtained effect size (Jeffreys, 1961), indicating that there is a single critical t-value with which we can gain a B that indicates good enough evidence for H1 and a p-value that is just significant (provided that the level of alpha is set to .05). Second, if the critical t-value is identical to the noncentrality parameter of our t-distribution then the observed t-values will be above the critical value half of the time. Hence, the long-term relative frequency of the design to reject the null or in the case of the Bayes factor to find good enough evidence for H1 is 50%.

We can improve the long-term relative frequency of our design to find good enough evidence for H1 by further increasing the sample size and so reducing the standard error. This shifts the noncentrality parameter of our t-distribution and so increases the area under the curve that is beyond the critical t-value. To obtain a specific expected long-term relative frequency such as 80%, first, we need to identify the noncentrality parameter of the t-distribution from which we would obtain t-values larger than the critical t-value 80% of the time. This t-value is 2.8. Next, we simply need to find the corresponding B of the noncentrality parameter of that the distribution. We can do this by calculating the B as if the noncentrality parameter is the effect size and the expected effect size at the same time, while standard error equals 1. The result of this is $B = 20$. To calculate the threshold we need to estimate the sample size if we have a different cut-off of good enough evidence

than 3, we would need to adjust the critical t-value. We can do this by finding the t-value that corresponds to the B we plan to use as the cut-off of good enough evidence. For instance, t-value of 2.2 roughly corresponds to B of 6. To get the B threshold that we can use to estimate the sample size with which our design has a specific long-term relative frequency to find good enough evidence, we need to follow the steps described above and use the newly identified t-value as the critical t-value.

It is to be noted that the arguments laid here underlying the sample size estimation method are related to fixed designs (i.e., data collection until we reach the pre-set sample size) in which we want to assess the evidence for H1. It is not straightforward however, how this procedure works out when one uses optional stopping or when one is assessing evidence for H0. For instance, optional stopping allows us to stop collecting the data before we reach the sample size we gained from the estimation process, once we have reached the cut-off of good enough evidence for either H1 or H0. Due to this, however, in some cases, we may conclude wrongly that the null is true (based on the good enough evidence), when in fact H1 is true. This feature of optional stopping may or may not deteriorate the long-term relative frequency of our design to find good enough evidence for H1 when H1 is true. To evaluate the performance of the sample size estimation method for these cases, we conducted simulations covering all the combinations of long-term relative frequencies, cut-offs of good enough evidence and assumptions whether H1 or H0 is true.

Testing the thresholds via simulations

For each combination, we generated 1000 studies that had identical properties on average to the case study of Example 2. Specifically, the average of the raw effect sizes and the means of the experimental and control groups were equal to those of Example 1, respectively (when we assumed that H0 is true, the samples had identical means). This practice is in line with the assumption that the real effect sizes and their standard deviations are about the same as we acquired in the pilot study. The reader should be aware that they can have different assumptions and, for instance, replace the raw effect size gained from the pilot study with one that seems more plausible to approximate the real effect size, it is up to their scientific judgment. To calculate the Bayes factor, we used the same assumptions and parameters as we did in the case of Example 2. In half of the cases, we used fixed design, so we calculated the Bayes factor based on all data in each study. In the other half of the cases, we used optional stopping: we calculated the Bayes

factor first with 10 participants and then we added 10 participants and calculated the Bayes factor repeatedly until we gained good enough evidence either for H0 or H1. Finally, we derived the performance of a specific threshold by counting how many studies delivered evidence stronger than or equal to the cut-off of good enough evidence. The R script of the simulation can be accessed at: https://osf.io/5bzdf/.

Assuming that H1 is true. The results of the simulations, when we assumed that H1 was true, are represented in Table S1. Using fixed design and optional stopping produced similar long-term relative frequencies of finding good enough evidence, and their performance is comparable to what was expected. However, when the cut-off of good enough evidence is set at 3, optional stopping appears to be inferior to fixed designs as it did not reach higher long-term frequency than 62%.

Table S1

Simulated Long-Term Relative Frequencies of Finding Good Enough Evidence for H1 when H1 is Assumed to be True

Cut-off of good enough evidence	Threshold of B to estimate sample size	Theoretical probability (%)	Simulated probability with fixed design (%)	Simulated probability with optional stopping (%)
B > 3	3	50	0.521	0.571
	20	80	0.778	0.672
	70	90	0.864	0.676
	220	95	0.899	0.694
B > 6	6	50	0.495	0.58
	40	80	0.762	0.787
	150	90	0.864	0.863
	520	95	0.936	0.9
B > 10	10	50	0.472	0.559
	85	80	0.771	0.808
	350	90	0.866	0.898
	1370	95	0.922	0.937

Assuming that H0 is true. When we assumed that H0 is true, the simulated long-term relative frequency of finding good enough evidence for H0 was higher in virtually all of the cases than the expected probability of finding good enough evidence. This seems to be true for both of the fixed design and for the optional stopping.

Table S2

Simulated Long-Term Relative Frequencies of Finding Good Enough Evidence for H0 when H0 is Assumed to be True

Cut-off of good enough evidence	Threshold of B to estimate sample size	Theoretical probability (%)	Simulated probability with fixed design (%)	Simulated probability with optional stopping (%)
B < 1/3	3	50	0.519	0.473
	20	80	0.948	0.927
	70	90	0.981	0.928
	220	95	0.992	0.93
B < 1/6	6	50	0.511	0.589
	40	80	0.951	0.963
	150	90	0.979	0.96
	520	95	0.994	0.947
B < 1/10	10	50	0.482	0.612
	85	80	0.944	0.964
	350	90	0.991	0.975
	1370	95	0.992	0.971

Conclusions. The results of the simulation support that the presented sample size estimation method approximates well the long-term relative frequency of our design to find good enough evidence for H0, when H0 is true. This conclusion holds for the case when we assume that H1 is true as well with the restriction that if we use B > 3 as the cut-off of good enough evidence for H1 and we apply optional stopping then our design may not be able to achieve a better long-term relative frequency to find good enough evidence for H1 than 70%.

Appendix A. Items of the experiment

Expectations

How strongly do you expect to experience the words as at least somewhat meaningless?

1. I know the meaning of the words on the screen will be completely clear to me
2. I am guessing that the meaning of the words on the screen will in some way be unclear
3. I am fairly sure that the meaning of the words on the screen will in some way be unclear
4. I am almost certain that the meaning of the words on the screen will in some way be unclear
5. I am certain that the meaning of the words on the screen will in some way be unclear

Subjective experience questions

On what percentage of the trials:

1. Was the meaning of the words on the screen completely clear to you?
2. Were you aware of only an unclear meaning of the words on the screen?
3. Were you just aware of the colour and had no idea of what script the words were written in?
4. Were the words on the screen written in a clear yet completely meaningless script?

Depth of hypnosis scale

How deeply hypnotised were you during that game (Stroop task)?

1. Normal state
2. Relaxed
3. Hypnotized
4. Deeply hypnotized

Appendix B. R script to calculate the Bayes factor

The following R script integrates the scripts of the Dienes and McLatchie (2018) and the Dienes (2008) Bayes factor calculators into a single R function that can calculate *B*s with normal and t likelihood function; and it can model the predictions of H1 with either a uniform, a normal, a t or a Cauchy distribution. Moreover, the R script allows to calculate *B*s with one-tailed models that have a non-zero mode.

```
Bf<-function(sd, obtained, dfdata = 1, likelihood = c("normal", "t"), modeloftheory= c("normal","t","cauchy", "uniform") ,lower =0, upper=1, modeoftheory = 0, scaleoftheory = 1, dftheory = 1, tail = 2)
{
 if(likelihood=="normal"){
  dfdata=10^10
 }
 if(modeloftheory=="normal"){
  dftheory = 10^10
 } else if(modeloftheory=="cauchy"){
  dftheory = 1
 }
 area <- 0
 normarea <- 0
 if(modeloftheory=="uniform"){
  theta <- lower
  range <- upper - lower
  incr <- range / 2000
  for (A in -1000:1000){
   theta <- theta + incr
   dist_theta <- 1 / range
   height <- dist_theta * dt((obtained-theta)/sd, df=dfdata)
   area <- area + height * incr
  }
  LikelihoodTheory <- area
 }else{
  theta <- modeoftheory - 10 * scaleoftheory
```

```
  incr <- scaleoftheory/200
  for (A in -2000:2000){
    theta <- theta + incr
    dist_theta <- dt((theta-modeoftheory)/scaleoftheory, df=dftheory)
    if(identical(tail, 1)){
      if (theta <= modeoftheory){
        dist_theta <- 0
      } else {
        dist_theta <- dist_theta * 2
      }
    }
    height <- dist_theta * dt((obtained-theta)/sd, df = dfdata)
    area <- area + height * incr
    normarea <- normarea + dist_theta*incr
  }
  LikelihoodTheory <- area/normarea
 }
 Likelihoodnull <- dt(obtained/sd, df = dfdata)
 BayesFactor <- LikelihoodTheory/Likelihoodnull
 BayesFactor
}
```

Appendix C. Instructions in the experimental conditions of Experiment 1

No strategy

"This time do not use any of the strategies we have instructed you in previous blocks.

We would now like you to respond to the colour of the word on the screen as quickly and as accurately as you can."

Looking-away

"We would like you to focus on the top-right corner of the screen throughout the following experimental block and use only your peripheral vision to identify the colour of the words that appear on the screen.

You can practice this strategy now on an example word."

In this condition, the participants were told that they can focus on a spot that is closer to the word if they found the top-right corner to be too far away to easily identify the color of the word.

Blurring

"We would like you to blur your vision throughout the following experimental block by focusing on the screen as if you were looking into the distance.

You can practice this strategy now on an example word."

Single-letter focus

"We would like you to attend to a portion of the last coloured letter of each word in the next experimental block.

You can practice this strategy now on an example word."

Goal-maintenance

"We would like you to internally repeat the phrase "displayed colour" whenever you see the fixation cross.

Please repeat the phrase until the target appears on the screen."

Appendix D. Protocol of the Pilot Experiment

This is an example protocol in which the order of the condition was: (1) Suggestion, (2) No suggestion, (3) Volition. Note that the order of these conditions was counterbalanced across participants.

1. Instructions and consent form

Start script and provide the participant with the consent form

2. Practice

5 minutes of practice Stroop. Ends up with a screen asking to wait for the experimenter.

3. Induction, suggestion and test of suggestion

(O). Induction by Eye Closure.

(1). Now, please seat yourself comfortably and rest your hands in your lap. That's right. Rest your hands in your lap. Now look at your hands and find a spot on either hand and just focus on it. It doesn't matter what spot you choose; just select some spot to focus on. I will refer to the spot you have chosen as the target. That's right... hands relaxed... look directly at the target.

I am about to help you to relax, and meanwhile I will give you some instructions that will help you to gradually enter a state of hypnosis. Please look steadily at the target and while staring at it, keep listening to my words. You can become hypnotized if you are willing to do what I tell you to, and if you concentrate on the target and on what I say. You have already shown your willingness by coming here today, and so I am assuming that your presence here means that you want to experience all that you can. Just do your best to concentrate on the target -- pay close attention to my words, and let happen whatever you feel is going to take place. Just let yourself go. Pay close attention to what I tell you to think about; if your mind wanders, that will be okay; just bring your thoughts back to the target and my words, and you can easily experience more of what it's like to be hypnotized.

Hypnosis is perfectly normal and natural, and follows from the conditions of attention and suggestion we are using together. It is chiefly a matter of focusing sharply

on some particular thing. Sometimes you experience something very much like hypnosis when driving along a straight highway and you are oblivious to the landmarks along the road. The relaxation in hypnosis is very much like the first stages of falling asleep, but you will not really be asleep in the ordinary sense, because you will continue to hear my voice and will be able to direct your thoughts to the topics that I suggest. What is important here today is your willingness to go along with the ideas I suggest and to let happen whatever is about to happen. Nothing will be done to embarrass you.

(2) Now take it easy and just let yourself relax. Keep looking at the target as steadily as you can, thinking only of it and my words. If your eyes drift away, don't let that bother you... just focus again on the target. Pay attention to how the target changes, how the shadows play around it, how it is sometimes fuzzy, sometimes clear. Whatever you see is all right. Just let yourself experience whatever happens and keep staring at the target a little longer. After awhile, however, you will have stared long enough, and your eyes will feel very tired, and you will wish strongly that they were closed. Then they will close, as if by themselves. When this happens, just let it happen.

(3) As I continue to talk, you will find that you will become more and more drowsy. When the time comes that your eyes have closed, just let them remain closed.

You will find that you can relax completely, but at the same time sit up comfortably in your chair with little effort. You will be able to shift your position to make yourself comfortable as needed without it disturbing you. For now, just relax more and more. As you think of relaxing, your muscles will actually begin to relax. Starting with your right foot, relax the muscles of your right leg..... Now the muscles of your left leg..... Just relax all over. Relax your right hand... your forearm... upper arm... and shoulder.... That's right.... Now your left hand.... and forearm.... and upper arm.... and shoulder.... Relax your neck, and chest.... more and more relaxed.... completely relaxed.... completely relaxed.

(4) As you become relaxed, your body will feel deeply at ease.... comfortably heavy. You will begin to have this pleasant feeling of heaviness and comfort in your legs and feet.... in your hands and arms.... throughout your body.... as though you were settling deep into the chair. Your body feels comfortable and heavy.... Your eyelids feel heavy too, heavy and tired. You are beginning to feel very relaxed and comfortable. You are breathing freely and deeply, freely and deeply. You are becoming more and more

deeply and comfortably relaxed. Your eyelids are becoming heavier, more and more heavy and difficult to keep open.

(5) Staring at the target so long has made your eyes very tired. Your eyes may hurt from staring and your eyelids feel very heavy. Soon you will no longer be able to keep your eyes open. Soon you will have stood the discomfort long enough; your eyes are tired from staring, and your eyelids will feel too tired to remain open. Perhaps your eyes are becoming moist from the strain. You are becoming more and more relaxed and comfortable. The strain in your eyes is getting greater and greater. It would be a relief just to let your eyes close and to relax completely, relax completely. The strain in your eyes will eventually be so great that you will welcome your eyes closing of themselves, of themselves.

(6) Your eyes are tired and your eyelids feel very heavy. Your whole body feels heavy and relaxed. You feel a pleasant warm tingling throughout your body as you become more and more deeply relaxed ... deeper ... deeper ... more relaxed ... completely relaxed and drifting down into a warm pleasant state of relaxation. Keep your thoughts on what I am saying; listen to my voice. Your eyes are getting blurred from straining. You can hardly see the target, your eyes are so strained. The strain is getting greater, greater and greater, greater and greater. Your eyelids are heavy. Very heavy. Getting heavier and heavier, heavier and heavier. They are pushing down, down, down. Your eyelids seem weighted and heavy, pulled down by the weight .... so heavy ... your eyes are blinking, blinking .... closing, closing ...

Your eyes may have closed by now, and if they have not, they would soon close of themselves. But there is no need to strain them more. You have concentrated well on the target, and have become very relaxed. Now we have come to the time when you may just let your eyes close. That's it, eyes closed now.

(7) You now feel very relaxed, but you are going to become even more relaxed. It is easier to relax completely now that your eyes are closed. You will keep them closed until I tell you to open them or until I tell you to become alert ... You feel pleasantly, deeply relaxed and very comfortable as you continue to hear my voice. Just let your thoughts dwell on what I'm saying. You are going to become even more relaxed and comfortable. Soon you will be deeply hypnotized, but you will have no trouble hearing me. You will remain deeply hypnotized until I tell you to awaken later on. Soon I shall

begin to count from one to twenty. As I count, you will feel yourself going down further and further into a deeply relaxed, a deeply hypnotized state... but you will be able to do all sorts of things I ask you to do without waking up... One... you are going to become more deeply relaxed and hypnotized.... Two... down, down deeper, and deeper... Three... Four... more and more deeply hypnotized.... Five... Six... Seven... you are sinking deeper and deeper into hypnosis. Nothing will disturb you... Just let your thoughts focus on my voice and those things I tell you to think of. You are finding it easy just to listen to the things I tell you. Eight... Nine, Ten... halfway there... always deeper... Eleven... Twelve... Thirteen... Fourteen... Fifteen... although deeply hypnotized you can hear me clearly. You will always hear me distinctly no matter how deeply hypnotized you become. Sixteen... Seventeen... Eighteen... deeply hypnotized. Nothing will disturb you. You are going to experience many things that I will tell you to experience... Nineteen... Twenty. Deeply hypnotized now! You will not wake up until I tell you to. You will wish to remain relaxed and hypnotized and to have the experiences I describe to you.

Even though you are deeply relaxed and hypnotized, I want you to realize that you will be able to write, to move, and even to open your eyes if I ask you to do so, and still remain just as hypnotized and comfortable as you are now. It will not disturb you at all to open your eyes, move about, and write things. You will remain hypnotized until I tell you otherwise... All right, then....

Very soon you will be playing a computer game. When I clap my hands once, meaningless symbols will appear in the middle of the screen. They will feel like characters of a foreign language that you do not know, and you will not attempt to attribute any meaning to them. This gibberish will be printed in one of four ink colours: red, blue, green or yellow. Although you will only be able to attend to the symbols ink colour, you will look straight at the scrambled signs and crisply see all of them. Your job is to quickly and accurately depress the key that corresponds to the ink colour shown. You will find that you can play this game easily and effortlessly. When I clap my hands twice, you will regain your normal reading abilities.

[Clap to activate: “Now you see meaningless words on the screen]

[Show an example word and ask the participant to open her eyes and read out loud the following question with the answer options]

How strongly do you experience the word as meaningless?

1) The meaning of the word on the screen is completely clear to me
2) The meaning of the word on the screen is a little unclear
3) The meaning of the word on the screen is unclear
4) The meaning of the word on the screen is completely unclear

[If the participant has chosen 1 or 2 then read the following script otherwise jump it through]

“Notice how as you look at the word on the screen, you can look at it with the meaning fading to the background of your mind. We have found even when people consciously experience some meaning after this suggestion, they still process the words differently at a deeper level. You know you are capable of not reading meaningfully, remember how you have zoned out while reading a book.”

[Clap twice to deactivate: “Now you see meaningful words on the screen”]

[Ask the participant to close her eyes]

Stay completely relaxed and pay close attention to what I'm going to tell you next. In a moment I shall begin counting backwards from twenty to one. You will awaken gradually, but for most of the count you will remain in the pleasant, relaxed state that you are now in. By the time I reach "five" you will open your eyes, but you will not be fully aroused. When I get to "one", you will be fully alert, in your normal state of wakefulness. You probably will have the impression that you have slept, because you will have difficulty in remembering all the things I have told you and all the things you did or felt, since you started looking at the target. In fact, you will find it so much of an effort to recall any of these things that you will have no wish to do so. It will be much easier simply to forget everything until I tell you that you can remember. You will remember nothing of what you did or felt from the time that you started looking at the target, until I say to you: "Now you can remember everything!" You will not remember anything you did until then. After you open your eyes you will feel fine. I shall now count backwards from twenty, and at "five", not sooner, you will open your eyes but not be fully aroused until I say "one". At "one" you will be awake ... Ready, now: 20...19...18... 17... 16... 15... 14... 13... 12... 11... 10, halfway... 9... 8... 7... 6... five... 4... 3... 2... 1. Wake up! Wide awake! Any remaining drowsiness which you may feel will quickly pass.

From now you won’t feel hypnotised at all, but the suggestion will powerfully affect you when it is activated by the clap."

4. Suggestion condition

[Say the following]

[Clap to activate suggestion: "Now you see meaningless words on the screen"]

[Get the expectancy rating. Read out loud the question and provide the participant with the text format. Explain in detail if the participant has a question (same procedure for all of the other self-report measures)]

[Start the Stroop task]

[Ask about the subjective experience]

[Ask them to recall the meaning of the words]

[Ask about the depth of hypnosis]

[Ask about the level of control]

[Ask about how did they produce the effect of meaninglessness]

[Clap twice to deactivate suggestion: "Now you see meaningful words on the screen"]

5. No suggestion condition

[Say the following]

"For this part of the experiment no suggestion has been activated. It is important that you make no attempt to make the words seem like gibberish or word of foreign language. We would now like you to respond to the colour of the word on the screen as quickly and as accurately as you can"

[Get the expectancy rating]

[Start the Stroop task]

[Ask about the subjective experience]

[Ask them to recall the meaning of the words]

[Ask about the depth of hypnosis]

[Ask about the level of control]

6. Volition condition

[Say the following]

"Highly hypnotisable individuals such as you have been shown to be able to eliminate the interference from the irrelevant word when under the influence of the post-hypnotic suggestion and even when the suggestion is given without hypnosis. We would like you to voluntarily strongly and clearly imagine the irrelevant words as gibberish, words of a foreign language so that no meaning can be taken from them. This is not a hypnotic suggestion and we have not hypnotised you for this part of the task. You'll notice we have not initiated a suggestion by clapping or giving any other cue. You have the ability to do that anytime you please, under your control, as effectively as you just did. Please now voluntarily remove meaning from the words. You can do this so that it is under your control, just by exercising your imagination. You can be aware it is your imagination at the same time as it produces powerful effects."

[Have the participants look at a Stroop stimulus on the screen and ask them to make the word seem meaningless and then meaningful again. Tell them they can turn the control on and off.]

[Get the expectancy rating.]

[Start the Stroop task]

[Ask about the subjective experience]

[Ask them to recall the meaning of the words]

[Ask about the depth of hypnosis]

[Ask about the level of control]

[Ask about how did they produce the effect of meaninglessness]

[Finish]

7. Debrief and thank the participant

Appendix E. Items of the Pilot Experiment

Expectations

How strongly do you expect to experience the words as at least somewhat meaningless?

1. I know the meaning of the words on the screen will be completely clear to me
2. I have a little confidence that the meaning of the words on the screen will in some way be unclear
3. I am somewhat sure that the meaning of the words on the screen will in some way be unclear
4. I am fairly sure that the meaning of the words on the screen will in some way be unclear
5. I am almost certain that the meaning of the words on the screen will in some way be unclear
6. I am certain that the meaning of the words on the screen will in some way be unclear

Subjective experience questions

On what percentage of the trials:

Was the meaning of the words on the screen completely clear to you?
Were you aware of only an unclear meaning of the words on the screen?
Were you just aware of the colour and had no idea of what script the words were written in?
Were the words on the screen written in a clear yet completely meaningless script?

Recalling the meaning of the words

If you were aware of any words, can you recall them?

Depth of hypnosis scale

How deeply hypnotised were you during that game (Stroop task)?

1. Normal state
2. Relaxed
3. Hypnotized
4. Deeply hypnotized

Level of control

How much control did you have over how meaningful the words appeared to you?

1. I had no control
2. I had some control
3. I had almost complete control
4. I had complete control

Experienced nature of meaninglessness

How did you produce the effect of meaninglessness?

1. The script appearing meaningless was just me imagining it was meaningless
2. The script appearing meaningless was me perceiving the script as really meaningless

Appendix F. Instruction in the Volition condition of the Pre-registered Experiment

“Highly hypnotisable individuals such as you have been shown to be able to eliminate the interference from the irrelevant word when under the influence of the post-hypnotic suggestion and even when the suggestion is given without hypnosis. Notice that when a hypnotic suggestion is given it is always you who creates the response; thus you can achieve the full effect of a suggestion any time you wish. We would like you to voluntarily, strongly and clearly create the experience that the irrelevant words are gibberish, words of a foreign language so that no meaning can be taken from them. You have the ability to do that anytime you please, under your control, as effectively as you did it during the hypnotic induction. This is not a hypnotic suggestion and we have not hypnotised you for this part of the task. You'll notice we have not initiated a suggestion by clapping or giving any other cue. Please now voluntarily remove meaning from the words. You can do this so that it is under your control, just by exercising your every-day capacity to consider the world in different ways, while still knowing how the world really is. You can have complete control over the strategy you used hypnotically and use it without being hypnotised and produce the same powerful effects.”

Appendix G. New items of the Pre-registered Experiment

Expectations 2.

How strongly do you expect that naming the colour of the words will be somewhat easy?

1. I know that naming the colour of the words on the screen will be hard to me
2. I have a little confidence that naming the colour of the words on the screen will in some way be easy
3. I am somewhat sure that naming the colour of the words on the screen will in some way be easy
4. I am fairly sure that naming the colour of the words on the screen will in some way be easy
5. I am almost certain that naming of the words on the screen will in some way be easy
6. I am certain that naming of the words on the screen will in some way be easy

Depth of hypnosis

On a scale from 0 to 5, to what degree did you enter a hypnotic state during the game? 0 means your general state of consciousness was just the same as normal, 1 means you were slightly hypnotized and 5 means you entered very deep hypnosis?

Normal State 0 – 1 – 2 – 3 – 4 – 5 Deep hypnosis

Level of control

How much control did you have over how meaningful the words appeared to you?

1. I had no control because the words were written in a meaningless script
2. I had some control because the words were written in a meaningless script
3. I had almost complete control over whether the script appeared meaningless or meaningful
4. I had complete control over whether the script appeared meaningless or meaningful

Experienced nature of meaninglessness

How did it seem the effect of meaninglessness came about?

The script appearing meaningless was me perceiving it as meaningless	1 – 2 – 3 – 4	The script appearing meaningless was me imagining it as meaningless