Introspection of complex cognitive processes

Gabriel Reyes

Université Pierre et Marie Curie – Paris VI

École Doctorale Cerveau Cognition Comportement (ED3C)

Laboratoire de Sciences Cognitives et Psycholinguistique / Brain and consciousness

Introspection of Complex Cognitive Processes

Gabriel Reyes

DOCTORAL THESIS

Cognitive Science

Advisor: Jérôme SACKUR

Presented and defended publicly on the 29st of September 2015

Jury :

Dominique Muller, Professeur, Université de Grenoble	Rapporteur
Axel Cleeremans, Professeur, Université Libre de Bruxelles	Rapporteur
Jérôme Sackur, Directeur d'Études, EHESS	Directeur de thèse
Jean Lorenceau, Directeur de Recherche, CNRS	Examinateur
Claire Sergent, Maître de Conférence, Université Paris V	Examinatrice

Abstract

In the last decade, there has been a huge effort in cognitive science devoted to the understanding of how individuals access their own cognitive productions. However, in contrast with the extensive philosophical work in this area, empirical research about the introspective capacity is still in its infancy. This thesis investigates which mental contents are accessible by introspection, and what constraints apply to introspection. Specific minimal conditions for an adequate introspective report are theorized and investigated, as well as individual factors that might alter the accuracy of introspection. Four experimental projects were developed. The first project, consisting of four experiments, investigates the possibility to introspectively access and discriminate complex cognitive processes in the context of a visual search paradigm (serial searches *vs.* parallel searches). The second project, consisting of one experiment, refines the results of the first project. We used a pre-conscious visual cue to alter a visual search, and collected introspective data showing that participants were sensitive to this alteration. The results in both projects converge on three main ideas: 1) introspection, under certain and particular experimental conditions, is capable of accessing complex cognitive processes; 2) introspection is permeable to different sources of information underlying the experimental task; 3) the focus of introspection can be experimentally controlled during a simple cognitive task. The third project, composed of two experiments, extends the results evidenced in projects 1 and 2 to another domain: working memory. The study shows that introspection can successfully access the type of cognitive process engaged during memory recovery (serial access to information *vs.* parallel access). Lastly, the fourth project, composed of two experiments, investigates an individual factor that might alter the precision of introspective reports: biological reactivity to stress. Results indicated that individuals with high reactivity to stress have a poorer introspective access of their mental states. In summary, all experiments presented support the idea that introspection, presents functional properties that are experimentally approachable. Accordingly, the present thesis presents a first systematic account of the functional architecture of introspective report.

Table of Contents

1. Introduction

1.1. Historical context

The term introspection literally means "(lat. *spicere*) to look (lat. *intra*) within" (Lyons, 1986; Butler; 2013). In psychological literature, the term is used to refer to the ability in humans, as well as in some other animals (Smith, 2009; Smith, Couchman & Beran, 2012), to monitor their own mind (Flavell, 1979). Even though there is a certain consensus in psychology regarding the conceptualization of introspection today, this situation was completely different in the origin of the discipline, and this is partly explained by the tradition in the philosophy of introspection (Lyons, 1986). Restricting exclusively to the psychological domain, W. Wundt (1832-1920) was the first to maintain that the objective of the study of scientific pure psychology (as opposed to "social psychology") was immediate experience, and introspection, its method[1]. In later years, his followers set up a movement known as *Introspectionism* (Costall, 2006), which proposed building a scientific method focused on training individuals to carry out scientifically controlled introspection processes. In spite of the initial enthusiasm, such movement was not well received by psychologists and philosophers of that time, mainly because introspection as a method for the investigation of cognitive functioning was largely unsuccessful in the scientific context of the first half of the 20th century (Boring, 1953; Danziger, 1980). The result was that for a long period in the history of experimental psychology, introspection formally disappeared from experimental context.

Recently, a combination of the rise of cognitivism and a rejuvenated interest for the topic of consciousness has resulted in the return of the study of introspection. Introspection now is considered as a legitimate field in cognitive psychology (Jack & Roepstorff, 2002; Schooler, 2002a) and even amenable to experimentation in neuroscience (Fleming & Frith, 2014). The renewed interest of scientific psychology for introspection responds to a change in perspective: if in its origins introspection was proposed as a possible method of access to mental life, introspection is now conceptualized like any other cognitive process, i.e., an object of scientific research

[1] It is important to note the influence of Franz Brentano (1838-1917) in this respect. Brentano, in his book *Psychology from an Empirical Standpoint* (1984, cited in Lyons, 1986), maintains that the objective of psychology was the mental phenomenon and its method of study, inner-perception (Lyons, 1986). It is important to mention that my own work accepts this theoretical premise: introspection is indeed inner perception. At the same time, I do not consider in any details some distinctions that were crucial for Brentano and Wundt, such as the distinction between inner perception (*Wahrnehmung)* and inner observation (*Beobachtung*). Such distinctions are important for a correct understanding of the historical development of introspection (see Butler, 2013), but I did not operationalise them in my experimental work.

(Dunlosky & Metcalfe, 2009). Therefore, introspection is not so much a method for psychology as a human psychological faculty (amongst many others). The change of perspective justifies the interest for investigation of the functional conditions surrounding introspection. In the following sections this research introduces the theoretical and experimental breakthroughs which allowed for the development of the science of introspection.

1.1.1. Before Cognitivism

An arbitrary starting point for the history of experimental introspection is W. James (1842-1910). In the context of his proposal for a science of psychology, understood as science of mental life, James argues that the study of psychology must partake both to description and explication. On the one hand, psychology should be descriptive, as individuals present a privileged access to their mental states: "*the introspective observation is what we have to rely on first and foremost and always"* (James, 1890). In strict rigor, James states that, through introspection, individuals can describe their mental states: "*everyone agrees that we there discover states of consciousness"* (James, 1890). On the other hand, psychology should be explicative as it is interested in the correlation between the descriptive level (phenomenological level) and the research about brain states. In general, introspection is postulated as part of a method for the investigation of the mind. In particular, introspection is required as an active attentional mechanism over an internal mental content, which is needed for sound phenomenological descriptions.

James' position contrasts with the positions of Brentano (1838-1917) and Wundt (1832-1920) regarding the conceptualization of introspection as active internal perception. These authors argued that active introspective judgments concurrent to any mental act would be inadequate for a description of this very same mental act, given that the act of self-observation could alter, or even destroy the experienced content[2]. Brentano's research not only forced James to specify the mechanism through which introspection acts, but also preluded one of the two milestones explaining the discreditation of introspection in psychology.

The first milestone is known as the paradox of Comte (1798-1857). The paradox states that it is impossible to disassociate the subject and the object of study in an introspective judgment. In essence, the observer and the

[2] In contrast, Wundt proposes a *passive inner perception* account for introspection. This consists on rigorously train individuals to make introspective reports about what takes place passively in their own minds as of passively perceived mental content (Lyons, 1986).

observed are one and the same entity, a situation which goes against the demand of all scientific tools in respect to the subject-object distinction[3]. Comte concludes that introspection as a method is, as matter of principle, faulty beyond repair. The solution to this paradox was given by John Stuart Mill (1806-1873): introspective access is given in memory, a moment after the mental act has been produced. James (1890) accepts this distinction and argues that introspection, understood now as retrospection, acts over mental contents suspended in memory, outside the current of mental events. The previous hypotheses raise a series of questions, for example: What type of memory does introspection access to recover mental data? Does the storage time of a mental content in memory have any impact on the quality of introspection? What differentiates the introspective mechanism from the simple act of remembering a mental act? Although the relation between working memory and introspection has not been sufficiently investigated (Bona, Cattaneo, Vecchi, Soto & Silvanto, 2013; Bona & Silvanto, 2014), most of the contemporary studies implicitly assume this distinction (Kosslyn & Thompson, 2000; Schooler, 2002a). Lastly, an important consequence of the James-Mill vision, which influences contemporary research, is that introspection would act as an internal attentional focus, i.e., a second-order conscious process to recover a *representation* of a mental content that has taken place in the past.

The second milestone refers to the second objection of Comte (1798-1857) about the low replicability of introspective reports: the introspective method does not always generate the same results under the same experimental conditions. In fact, many studies with introspective reports not only presented a high inter-individual variability, but also when being re-evaluated in the same individuals or similar experimental conditions, a poor within-individual stability was observed (Overgaard, 2006). In response to such objections, several researchers of that period, for instance Titchener (1867-1927) and Külpe (1892-1915), attempted to train participants to make reliable introspective processes in rigorously controlled experimental contexts, however, without much success. The second objection of Comte was responsible for discrediting introspection as a method of investigation in psychology, which was even further expanded with the emergence of behaviorism (Watson, 1878-1958). However, it can be argued that the experimental development exhibited only an *apparent* disappearance. Overgaard (2008) for instance argues that in many experiments in cognitive sciences, it is possible to implicitly distinguish the interest for introspective reports. The pioneering work of (Sperling, 1960) is a prototype in this respect as it can be seen both as a work in information processing cognitive science and as a masterpiece of introspection. In this line, it can be argued that introspection as a method of study continued to

[3] In the words of T. Nelson (1996, p. 104): "*How can one and the same organ be the organ doing the observing and the organ being observed*".

have an influence during the both the 20th Century and the beginning of the 21st Century, despite the fact that the term as such, nearly completely vanished (Costall, 2006; Sackur, 2009).

In perspective, Dunlosky and Metcalfe (2009) state that the fundamental problem with introspection at the beginning of the 20th century was that researchers unsuccessfully attempted to use introspection as a tool to describe how the mind works; however, they were not interested in investigating introspection as a psychological phenomenon *per se*. The difference between "*research method*" and "*object of study*" is the change in focus that characterizes the investigation of introspection in the following decades.

1.1.2. Cognitivism

The return of introspection on the experimental scene can be traced back to the emergence of metacognition as a theme in the cognitive psychology of memory and in developmental psychology. "Introspection" and "metacognition" can be seen as synonyms, in a first and crude sense (Overgaard & Sandberg, 2012); one could wonder whether the latter term was preferred simply because it did not evoke unwanted memory of psychology's pre-behaviorist past. The renewed interest in introspection or metacognition is parallel to the appearance of cognitivism[4], i.e., the study of the mind as mechanism of information processing (Mandler, 2007). Of particular interest is the proposal of Atkinson and Shiffrin (1968) in this period. The authors assert that information stored in short term memory, previously recovered from sensory memory would be available for diverse processes of cognitive control. This idea is not only the prelude to the notion of "*control*" and "*monitoring*" in metacognition studies (Nelson and Narens, 1990), but also updates the conceptualization of introspection in as retrospection of a mental content suspended in the memory (James, 1890)[5]. Indeed, Atkinson and Shiffrin's model supposes that, by default, information about past processes is available in working memory, and that it could be read out by monitoring processes.

[4] The literature describes two reasons why conductivism loses popularity facing the study of mental processes. The first critical factor is the incapacity of fully covering the study of the animal behavior, beyond a simple stimulus-response explanation (Dunlosky & Metcalfe, 2009). The second factor refers to the contrast with the cognitivist interpretation of the behaviour, based on emerging computational models of the mental processing (Hunt & Ellis, 2004). In fact, this historic period (1950-1960) is characterized by the appearance of seminal cognitivist models of the mental architecture (e.g., Broadbent, 1958; Neisser, 1967).

[5] In subsequent years this idea would constitute the first condition of validity in introspection: an introspective act will be valid only if the requested mental information still resides in the short term memory (Ericsson & Simon, 1980).

During this period the first explicit attempts to renew introspective investigations appeared: the work of Joseph T. Hart (1965) is seminal. The author introduced in experimental psychology the concept of *"tip-of-the-tongue"* (TOT), an introspective phenomenon which occurs when one has the sensation of having the response to a question without being able to access it, or retrieve it from working memory (Brown, 1991). Hart in several studies (1965, 1966 and 1967) provided evidence that individuals who showed signs of TOT systematically recognized the correct answers in a 2AFC (two-alternative forced choice) task, in comparison to the group which did not have signs of TOT. The fundamental question uncovered by the research is: *how can individuals know something about their memory that in fact they cannot recall?* In order to provide a response; the concept of introspection could be suggested. In the following years (1970-1990), several introspective phenomena were introduced: Judgments of Learning (JOLs), Feeling of Knowing (FOKs). Confidence in decision making was re-discovered (Dunlosky & Metcalfe, 2009; Dunlosky, Serra & Baker, 2007), after it had been used in influential studies at the very beginning of experimental psychology (Pierce & Jastrow, 1884). Introspection was not only introduced in the guise of *monitoring*, but also in the guise of *control*: for instance, in Kroll, Kellicutt & Parks (1975), participants were asked to actively control their rehearsals in verbal memory, a form of control which is purely introspective. Motivated by these studies, the 70's were characterized by the debate regarding the validity of an introspective report (Ericsson & Simon, 1980; Lieberman, 1979; Nisbett & Wilson, 1977). It is important to note that this was the historic moment where introspection began to be considered not principally as a *method* for studying mental processing. Instead of that, introspection begins to be conceptualized as a psychological *object* of research. The main inquiry was: under which subjective and experimental conditions is it possible to trust an introspective report?

Formally, the metacognitive school rose up in the 1980's with the publication of the influential study *"Metacognition and Cognitive monitoring"* (Flavell, 1979)[6]. In a previous paper, the author defines metacognition as "*the knowledge of each individual regarding its own cognitive processes and products* […] *Metacognition refers to active control and subsequent regulation of the objects of knowledge* (mental contents)" (Flavell, 1976, p. 232). From this definition two key concepts can be extracted: i) it is possible to distinguish *cognitive processes* and their *products* when we talk about metacognitive access (to be discussed subsequently) and ii) metacognition is not only a *knowledge process*, but also plays a role in the *cognitive control* of mental

[6] It is important to highlight that the term "Introspection" is changed in favor of the "Meta (*μετά*: after or beyond) – Cognition", defined as "knowledge and cognition about cognitive phenomena". In the literature there are no conclusive reasons to abandon the term "Introspection". It is possible to speculate that the reason is the still present fear of being associated with the introspectionism of the Würzburg School, criticized earlier for its unreliability.

states. Flavell (1979) emphasizes the importance of two central concepts in the scientific investigation of metacognition: individuals can *monitor* their own cognitive activity, which leads to the possibility of *controlling* this activity. The dialectic between "control" and "monitoring" generates a series of metacognitive experiences which can be studied experimentally (Flavell, 1979). Starting from this conceptualization, the study of metacognition was diversified to almost all areas of psychology: social cognition (Bang et al., 2014; Frith, 2012), education (Zohar & Barzilai, 2013), child development (Lyons & Zelazo, 2011), mental disorders (David, Bedford, Wiffen & Gilleen, 2014), among many other domains, each one with a particular development (Wilson, 2003). Finally, in the 90's Thomas O. Nelson and Louis Narens (1990, 1994) presented a framework about the *metacognition-cognition* distinction (Figure 1), which generated the dominant interpretation in the discipline. The authors proposed a model with two levels: *object*-level and *meta*-level. The first order level, or object-level, refers to the mental processes involved in a cognitive task, such as memory, attention, learning, language, etc. The second order level, or meta-level, refers to a model or representation of the individual with respect to the cognitive process implied in that task. It is proposed that the meta-level monitors the object-level, and may possibly modify it.

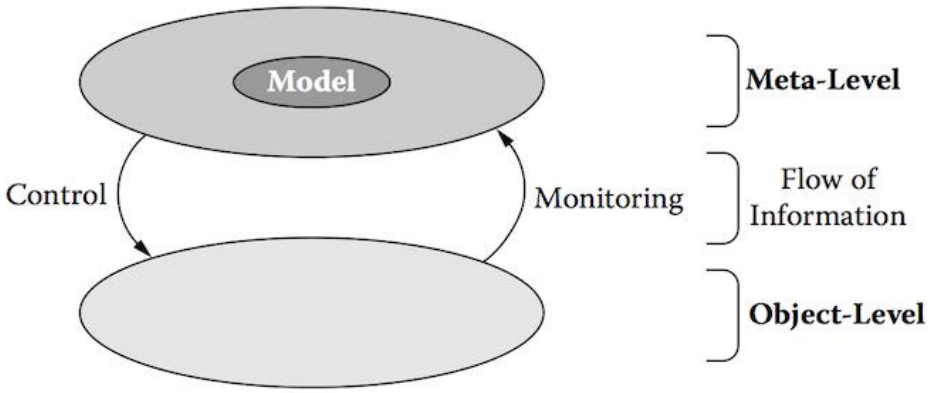


Figure 1. Nelson and Narens's model (1990).

1.1.3. Neuro-Cognitivism

The last two decades of research have considered introspection as a new cognitive process that can be scientifically researched[7], and not as an alternative method of access to mental processes. Recently, interest has focused upon the development of introspection. Recent cohort studies suggest that metacognitive capacity would increase during adolescence and stabilize during adulthood (Palmer, David & Fleming, 2014; Weil et al., 2013). In old age, a deterioration of metacognitive capacity would be observed (Palmer et al., 2014), which goes above and beyond the normal deterioration of executive functions during this period. Although future longitudinal studies that evaluate this matter within participants are required, it all seems to indicate a curved development of metacognitive capacity. The previous point supports the hypothesis of introspection being an independent cognitive process. At the same time, if introspection is an independent mechanism, as James (1890) insisted, it should also be possible to observe specific neural states indicative of its functioning (Fleming & Dolan, 2012). In this line, recent studies associate both medial and lateral regions of the anterior PFC (Baird, Smallwood, Gorgolewski & Margulies, 2013; de Martino, Fleming, Garrett & Dolan, 2013; Fleming, Huijgen & Dolan, 2012; Fleming, Weil, Nagy, Dolan & Rees, 2010; Rounis, Maniscalco, Rothwell, Passingham & Lau, 2010; Yokoyama et al., 2010), as well as with the dorsal premotor cortex (Fleming et al., 2015), with the precision of metacognition. In addition, neuro-imaging studies have identified a high inter-individual variability related to specific structural markers (Baird et al., 2013; Fleming et al., 2010). Interestingly, within the same cognitive domain, such metacognitive sensitivity is independent of the cognitive complexity of the experimental context (Meuwese, van Loon, Lamme & Fahrenfort, 2014; Song et al., 2011). In summary, these breakthroughs in the neuroscience of metacognition not only point to the relative independence of introspection from cognition, but also highlight the presence of intra-individual factors, that seem out of the reach of experimental control. Finally, notice a set of two highly debated and open questions in the current field: First, the unity of introspection has been questioned by recent studies that evidenced differences in the introspective accuracy in perceptual tasks *vs*. memory tasks (Baird et al., 2013; Baird, Mrazek, Phillips & Schooler, 2014; Fleming, Ryu, Golfinos & Blackmon, 2014). Second, the reliability of introspective measures has been questioned, (Overgaard & Sandberg, 2012; Sandberg, Timmermans, Overgaard & Cleeremans, 2010), as well as their physiological

[7] Methodologically speaking, contemporary research has been interested in three types of introspective reports: (i) measures related to meta-memory (Dunlosky & Bjork, 2008); (ii) the introspective evaluation of the response time in a cognitive task (Bryce & Bratzke, 2014; Corallo, Sackur, Dehaene & Sigman, 2008; Marti, Sackur, Sigman & Dehaene, 2010); and (iii) the judgment of confidence in decision making (Fleming & Frith, 2014). These reports have presented different strategies of analysis, the latter field being the one that has been the focus of attention recently (Fleming & Lau, 2014; Maniscalco & Lau, 2012).

markers (e.g., metacognitive accuracy would be negatively related to pupil dilatation (Lempert, Chen & Fleming, 2015).

The new topics of research define introspection as a cognitive process with its own functional properties. When conceptualizing introspection this way, the objection about the lack of generalization or reproducibility of introspective data, no longer makes sense (Piccinini, 2003). As a consequence, the new challenge for experimental introspection consist in establishing which mental contents (objectively described) correlate with which introspective reports (subjectively described) and under which psychological and experimental conditions. Along these lines, the motivation of this thesis is to re-evaluate the debate about the extension of introspective reports: *which mental contents are accessible by introspection?* In the following section this scientific question is discussed further.

1.2. Introspection of high-order cognitive processes

Contemporary literature in philosophy of the mind (Smithies & Stoljar, 2012) and experimental psychology (Jack & Roepstorff, 2003, 2004) converge on the idea that individuals can access, to some extent, their own mental states. The dominant theory (Butler, 2013) maintains that such access is characterized by a type of inner perception. In other words, individuals would access their mental contents through an inner self-perception process in a manner analogous to their access to external objects. It is important to indicate that this is not the only conception of introspective access found in literature (e.g., phenomenological account[8]), however, this account has received the most acceptance in experimental contexts (Jack & Shallice, 2001). Contemporary research has assumed the task of revealing the functional mechanisms of inner perception.

Expanding on the research into inner perception, Armstrong (1968) argues that this type of inner-observation is a form of self-scanning process, a mechanism of monitoring mental states. Supporting research (Armstrong, 1968; Churchland, 1984; Lycan, 1987) suggest that, like any other cognitive mechanism, self-observation is not infallible: as visual perception can differ from what happens in reality, the introspective scan may also fail on internal mental contents. The previous point responds to the fact that access is not direct and transparent, but representational. Along these lines, Lycan (1987) suggested that the process of introspective monitoring would have as an end-product certain second-order representations regarding the mental content of interest (or first-order information). This idea is central in the history of introspection, since it places the validity of the access mechanism in the process of re-representation (Jack & Roepstorff, 2002) or translation (Schooler, 2002a, 2002b) of a mental content. The question of the extension of introspective reports has been debated in the contemporary philosophy of introspection (Carruthers, 2010; Schwitzgebel, 2011), a topic that seems forgotten by experimental introspection (Overgaard & Sandberg, 2012). Two main open questions are: Firstly, are all mental contents equally accessible or (second order) representable? Secondly, what are the conditions of introspective representations?

The concern for this matter was inaugurated by cognitive psychologists in the 1960s, who – with insufficient experimental evidence – suspected that individuals did not posses direct access to higher order mental processes

[8] Butler (2013, p. 62) asserts that "*we know the phenomenal character of what our conscious states are like by actually being in them* […] *because* [it] *consists in the actual embodied experience of the state in the world, in virtue of one's being the existential subject of experience that undergoes the conscious state*. *The knowledge is constituted by the experience of the state itself* ".

(Mandler, 1975; Miller, 1962; Neisser, 1967). Thus they deduced that purported access were in fact confabulations. In the words of these authors (quoted by Nisbett & Wilson, 1977, p.232): "[…] *people's ability to observe directly the working of their own mind* […] *is the result of thinking, not the process of thinking, that appears spontaneously in consciousness*" (Miller, 1962, p.56); "*The constructive processes of encoding perceptual sensations themselves never appear in consciousness, their products do*" (Neisser, 1967, p.301). From a general perspective, the literature of that time maintains that introspective access would be restricted to the cognitive *states* (e.g., perceptual mental state) accompanying or preceding a decision. On the contrary, information related to the cognitive *processes* which precede and inform such decisions (i.e., sensory computations at the basis of the decision), would be inaccessible to participants. In the following section the empirical evidence for this thesis is discussed further with regard to the contemporary literature.

In the philosophy of introspection, the dominant vision maintains that the veracity of introspective data depends on the nature of the mental content under scrutiny (Carruthers, 2010, 2011). In effect, the distinction would be the product of an epistemic difference (*in kind*) in the access process. Whilst access to complex cognitive processes (access to propositional states) would require an interpretative self-attribution process, in contrast, access to perceptual cognitive states would be mainly inferential in the sense that it would proceed through a representation (e.g., Rosenthal, 2005), i.e., without an interpretative mechanism. In accordance, it is convenient to distinguish between introspective models (Goldman, 2006; Hill, 2009; Prinz, 2007) and non-introspective models (Bem, 1972; Nisbett & Wilson, 1977; Wilson, 2002) of access to inner mental activity (cognitive states and processes). It is important to point out that the distinction between these two classes of models is experimentally justified as long as there is a correct correlation between subjective report and objective indices of the mental content of interest (Schwitzgebel, 2012). As an example, it is accepted that individuals are able to introspectively access their response time in a cognitive task, because there is a positive correlation between the subjectively estimated response time and the objective response time in the same task (e.g., Corallo et al., 2008). In opposition to this theory, non-introspective models have systematically denied the existence of such relations, mainly because of the complex nature of the interpretative mechanisms of access.

Some researchers (Schwitzgebel, 2011) have recently questioned the pertinence of the above distinction. Schwitzgebel defends the idea that during an introspective judgment, individuals make use of an indiscriminate

array of multiple monitoring mechanisms over different sources of information (decisions and perceptions)[9]. This inaugurates the hypothesis that the validity of access to a mental state might directly relate to how introspective mechanisms select the sources of information (Prinz, 2004, 2007, Goldman, 2004). This hypothesis has not currently been experimentally investigated. It supposes that the difference between introspective and non-introspective access would be related more to a difference in the *degree of access* than to a difference in the *kind of access.* In other words the lack of introspective process could be related to a deficient discrimination of the mental content of interest from other irrelevant sources of information. This suggests a certain *gradualism in introspective access:* the higher the complexity of the mental content of interest, the higher the probability of an interpretative mechanism. At the same time it is also possible that introspective mechanism should be more or less *specific*: the more detailed the introspective demands, the lower the probability of confabulatory interpretation of internal data. According to these ideas, the present thesis investigates the functional conditions that determine the validity of introspective data. The specificity of introspection is manipulated, and in parallel, the diversity of introspective information that is available to participants within one particular task is estimated. The next section revises the empirical evidence for introspective inaccessibility of complex cognitive processes.

1.2.1. The Nisbett and Wilson Canon (NWC)

The renewed interest for introspection in the early days of cognitivism was accompanied by the question about its reliability. Nisbett and Wilson (1977) famously suggested that it is only possible to trust those mental contents available to the participants' consciousness. The authors proposed the distinction between accessible mental contents (cognitive *states*) and inaccessible ones (cognitive *processes*). The research comes from a systematic review of experiments in social psychology (research areas included: attribution, cognitive dissonance, subliminal perception, problem solving, etc.), from which Nisbett and Wilson concluded that individuals might be: (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response. These points form what we shall call Nisbett and Wilson's canon (henceforth, NWC): introspective access is restricted to mental products or perceptual contents; the underlying sensory transformations are inaccessible to the

[9] The pessimist version to this thesis (Schwitzgebel, 2011, 2012) suggests that such multiplicity of sources of information would be the cause of inter and intra individual variability, traditionally evidenced in a report of this type. Optimistic versions of this thesis (Bayne, 2015), converge on the idea that not all sources of information are equally weighted in the processing of introspective data.

individuals' consciousness[10]. In summary, the NWC maintains that cognitive processes would not be accessible to introspection; when participants try to introspect on processes, they make use of *a priori* theories about the causal relationship between stimulus and response.

In order to illustrate the above, Nisbett and Wilson present an experiment where participants are shown four pairs of socks and are asked to choose the best pair. Participants were not informed that all four pairs were identical. After the participants' decision, they were requested to justify their choice. The experiment presented a robust linear effect of position: as the position of the pair of socks advanced from left to right, the preference increased. The experiment does not investigate or speculate about the origin of this effect. The point of interest is that, when asking them for the cause that motivated their preference, none of the participants mentioned position as a determining factor. Even after asking them directly if position could have affected their preference, all participants denied this possibility. On the contrary, all participants confabulated regarding the motives of their choice.

It is important here to distinguish the *cognitive process* from the *cognitive state*. Individuals successfully access their sensory experience regarding the apparently "better quality" of the selected pair of socks; however, they do not access the cognitive process that motivated this decision. According to Schwitzgebel (2011): "*They* (Nisbett and Wilson) *are skeptical about our knowledge of why we selected a particular brand of socks, not about the fact that we do judge them to be superior or about our sensory experience as we select them*". According to this, Nisbett & Wilson do not reject all kinds of introspective access; the authors accept that it is possible to introspectively describe mental contents such as perceptual experiences; however, there would be no introspective knowledge of the causal process underlying and driving our judgments, decisions, emotions and sensations. The difficulty in this (and other) experiment is that it is nearly impossible to tell apart the veridical introspection of hallucinated properties of the stimulus caused by a mechanism the participants are not aware of from a pure confabulation, that is: the invention of a justification for the choice that is rationally and socially acceptable in the absence of any underlying cognitive difference. Our studies will try to offer solutions to this issue.

---

10 The common definition of a cognitive state refers to a mental content that accompanies or precedes the decision (e.g., the level of confidence in a decision or the perception of the response time on a cognitive task). On the other hand, the cognitive processes refer to the cognitive treatments or sensorial transformation that both precedes and determines the decision (Rich, 1979). Another way to introduce the concepts of cognitive *"states"* and *"processes"* is proposed by White (1980) through the distinction between "*knowing that*" and "*knowing how*", respectively.

The contemporary formulation of the NWC (Wilson, 2002, 2003; Wilson & Dunn, 2004) maintains that introspection would have retrospective access only to consciously processed mental contents (cognitive *states*) and not to those contents with *unconscious* processing (cognitive *processes*). This would constitute a central characteristic of our cognitive architecture. Relevant information about mental states is maintained in consciousness and superfluous information is kept underneath: this is precisely the notion of the *adaptive unconscious*. Moreover, recent studies in social cognition (Johansson, Hall, Silkström & Olsson, 2005; Johansson, Hall, Silkström, Tärning & Lind, 2006) have supported the NWC through a phenomenon called *choice blindness* (see also Froese, 2013; Jack, 2013; Petitmengin, Remillieux, Cahour & Carter-Thomas, 2013). A famous experiment on this phenomenon consists in presenting to participants two pairs of cards with two faces. Immediately afterwards, participants were asked to determine which one they preferred. In certain trials, researchers asked participants to report the cause that motivated their preference. The critical point of the experiment is that in some trials and immediately after the individuals stated their preference, researchers, by means of legerdemain, presented to participants the card they did *not* select, without their noticing it (Figure 2). Surprisingly, participants did not detect that the researcher had handed them a different card and furthermore they continued to articulate justifications for their decisions that are related to the new card handed to them by the researcher. Participants are blind to the causes that guided their original choice. The results of this study support the NWC model, as it provides evidence for a lack of conscious access to the causes that motivate behaviour. Importantly, recent studies (Hall et al., 2013; Johansson, Hall, Tärning, Sikström & Chater, 2013; McLaughlin & Somerville, 2013) confirm that this phenomenon is independent from the experimental context. An important aspect to note is that all studies commented upon by Nisbett and Wilson, in addition to the more recent evidence in favor of the NWC, comes from social psychology experiments. There is no evidence from basic cognitive experimental psychology that shows blindness to cognitive processes thus far. All evidences seem to comprise of crucial psychosocial contexts or researcher-participant relationship as driving factors (Petitmengin et al., 2013).

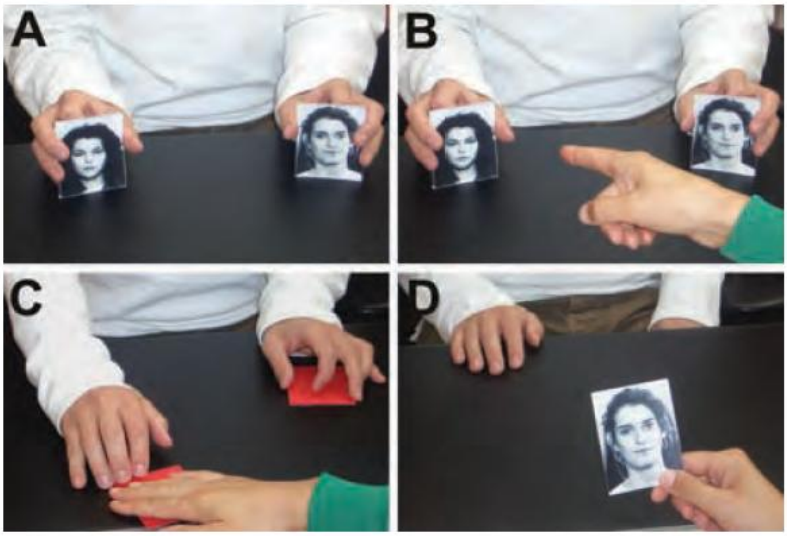


Figure 2. Choice blindness phenomenon (from: Johansson, Hall, Silkström & Olsson, 2005): individuals tend to justify their choice of *Y*, even when their original choice was, in fact, *X*. In first place (A) the participant is shown two faces, then (B) participants are asked to select the card with their highest preference. Immediately afterwards, (C), only on certain trials the experimenter switches the card previously selected by the participant. Finally (D), the experimenter hands the participant a card (card with the face not originally selected by the participant) and asks participants to justify their preference.

In opposition to the NWC, several conceptual and methodological objections have been proposed (Ericsson & Simon, 1980; Smith & Miller, 1978; White, 1988). Among the most important objections are the following: Firstly, (i) the NWC does not propose a clear distinction between a cognitive *process* and a cognitive *state*. Even though Nisbett and Wilson (1977) specify a list of contents or introspectively accessible mental states, the authors only offer a vague definition of a cognitive process: "*causes that guided, lead to or motivate a decision*" (p. 232). In a subsequent article, Nisbett and Ross (1980) define a cognitive process as "*the causal relation between events or mental contents*". Notice in both cases the pivotal notion of a *cause*, which is, in itself a very elusive concept. Since Hume (1711-1776), many philosophers have criticized causality as a relation or theoretical construct that is inaccessible to individuals' consciousness. Perhaps introspection of processes defined as causal is impossible because mental causation is difficult to define. Defenders of introspection (Engelbert & Carruthers, 2010), in line with the posture of the present thesis, in no case argue in favor of an introspective access of that type. My experiments will investigate access to the ongoing mental states occurring

before, during and after the decision, independently of their causal status with respect to our behavior. In fact, Nisbett and Wilson (1977) only distinguish states and processes as a function of their introspective accessibility, generating an arbitrary (Smith & Miller, 1978) and circular argument (White, 1980)[11]. A conceptual strategy to avoid this circularity, and that additionally facilitates the design of experiments, is to define a cognitive process simply as "*the sensory and representational transformations that occur between a stimulus and a response*". This operational redefinition has a double objective: eliminating *accessibility* as a criterion distinguishing cognitive states and processes, but also to simplify participants' introspective tasks. Indeed, when requesting participants to report "*the causes that motivate behavior*" it is highly probable that they would resort to *a priori* theories (because folk epistemology associates "causes" with "theory"). This effect should be decreased when it is demanded to simply report "*the sensory transformation*" that precedes a decision. This formulation can also be seen in many other authors (Engelbert & Carruthers, 2011; Schwitzgebel, 2008, 2011). The experimentation in this research will propose operational definitions for "sensory transformation", appropriate for the first order task.

Secondly, (ii) the NWC presumes that even in the case of evidence of a correct description of a cognitive process, this must not be considered as a real access but rather an *ad-hoc* theory, one that is only accidentally correct[12]. Indeed, the NWC is presented as a hypothesis that cannot be rejected in principle (Smith & Miller, 1978). In subsequent developments, (Nisbett & Bellows, 1977; Nisbett & Ross, 1980) the authors argue that introspection of complex processes must be studied through an *actor-observer* paradigm. That is, introspection should be investigated in an actor-group (*actors*: individuals that evaluate the causes of their own behavior), but also in an observer-group (*observers:* individuals that evaluate the causes that motivate the actors behavior). Under the premise that only actors have a privileged access to their cognitive processes by introspection, this group's report not only should be adequate (that is, agree with the real causes of behavior, known to the researcher), but also present a higher accuracy than the observer-group reports, which does not possess

[11] In terms of Smith and Miller: *"*[…] *the only difference between these two kinds of information* (states and processes) *is that people may in general be able to report on the first but not on the second. There is no other reason to consider one content* (state) *and the other process* […]" (Smith & Miller, 1978, p. 360). This circularity is also present in the new version of NWC, the adaptive unconscious (Wilson, 2002).

[12] In terms of the authors: "*subjective reports about higher mental processes are sometimes correct, but even the instances of correct report are not due to direct introspective awareness. Instead, they are due to the incidentally correct employment of a priori causal theories*" (Nisbett & Wilson, 1977, p. 233).

privileged access to such information[13]: Therefore, this was the reliability test for introspective data. Many authors criticized the use of between-subject designs to study introspection of cognitive processes (Kraut & Lewis, 1982; White, 1980; Wright & Rip, 1981). Smith and Miller (1978) argue that consciousness of a mental process is something that happens within the individual, therefore the evaluation of accuracy should be done on a case-by-case basis. Consequently, a hypothesis regarding consciousness of any mental content is always best evaluated with within-subject designs. In line with this suggestion, the vast majority, if not all contemporary research programs interested in introspective access consider within-subject designs (e.g., Corallo et al., 2008; Fleming et al., 2010; Marti et al., 2010; Palmer et al., 2014). Surprisingly, as far as we know, there are still no experimental studies that evaluate the NWC with such a design. This thesis is partly motivated by a desire to fill this gap in the literature.

Thirdly, (i) not only does the conceptual distinction between cognitive states and processes remain unclear, but (ii) also it is largely unknown whether the experimental context has a preponderant role in introspective inaccessibility. In effect, (iii), Nisbett and Wilson do not describe thoroughly the experimental conditions of the studies reviewed. Along these lines, the most problematic aspect[14] is the response mode[15], which is often based on verbal reports. The problems associated with introspective verbal reports have been extensively discussed in the literature surrounding it (Ericsson & Fox, 2011; Ericsson & Simon, 1980; Fox, Ericsson & Best, 2011; Schooler, 2002b, 2011). Many authors converge on the idea that verbal reports must involve a translation of the information of interest (Schooler, 2002a). This opens the possibility that cultural factors, *a priori* beliefs, personality factors or aspects of the researcher-participant relationship (demand characteristics) could alter the formation of introspective judgments[16]. Indeed, the problem of translation is particularly relevant in the NWC. Most, if not all, cognitive processes investigated by Nisbett and Wilson have high cognitive complexity; as a

[13] In summary, "[…] *if actor consistently gave more accurate reports about the reasons for their behavior than observers did, then this would indicate privileged sources of information underlying these reports. If not, then the position of NWC would be further supported* […]" (Johansson et al., 2006, p. 674).

[14] For instance, White (1980, 1988) suggests that most of the studies reviewed by the authors favor the automation of processing, which leads to a systematic loss of attention about the task. The introspective deficit could be due to this effect. Ericsson and Simon (1980) note that Nisbett and Wilson provided considerable background information to their participants in their experiments, so that they could have chosen to base their reports on the background information rather than on internal information.

[15] Particularly important is the fact that introspective task in a most of Nisbett and Wilson's experiments was applied not immediately after the cognitive process took place. In consequence, and under the conceptualization of introspection as retrospection, individuals when asked about the nature of their cognitive processing might simply not have such information in working memory (Bona & Silvanto, 2014). Therefore, they probably resorted to other sources of information in the elaboration of the introspective judgment (Ericsson & Simon, 1980).

[16] According to Schooler: "*If meta-consciousness* (introspection) *requires re-representing the contents of consciousness, then, as with any recording process, some information could get lost or become distorted in the translation. The likelihood of noise entering the translation process is particularly great when individuals (1) verbally reflect on inherently non-verbal experiences, and/or (2) assess ambiguous or subtle visceral signals*" (Schooler, 2002a, p. 342).

consequence, they demand that participants execute high-level forms of reasoning or at least integrate many information sources (e.g., selection of the best quality product). The high complexity of the task, added to the low experimental control of introspective verbal-report, are factors that favor confabulation (Ericsson & Simon, 1980). In fact, Ericsson and Simon observed in many studies that verbalization tends to be incomplete when participants are under high cognitive load. In response, recent advances in the field of introspection of mental states (Corallo et al., 2008; Marti et al., 2010) have all been achieved by focusing on elementary cognitive tasks, and replaced verbal reports by quantified reports.

The general objection can be summarized as the absence of a model which explains under which conditions an introspective report is reliable information and under which it is the end-product of inferential reasoning. Synthesizing the three criticisms, as well as considering new paradigms in experimental introspection, our proposition consists in re-evaluating the NWC from an (i) operational redefinition of cognitive processes and states, (ii) the implementation of a within-subject design to evaluate the introspective access to complex processes and (iii) a simplification of the experimental context with the purpose of avoiding translation problems.

1.2.2 A new framework for the research on Introspection

Our conceptual and experimental work is guided by the idea that the perceived inaccessibility of cognitive processes to introspective reports advocated by the NWC, derives from a poor representation of the mental content of interest, and this in turn is the product of a poor control over the multiple sources of information that contribute to the formation of an introspective judgment. In summary, a poor representation of a mental state is not necessarily evidence of a lack of access (White, 1988). Therefore, our objective is two-fold. On the one hand, we will try to experimentally re-evaluate the NWC (taking into account the previous critique). On the other hand, we will try to conceptually design a general model of introspective access from recent evidence in experimental metacognition. In order to do so, we need to determine the structural and functional factors that contribute to the limits of introspective access. The first process will be based on two functional and structural properties of introspection that can be inferred from the recent advances in studies on metacognition: (i) Evidence suggests that the formation of introspective judgments results from the accumulation of information from multiple (inner)

sources. In this sense, introspections are decisions, much like perceptions. (ii) There is also evidence of a high inter-individual variability in the introspective capacity.

Firstly, recent experimental evidence indicates that introspection would act as an information accumulator (Yeung & Summerfield, 2012), based on many sources of information, that may not strictly depend on the cognitive processes leading to the first order decision. For instance, Schwartz and Diaz (2014) showed that there is not a 1:1 correspondence between first order processes and the processes that drive introspection. Multiple sources of information contribute to the elaboration of a metacognitive judgment. A tip-of-the-tongue (TOT) experience (an indication of introspective judgment, as there is absolutely no external fact that can validate the feeling), would be determined by the familiarity of the stimuli, the combination and intensity of the surrounding information, the force of target activation (i.e., items of interest) by itself, among other factors. Schwartz and Diaz hypothesize that there would be multiple cognitive processes (i.e., sources of information) that underlie each introspective judgment (TOT, JOL or confidence judgments). Similarly, Lempert et al. (2015) find evidence for the dissociation between the mechanism through which individuals build an introspective judgment and first order decision, in that an increase in pupil dilatation correlates negatively with the introspection precision (confidence in decision making), independently of the difficulty during an auditory decision task. Thus, introspective judgments seem to integrate multiple sources of information, so as to create a new decision variable. Admittedly, it is unclear whether there is a single introspective mechanism or if such property requires multiple introspective focuses (Prinz, 2004) accumulating information in parallel (Merkle & van Zandt, 2006).

Moreover, the *locus* of the introspective decision, relative to the first order processes, is still highly controversial: is it the case that introspective evidence accumulates up until the moment of the first order decision (Decisional Locus Models (DLM): Kiani & Shadlen, 2009; Zylberberg, Barttfeld & Sigman, 2012)? or on the contrary, does it continue to integrate information after the decision (see Post-Decisional Locus Models (PDLM): Pleskac & Busemeyer, 2010; Resulaj, Kiani, Wolpert, & Shadlen, 2009)? Recently, Fleming et al. (2015) contributed evidence supporting the second alternative. The authors, through transcranial magnetic stimulation (TMS) altered the process of information accumulation, in certain cases prior to the participants' decision to a visual discrimination task or after the decision to the same task. They showed similar effects on introspective judgments when the TMS pulse was applied before, or after the decision, suggesting that the accumulation process is extended after the participants' response (PDLM).

Both the multiplicity of sources of information that form an introspective judgment, and the temporal extension of the accumulation process itself, allows us to suggest that participants could be able to introspectively recover information from different moments during first order processing. We hypothesize the presence of both early and late introspective integration. Thus, it could be possible to imagine that participants can flexibly activate an early recovery process of the information related to perceptual load during a cognitive task, but also a late recovery process related to information on performance in the same task. Previous authors have implicitly assumed that introspection can be guided to different points in the development of the task (Goldman, 2004). Hurlburt and Heavey (2004), in the context of experience sampling protocols, suggest the use of a marker (the beep procedure) to signal to participants the moment when they must introspect. This method would specify the cognitive stage where the hypothetical mental content of interest is generated and to which introspection must be applied. According to this view, the distinction between DLM and PDLM could be understood in terms of a flexible introspective focus, either on early or late task processes. In addition, some theorists (Schwitzgebel, 2011) have recently suggested that there is little control from researchers and participants regarding how different sources of information impact introspection. Indeed, if the focus of introspection is labile, it is also possible to imagine that "confusion of the source" may occur: participants may have difficulty in discriminating the mental content of interest from irrelevant information that would lead to confabulations. Ericsson and Simon's (1980) *think aloud* procedure was precisely meant to track and follow as closely as possible the evolving content of the stream of consciousness, so that at each time point, the content of the report would exactly match the mental state. But of course, the temporal resolution of verbal discourse is poor compared to the stream of thought, notably because of the syntactic constraints in language. Therefore, in order to advance on this issue, the introspective recovery process towards certain mental content is manipulated experimentally in this research (regardless of the time when these were generated in the experimental task).

Another important constraint on introspection that has been put forward recently is the high inter-individual variability in the accuracy of introspection. These inter-individual differences have been linked to structural variability in brain regions that subserve introspection. Indeed, recent studies associate both medial and lateral regions of the anterior PFC (Baird et al., 2013; de Martino et al., 2013; Fleming et al., 2010; Fleming et al., 2012; Rounis et al., 2010; Yokoyama et al., 2010), as well as the dorsal premotor cortex (Fleming et al., 2015), with the precision of metacognition. Importantly, variations in participants' introspective sensitivity have been

shown to covary with morphological variability in these regions. The importance of these findings is that if adequate control over the introspection mechanisms is achieved, there would still be a structural modulator that determines the precision of introspection.

1.3 Experimental studies

The main aim in this thesis was to investigate whether one could have introspective access to cognitive processes, as opposed to introspective access to the cognitive states or to the behavioral consequences that they generate. The NWC model will thus be assessed; but we shall do so by means of quantified introspection over well studied and constrained simple cognitive tasks. The objective of the study has not been re-evaluated recently (however, see Petitmengin et al., 2013), as most recent advances on metacognition has focused on the mechanisms of confidence judgments. However, confidence would clearly be classified as a "cognitive state" in Nisbett & Wilson's classification. In addition to confidence judgments (Fleming et al., 2010; Pleskac & Busemeyer, 2010), judgments of duration of perceptual decisions (Corallo et al., 2008; Marti et al., 2010; Miller, Vieweg, Kruize & McLea, 2010, Bryce & Bratzke, 2014) have also attracted some attention. In order to go beyond reports on cognitive states, a new introspective measure, the "Subjective Number of Scanned Items" (SNSI) was devised, which is applicable whenever the first order task asks for the processing of more than one item. Then, the SNSI consists in the number of items that the participant consciously processed before the decision. The methodology is applied in visual searches (e.g. "how many items did you scanned before you found the target?") and in working memory searches (e.g. "how many items of the list did you review before deciding whether the target was present in the list?"). The hypothesis of this research is that this number is an index of participants' access to the processes of the searches themselves.

A problematic aspect of subjective reports is that they are highly dependent on the experimental context (Goldman, 2004). Recent conceptualizations suggest that access to different mental states would require different *species of introspection* with different forms of measurement (Prinz, 2004). It has even been suggested that different experimental approaches (e.g., *think aloud procedure*: Ericsson, 2003; Ericsson & Simon, 1980; *experience sampling*: Hurlburt & Heavey, 2004; *quantified introspection method*: Corallo et al., 2008; Marti et al., 2010; *script-report procedure*: Jack & Roepstorff, 2002, etc.), could be associated with different levels of metacognitive access (Overgaard & Sandberg, 2012; Sandberg et al., 2010). Taking into account this introspective pluralism (Schwitzgebel, 2011), the procedure in this research (SNSI), was designed to reduce the possibility of obtaining confabulatory introspective responses: Firstly, my research is not focused on introspection of higher level processes (as are almost all experiments examined by Nisbett and Wilson (1977). The first order tasks that we study are complex in the sense that they are potentially multi-step tasks (a search

might be fully serial, with examination of one item at a time). But they are not high level tasks, in the sense that the signal that contributes to the decision is precisely defined and controlled. Secondly, verbal re-coding of the experience was not required, as it has been suggested that this could favor over-interpretations (Ericsson & Fox, 2011; Fox et al., 2011; Schooler, 2002a; Schooler, 2011): SNSI reports are given on quantitative scales. Thirdly, this research procedure evaluates introspection on a trial-by-trial basis, an important aspect criticized in the studies examined by Nisbett and Wilson (White, 1988). Finally, the experimental strategy, inspired by the Jack and Roepstorff (2002) procedure, investigated whether introspection is capable to detect changes in the cognitive process *strategically* generated by the experimental conditions. The general prediction in these experiments was that participants' introspection will be sensitive to such changes, suggesting that this ability can access the complexity of the cognitive process deployed in each case.

In the first paper, the investigation focused on whether introspection can access and distinguish two different cognitive processes that underlie two kinds of visual searches. A series of experiments were designed which contrasted "pop-out" and "serial" visual searches. The strategy was to evaluate whether introspection could track whether the search had been serial or parallel. The results showed that participants were able to introspect the search processes, provided that they had enough time to deploy their introspection on the decision. In addition, indications showed that participants were in fact reporting how many attentional switches they performed before one decision.

In the second paper, the hypothesis that participants had access to the attentional guidance during visual searches was specifically tested. In order to do so, a brief cue was introduced, unbeknownst to participants, before the search array, so as to manipulate exogenously their attention. Importantly, our cue was so subtle that participants did not spontaneously notice its presence. The important point here is that it is possible to avoid the impact of potential contaminants related to the experimental setting on the introspective judgments. As participants did not know that there was a cue, attentional guidance was manipulated during the search, while participants' perception was not affected. The findings demonstrated that while participants were unaware of the cause of the modification of their search processes, their introspective reports showed that they accurately access the processes themselves.

In the third paper, the aim was to extend these findings to another cognitive domain. Following the same strategy, participants' ability to introspectively distinguish the cognitive processes underlying two working memory tasks was evaluated. Participants were instructed (on a trial-by-trial basis) to report the number of items that they scanned during their working memory retrieval (SNSI), whilst they performed either a Judgment of Recency (JOR, Hacker, 1980; Muter, 1979) or an Item Recognition (IR, Sternberg, 1966) task. The literature converges on the notion that JOR task are serially processed, while IR task are processed in parallel. Again, results showed that participants' introspection could track this distinction, showing that the process of the memory searches was accessible to them, and not only its end result.

Finally, in the fourth and fifth papers, a new factor of inter-individual variability in introspection was investigated: the biological reactivity to stress. Motivated by recent experimental models (Hermans, Henckens, Joëls & Fernández, 2014) which suggest a negative relation between stress and high cognitive functions, it was hypothesized that individuals with high reactivity to stress would present an alteration in introspective sensitivity. Firstly the impact of stress on introspection in the context of the visual search tasks was evaluated. Next, the impact of stress reactivity on the precision of confidence judgments in a perceptual decision task was analyzed. In both cases the results demonstrated a high stress reactivity was associated with poorer introspection.

2. Experimental Studies

2.1 Introspection during visual search

Reyes, G., & Sackur, J. (2014). Introspection during visual search. *Consciousness and Cognition, 29*, 212-229.

2.1.1 Introduction

As we mentioned in the introduction, recent advances in the area of experimental introspection suggest that individuals could access several cognitive states. However, the cognitive processes that underlie these cognitive products would be inaccessible to the consciousness of the participants. Attempt at accessing such mental contents would create confabulatory judgments (e.g. inferences or *ad-hoc* theories). This thesis, that we termed as Nisbett and Wilson's Canon (1977) hasn't been experimentally studied. The paper investigates the possibility to introspectively access and describes a cognitive process.

To this aim, we investigated whether participants could introspectively access the attention shifts (the cognitive processes) in two types of visual searches: feature and conjunction searches (Treisman & Gelade, 1980). In the line of the quantified introspection method (Corallo et al., 2008; Marti et al., 2010), we instructed participants to give, on a trial-by-trial basis, an estimate of the number of elements scanned before the perceptual decision (the Subjective Number of Scanned Item, SNSI). The experimental design implemented allows contrasting the variability of first order indicators (e.g. response times, error rate, eye movement, etc) with the variability of introspection. If access to cognitive processes is possible, there should be evidence of an introspective variability according to experimental control. At the same time, such effect should not be completely explained by the variability in the first order indicators.

2.1.2. Results

Results show that participants gained access to the nature of the search process through introspective estimation of the number of attention shifts. In Experiment 1, a cognitive-process introspective measure (SNSI) and two cognitive-state introspective measures (Introspective Response Times (Corallo et al. 2008) & Confidence Judgments) were implemented. Results indicate that the access to a cognitive *process* is highly limited , in contrast to what happens with the access to cognitive *states* (decision times and confidence in the decision).

Experiments 2 and 3 investigated the pertinence of the experimental context. We observed that access to the nature of a cognitive process is highly influenced by both the design of the task and the first-order surrounding information. Experiment 4 (overt attention setting) indicate that SNSI, regardless of the fact that it is able to capture changes in the cognitive process, is systematically influenced by search time (response time, RTs) and the number of saccades during the visual search task (eye movements). Experiment 5 controls ocular movement (covert attention setting). Results showed that participants' introspection (SNSI) was identical on both experiments (Exp. 4 & Exp. 5).

2.1.3. Discussion

This paper is the first attempt in experimental psychology of evaluating Nisbett and Wilson's canon in detail. From these series of experiments it is possible to maintain that, under precise experimental controls, introspection of cognitive processes is possible. We propose the hypothesis that introspection is relies on information accumulation both from internal (mental monitoring) and external (self-observation) sources.

2.1.4. Paper

Introspection during visual search

Gabriel Reyes [1, 2] and Jérôme Sackur [1, 3]

[1] Laboratoire de sciences cognitives et psycholinguistique, CNRS/EHESS/ENS, Paris, France

[2] Université Pierre et Marie Curie, Paris, France

[3] Institut Universitaire de France, Paris, France

Corresponding author: G.R. and J.S., Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 29 rue d'Ulm, 75005, Paris, France. E-mails: gureyes@uc.cl; jerome.sackur@gmail.com. Tel: + 33 (0) 1 44 32 26 25.

Word count (abstract + main text): 157 + 12.151 = 12.308

Abstract

Recent advances in the field of metacognition have shown that participants are introspectively aware of many different cognitive states, such as confidence in a decision. Here we set out to expand the range of experimental introspection by asking whether participants could access, through pure mental monitoring, the nature of the cognitive processes that underlie two visual search tasks: an effortless "pop-out" search, and a difficult, effortful, conjunction search. To this aim, in addition to traditional first order performance measures, we instructed participants to give, on a trial-by-trial basis, an estimate of the number of items scanned before a decision was reached. By controlling response times and eye movements, we assessed the contribution of self-observation of behavior in these subjective estimates. Results showed that introspection is a flexible mechanism and that pure mental monitoring of cognitive processes is possible in elementary tasks.

Keywords: introspection, perceptual decision, cognitive processes, consciousness, metacognition.

Introduction

Humans are endowed with introspection, which is the ability to monitor their own mind. For a long period in the history of experimental psychology this ability was viewed with some suspicion, mainly because introspection as a method for the investigation of cognitive functioning was largely unsuccessful (see a review in Boring, 1953; Danziger, 1980; Lyons, 1986; Costall, 2006; Sackur, 2009). However, since the recent re-conceptualization of introspection as an intrinsic feature of consciousness (Piccinini, 2003; Goldman, 2004; Feest, 2012), it has been reconsidered as a legitimate field in cognitive psychology (Jack & Shallice, 2001; Schooler, 2002; Schooler & Schreiber, 2004) and amenable to experimentation in neuroscience (Jack & Roepstorff, 2002; Fleming, Weil, Nagy, Dolan, & Rees, 2010; Fleming & Dolan, 2012; Baird, Smallwood, Gorgolewski, & Margulies, 2013).

Despite great progress in the science of introspection in recent years, an issue not yet resolved is: what mental content is accessible to introspection? In the wake of Nisbett and Wilson's seminal paper (Nisbett & Wilson, 1977), researchers have been very wary of the kinds of introspective reports they should elicit from their participants. Nisbett and Wilson gathered considerable empirical evidence and theoretical arguments to the effect that one should clearly distinguish reports on internal cognitive *states* as opposed to internal cognitive *processes*. While the former may, in some context, be introspectively accessed, the latter were deemed, by and large, inaccessible. Thus, asking participants about them would most often lead to confabulations. Nisbett and Wilson held that the process that links a stimulus and the response does not reach participants' consciousness, and that only cognitive products or states are consciously accessed (see also Neisser, 1967). Despite initial substantial objections (Smith & Miller, 1978; White, 1980, 1987, 1988; Ericsson & Simon, 1980), and recent reformulations (Wilson, 2002, 2003), this idea is considered as a canon of the literature on metacognition (Johansson, Hall, Sikström, & Olsson, 2005; Overgaard, 2006; Overgaard & Sandberg, 2012).

In recent years, the set of responses that may qualify as introspective has considerably increased. Among these, traditional confidence ratings (e.g., Pleskac & Busemeyer, 2010; Fleming, et al., 2010; Song, Kanai, Fleming, Weil, Schwarzkopf, & Rees, 2011) have been reconsidered in depth, and new ones, such as judgments of duration of perceptual decisions (Corallo, Sackur, Dehaene, & Sigman, 2008; Marti, Sackur, Sigman, & Dehaene, 2010; Miller, Vieweg, Kruize, & McLea, 2010) have come to the fore. However, it is important to note

that all these new forms of introspection are reports on internal cognitive *states*, and thus all abide by Nisbett and Wilson's canon. In this paper, we seek to put this limitation under experimental scrutiny.

It is interesting to note that most cognitive processes that Nisbett and Wilson target are very complex, high-level forms of reasoning. Recent advances in the field of introspection of mental states have all been achieved by focusing on elementary cognitive tasks. For instance, Corallo et al. (2008) and Marti et al. (2010) selected the well-studied *Psychological Refractory Period* phenomenon, as a first order cognitive task, and asked participants to report on the durations that they introspectively perceived while performing this task. Here, we ask whether participants are introspectively aware of a difference in the kinds of processes triggered by two well-attested first order experimental tasks.

We relied on the following basic paradigm: we instructed participants to perform a visual search task in two different conditions, one simple and fast, in which the target "pops out", the other being more difficult and requiring an effortful exploration of the visual scene. Concurrently, on a trial-by-trial basis, we collected quantitative introspective reports. Our aim was to assess whether these introspective reports correlated with differences in processing that we could infer from a third-person, external standpoint.

We chose visual search as a first order task, as it is known that in this task minimal changes in the stimuli induce important changes in performance profiles, indicative of a switch between two modes of processing. Traditionally, searches were construed as either *parallel* or *serial* processes (Sternberg, 1966; Townsend, 1990). In visual search, Treisman's seminal Feature Integration Theory (FIT, Treisman & Gelade, 1980) contrasted *feature searches* and *conjunction searches*; the former producing parallel searches and the latter serial searches. This difference was meant to account for the empirical finding that in feature searches, mean Response Times (RTs) do not increase as the number of distractors is increased, while in conjunction searches, mean RTs increase linearly as a function of the number of distractors. FIT asserts that in feature searches the visual system extracts in parallel, pre-attentively, the set of basic characteristics of the scene, which are necessary and sufficient to select the response. On the contrary, in *conjunction searches* attention is deployed serially one item, or group of items, at a time.

A strict dichotomy between parallel and serial searches is no longer tenable (Eckstein, 2011). First, it has been known for a long time that linear increase in RTs is not diagnostic of serial processing (e.g., see *model mimicking* in Townsend & Wenger, 2004). Second, it appeared that there is a continuum of more or less efficient searches (Wolfe, Cave, & Franzel, 1989; Wolfe, 1994, 2007; Thornton & Gilden, 2007). The current consensus is that inefficient visual searches exhibit prominently *capacity limits*, whereas efficient searches do not incur such limits. Furthermore, it is also widely admitted that easy, efficient searches evade capacity limits because they benefit from *guidance of attention* by features extracted from non-selective pathways (Wolfe, 2003; Wolfe, Horowitz, 2004; Wolfe, Võ, Evans, & Greene, 2011, but see McElree & Carrasco, 1999; Cameron, Tai, Eckstein, & Carrasco, 2004). Our objective was to test whether participants can introspectively access the presence or absence of capacity limits and of attentional guidance.

Of course, no decision process is ever absolutely without "capacity limits", and visual searches are no exception to this rule. For instance, Joseph, Chun, and Nakayama (1997) showed that even highly efficient pop-out searches are subjects to capacity limits when performed in conjunction with an attention depleting dual task. This feature is nicely accounted for by dual stage models of visual search (Wolfe, 2003) where the second, response selection stage is viewed as a central decision stage, subject to bottleneck effects. The key point for us is that, in the absence of concurrent tasks, in efficient searches the response selection stage can benefit from the parallel feature extraction in the first stage, through attentional guidance. Inefficient searches cannot benefit from attentional guidance, and thus always exhibit bottleneck effects that result in slower RTs with increasing set-size.

In all our experiments distractors were schematic Ts, while targets were either an X or an L. These stimuli are known to produce two clearly different search profiles. Without theoretical commitments, we will refer to searches of an X among Ts as Feature Search (FS, targets defined by a single orientation feature), and to searches of an L among Ts as Conjunction Search (CS, targets defined by the specific conjunction of two features that are also present in the distractors). After each decision on the search task, participants were instructed to report the number of items that they had scanned before giving their response, a measure that we termed "Subjective Number of Scanned Items" (SNSI). We predicted that participants' estimations would be constant and close to one item in FS, independently of the number of distractors on the screen. In contrast, we predicted higher SNSI scores in CS than FS, and crucially, an increase as a function of set-size. One may think of this measure as the subjective counterpart to the "scanning process" of Sternberg's (1966) pioneering work on memory search.

Two important aspects of the SNSI measure should be emphasized here: first, this measure is an *index* of putative differences in processing. We did not ask our participants to report directly on the type of processes involved in a particular trial, but we reasoned that if there were any such introspectively accessible differences, they should show in the number of subjectively scanned items before the decision. Second, we expect our index to be analytical or pure (Sternberg, 2001), to the extent that it captures only one among presumably many different kinds of introspective information. That is, our SNSI index attempts to selectively isolate the introspective contribution of capacity limits in visual search. Note that Miller et al. (2010) already tried for a pure measure of subjective (introspective) decision *duration*.

However, even with this framing of the introspective task, we cannot rule out the contamination of SNSI responses by other introspective information (Piccinini, 2003; Goldman, 2004; Prinz, 2004). Contamination could occur strategically because, for instance, participants notice that SNSI correlates with duration and find duration easier to access; or it could occur unconsciously, as a bias in SNSI reports. Furthermore, we should allow for the possibility that the information that the SNSI targets might simply *not* be introspectively accessible. In this case, data from the SNSI scale would be purely experimental artifacts: since we force participants to select a value on the scale, they might comply and simply report something which they think (i.e., according to *their* theory of search processes). Indeed, this is the straightforward prediction from Nisbett and Wilson's confabulation model.

In order to meet these challenges, we adapted a multi-level mediational approach (Bauer, Preacher, & Gil, 2006) as an analytic strategy of *reliability* (Piccinini, 2003), trying to detect whether any effect found on the SNSI scale is explained away when behavioral variables (i.e., RTs and eye-movements) are taken into account. This approach distinguishes *self-observation* and *mental monitoring*. Knowledge about oneself, even about one's own mental processes, can derive both from direct access to mental processes, or through inferences based on self-observation of behavior. Both qualify as introspection in a broad sense, but only the first is pure introspection, which may be more adequately termed *mental monitoring*. While this distinction was clearly stated in Nisbett and Wilson's seminal paper, it may have been under-appreciated in more recent experimental studies of introspection. The mediational approach is aimed at weighting the relative contributions of mental monitoring and self-observation in an introspective task.

The first three experiments delineate the conditions under which participants are able to introspect on the search processes. We show that even though self-observation of response times could account for a significant portion of the introspective judgments, we can set-up experimental conditions that permit mental monitoring of the processes themselves. Next, in the last two experiments we try to insulate introspective judgments from the contaminants that we identified or suspected in the first experiments. In Experiment 4a, we factor out response times and we measure eye movements, while in Experiment 4b, we both control response times and eye movements.

2.1.4.1 Experiment 1

In this first experiment, we asked participants to detect a visual target in an array of distractors, and after each response, we asked them to report on a quantitative scale the number of items they felt they had scanned before they reached their decision (Subjective Number of Scanned Items, SNSI). In addition we also collected traditional introspective measures: confidence judgments and introspective response times (iRT).

Material and methods

Participants

Thirteen normal adults, French speakers (10 women), aged between 20 and 29 (mean age: 24.3 years, *SD*: 3.5) participated in the study. In this, as in the experiments which follow, informed consent was obtained before the experimental session, and participants received compensation of €10 for each 1-hour session. None of the participants had any knowledge regarding the study and all had normal or corrected to normal vision.

Stimuli

Stimuli (see Figure 1) consisted of a set of black letters (T, L or X, size: 0.8° x 0.6°, luminance: 0.09 cd/m²) on a uniform grey background (luminance: 3.11 cd/m²), presented on an imaginary circle (radius: 6.2°) around a central fixation spot at the center of the screen. Individual orientation for each letter was randomized (0, 90, 180, 270°). Stimuli were equally spaced on the imaginary circle, while its overall orientation was randomized for each trial. Stimuli were presented on a CRT screen (size 17", resolution of 1024 x 768 pixels, refresh rate of 100 Hz, viewing distance ~ 55 cm). The experiment took place in a dark booth with the monitor as the only source of light.

Task and Procedure

Stimuli were presented for 200 *ms*, preceded by a fixation spot presented for a duration drawn from the interval of 400-700 *ms*. Participants were instructed to decide on the presence or absence of a target (L or X) within the set of distractors (Ts), by pressing as quickly and accurately as possible, with the index and middle fingers of

their left hand, either the “A” or “Z” key on a standard AZERTY French keyboard. Half of the trials were target absent trials, with only distractors. Target present trials contained one “L” or one “X”. Set-size (2, 4, 8 or 12 items, including target if present) and presence or absence of a target were fully crossed. Immediately after the perceptual decision, three continuous introspective scales were presented within the same display: i) *Confidence*: Are you certain of your decision? Labeled at the two extremes with “guess” and “absolutely certain”; ii) *Subjective Number of Scanned Items* (*SNSI*): How many items do you think you examined before reaching your decision? This scale ranged from a minimum of “0”to a variable maximum, equal to the set-size of the trial; iii) *Introspective estimate of the response time* (*iRT*): How long do you think that it took you to determine whether the target was present or absent? This was a graduated scale ranging from 200 *ms* to 1200 *ms* with marked intervals of 100 *ms*.

Position of the scales on the screen was constant during the experiment. Participants used their right hand to move the cursor with the computer mouse, and click on the scales to give their quantitative introspective estimates. Meaning and use of the introspective scales was explained before the main experiment, while during the experiment instructions were presented in an abbreviated manner below the scales. Participants were instructed to avoid fast or automated responses.

Before the experimental blocks participants received two-stage training. During the first stage of 16 trials the visual search task, with a lengthened duration of 800 *ms*, was presented without the introspective scales but with audio feedback on correct and incorrect responses. This phase was repeated until participants reached a performance of 90% correct. The second training, also comprising 16 trials, introduced the introspective scales. Feedback was given on the response time estimate: a blue bar above the scale, which indicated the objective response time, after the participant's estimate had been given. During the second stage, the primary task was presented at 200 *ms* and participants proceeded to the main experimental block without the performance criterion. The experimental session comprised 480 trials (120 repetitions per search condition) in 10 blocks with a 60 second pause between each block. The experimental session lasted ~1 hour.

Training session

The day before the experimental session, participants took part in a training session (480 trials, one hour) which was in all respects identical to the main experimental session with the exception that target types were blocked.

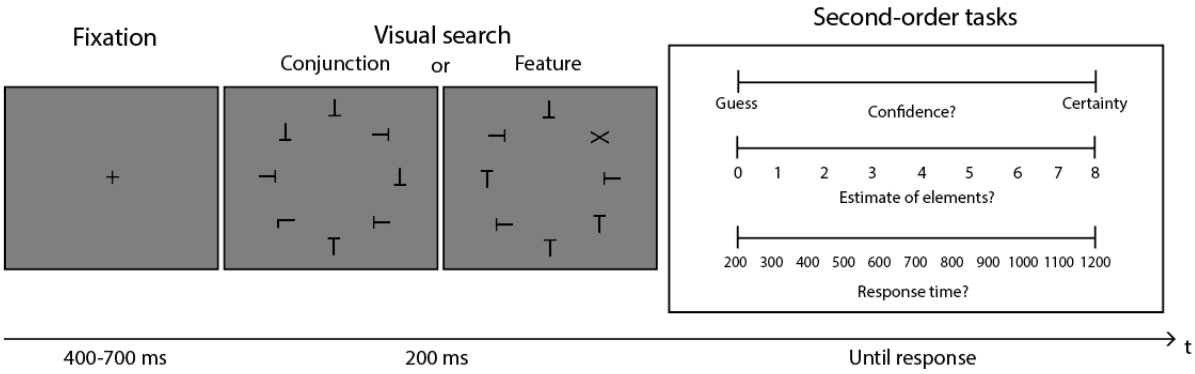


Figure 1. General structure of the task in Experiment 1. The scales appeared immediately after participant's response, and their position on the screen was fixed throughout the experiment. Instructions insisted on the fact that the "subjective number of scanned items" scale required a report of the number of items scanned *before* the identification of the target.

Results

First order task

First, we wanted to verify that the two search conditions were opposed as regards capacity limitation, as is classically reported in the literature. We excluded trials with response times below 200 *ms* and trials with response times 3 *SD* above the median (3.8%).

Here, and in all following analyses, we used Linear Mixed Models (LMMs) with fixed effects of search type (feature search, FS *vs.* conjunction search, CS), set-size (2, 4, 8, 12) and their interactions. As random effects the models included intercepts and a random slope for set-size for each participant. In all LMMs we used the restricted maximum likelihood (REML) as fitting method.

Response times and error rates were correlated (present target trials: $r^2(104) = .30$, $\beta_{Stand.} = .55$, $t = 6.58$, $p < .001$; absent target trials: $r^2(52) = .21$, $\beta = .46$, $t = 3.64$, $p < .01$, see Figure 2A). Thus, we computed an Inverse Efficiency Scores (IES: ratio of median RTs over proportion of correct responses, see Townsend & Ashby, 1983; Austen & Enns, 2003; Bruyer & Brysbaert, 2011), which provides a concise summary of the first-order results. [L]ower values correspond to better performance. Before calculating IES, RTs were log-transformed to approximate normal distribution.

We found the pattern of interaction between target type and set size (see Figure 2B), which is typical of the opposition of capacity limited and non-capacity limited searches. We ran an LMM on IES on target present trials, and found that the two main effects were significant (search type: $F(1,84.5) = 4.72$, $p < .05$, the set-size: $F(1,11.6) = 4.73$, $p < .05$) as well as the interaction ($F(1,84.8) = 8.22$, $p < .01$). A more detailed examination indicated that while in CS, IES increased as a function of set-size ($F(1,12.0) = 7.79$, $\beta = .28$, $p <. 05$), it was constant in FS ($p >.53$). When the analysis was repeated on the trials without a target, a significant increase of IES by set-size was shown ($F(1,13.8) = 7.19$, $\beta = .72$, $p < .05$). IES in these trials was higher than in target present trials ($F(1,88.7) = 38.12$, $p < .001$). In sum, these results validate the choice of targets and distractors: searching an L among Ts is increasingly difficult with increasing set-sizes compared to searching an X among Ts. This lends support to the idea that searching an L is capacity limited as opposed to the search for an X.

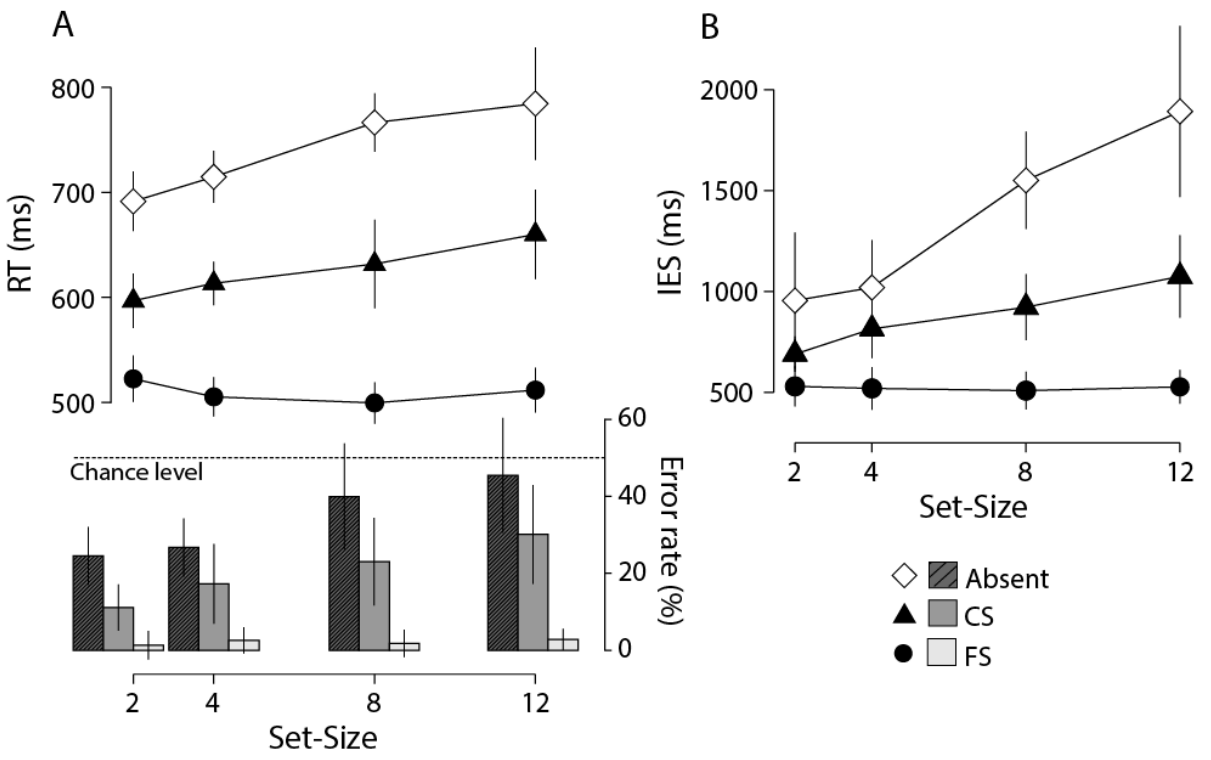


Figure 2. A) Response time and Error rate as a function of set size in both visual search conditions and absent target trials, in Experiment 1. Error bars, here and in the following experiments, are Cousineau-Morey within-subjects 95% confidence intervals (Cousineau, 2005; Morey, 2008), calculated separately for present / absent target trials. B) Inverse efficiency scores (IES) as a function of set size in CS, FS and absent target trials, in Experiment 1.

Second order task

After each first order response, participants gave three second order responses: confidence, introspective response time (iRT), and subjective number of scanned items (SNSI), this last response being the focus of our investigations. All *p* values were Bonferroni corrected (*p*(cor)), to account for the 3 dependent variables.

Confidence decreased in CS as a function of set-size, but stayed high for FS at all set-sizes (see Figure 3A). This was confirmed statistically: we ran the previous LMM on mean confidence index (anchored 0 / 1), which showed a significant main effect of set-size ($F(1,12.2) = 11.32$, $p(\text{cor}) < .05$), no effect of search type, ($p(\text{cor}) > .45$), and a significant interaction between these factors ($F(1,81.4) = 16.43$, $p(\text{cor}) < .001$). Importantly, we found

a significant and negative slope for CS ($F(1,12.0) = 13.40$, $\beta = -.01$, $p(\text{cor}) < .01$), while it was not significant in FS ($p(\text{cor}) > .54$). In absent trials, the confidence index significantly decreased as a function of set-size ($F(1,90.0) = 16.13$, $\beta = -.02$, $p(\text{cor}) < .001$).

Next, we regressed iRT (see Figure 3B & 3C) on RT, across correct-present individual trials, for each search condition separately. The regression slope was different from zero in both conditions (CS: r^2 (1199) = .13, β = .34, SE = .02, $p(\text{cor}) < .001$; FS: r^2 (1517) = .16, β = .58, SE = .03, $p(\text{cor}) < .001$), indicating that participants could access their response times in each search conditions. The difference between these two slopes was significant, ($F(3,2717) = 173.1$, $\beta = .23$, $SE = .04$, $t = 5.60$, $p(\text{cor}) < .001$), which indicates a better introspective access of RTs in FS than CS. The same regression on trials without a target yielded a significant slope (r^2 (1956) = .07, $\beta = .24$, $SE = .07$, $p(\text{cor}) < .01$). In sum, results on these two second-order tasks show that participants have introspective knowledge about their performance.

Now we come to the subjective number of scanned items (SNSI), which tracks the subjective accessibility of capacity limitations during the search. SNSI increased as a function of set-size (see Figure 3D), but did not reveal any difference between search conditions. Set-size effects were found both in target present trials ($F(1,12.0) = 6.73$, $p(\text{cor}) < .05$) and in target absent trials ($F(1,12.0) = 8.53$, $p(\text{cor}) < .05$). No other main effects or interaction were significant. This suggests a general effect of the number of items displayed, without introspective access to the difference in the search processes involved.

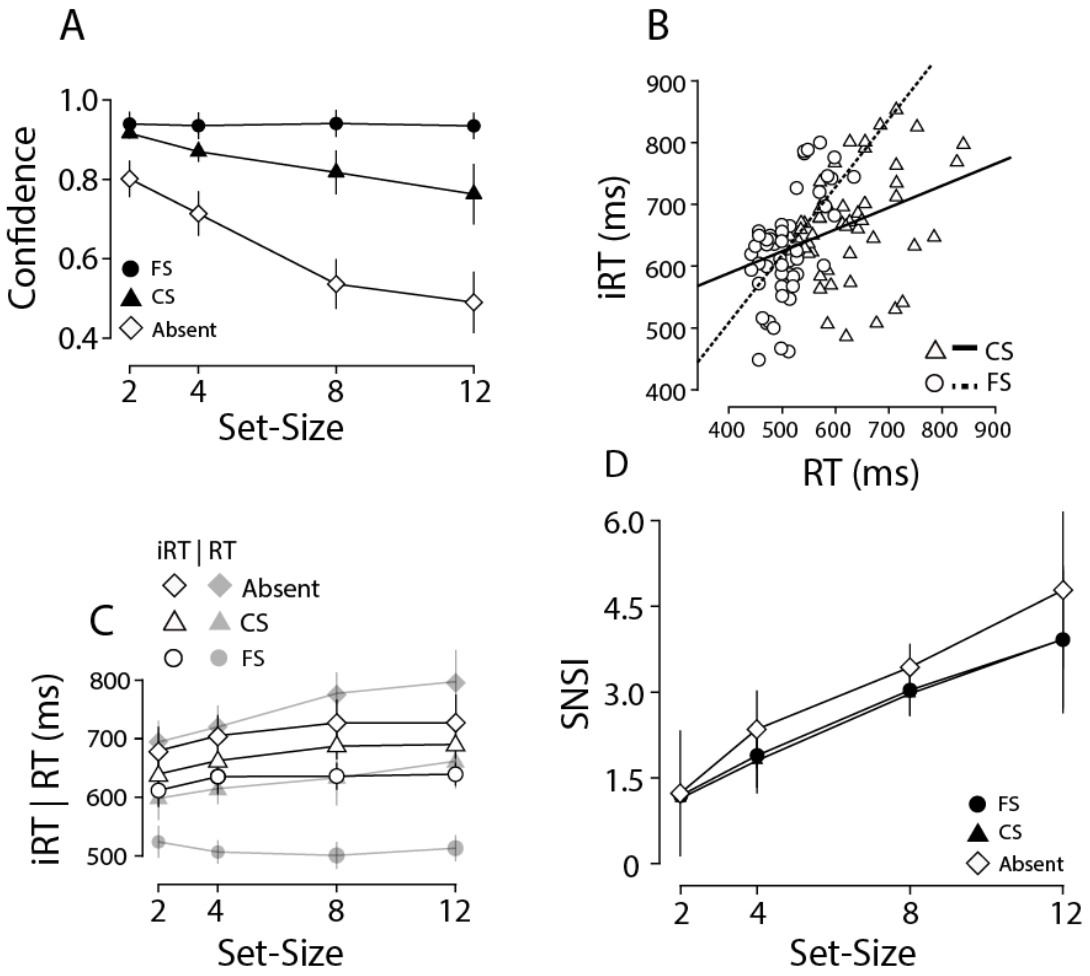


Figure 3. A) Confidence index as a function of set-size in both search conditions and absent target trials, in Experiment 1. B) Linear regression of mean RT on mean iRT in CS and FS condition and C) iRT (black lines) and RT (grey lines) as a function of set-size in both search conditions and absent target trials, in Experiment 1. D) SNSI as a function of set-size in CS, FS and absent trials, in Experiment 1.

Discussion

In agreement with the extensive literature on visual search, we found that a target with a distinctive feature (an X among Ts) gave rise to an efficient, pop-out search, evidenced by a flat slope in all first order measures (i.e., RTs, Error rate and IES) with increasing set-sizes. In contrast, the search for a difference in conjunction of the same two features (an L among Ts) yielded inefficient searches: an increased number of distractors decreased performance. Thus, our conjunction search stimuli did create capacity limitation which is not present in feature search.

Results on the introspection of the number of scanned items do not parallel the objective, first order results. Our prediction was a flat slope for FS as a function of set-size and a steeper SNSI slope for CS. We found that the number of items scanned increased in both search conditions as a function of set-size, without significant differences between them. The absence of any reported subjective difference between the two searches forces us to conclude that participant have no introspective access to capacity limitation.

Furthermore, as demonstrated by the results on the iRT and confidence scales, our participants were able to report well-established second order parameters: Confidence correctly tracks task difficulty, and iRT follows objective RT. Both subjective measures reveal introspective knowledge of the general structure of the experimental control, indicating that, after the decision has been made, participants are aware of some general properties of their decision processes.

The pattern of results we find runs directly counter to what previous quantified introspection paradigms would lead us to predict. Indeed, results using the Psychological Refractory Period paradigms (Corallo et al., 2008; Marti et al., 2010) pointed to a greater subjective availability of central decision processes as opposed to perceptual stages in an elementary cognitive task. Here, we found the opposite: set-size factor which is the more perceptual of the two, gives rises to differentiated introspection, whereas search type, which is more central, as it directly modifies the nature of the decision process, does not. Notice that in a sense this null result, ironically, is a good defense against the charge that high level introspective questions should not be used, because reports will be tainted by confabulations (Nisbett & Wilson, 1977). While the stimuli were easily differentiated retrospectively and their impact on the difficulty of the decisions was accessed through confidence and the duration of the search, participants did not confabulate.

The increase in SNSI with set-size may indicate that participants maintain a fixed width attentional window, irrespective of guidance (Wolfe, 1994, 2007). By necessity, such a window would encompass more items as set-size increases (Young & Hulleman, 2012), because the imaginary circle on which our stimuli are positioned has a fixed radius. According to this hypothesis, SNSI indexes the quantity of information recovered in parallel during the first pre-attentional stage of the search, but would not selectively distinguish the type of attentional control specific to each type of search.

If one takes into account both the effect of set-size on SNSI and the results on the confidence and iRT scales, our results are in overall agreement with Nisbett & Wilson: on a trial-by-trial basis, participants are introspectively aware of the perceptual load of the stimulus; they are also introspectively aware of some state consequences of the cognitive processes involved (confidence and self-observed global response duration); but they are mainly unaware of the processes themselves. However, this interpretation is open to methodological objections, as it rests on a null result. This could be the consequence of a deficiency at any of the following levels: (i) the cognitive difference targeted might not exist; (ii) introspection might not be able to access it; (iii), the means we give our participants to report their introspection might be inadequate.

This third possibility seems ruled out by the fact that there is a significant impact of set-size. However, before we can proceed any further, we first need to address the first objection, namely that we did not find any introspective difference between the two search types because they did not generate different processes: Both might be equally guided and capacity limited. Thus, we need independent empirical evidence of differential capacity limitations in our two search conditions, in the exact context of our stimuli and tasks. We designed the next experiment to address this issue.

2.1.4.2 Experiment 2

In the second experiment, we re-assess whether our stimuli generate distinct search processes, so that it may make sense to look for our participants' ability to gain introspective knowledge of them. We appealed to the following *objective* method: we introduced trials with two identical targets and participants had to report whether there were 1 or 2 targets. We reasoned that if capacity limits in CS are to be accessible in an introspective task, they should at least generate a bottleneck, and thus impair objective detection of an extra target. As opposed to that, in FS, the difference between one and two targets trials might be present right from the first sensory stage (Wolfe, 2003, 2007), and therefore it should be correctly detected. Additionally, if FS are done in a non capacity limited mode, performance should be independent from set-size and from the number of targets, while this would not be the case in CS. Failure in any of these predictions would suggest that the absence of introspection we found in Experiment 1 is a faithful introspection of an absence.

Material and methods

Participants

Seventeen normal adults, French speakers (12 women), aged between 20 and 32 (mean age: 23.8 years, *SD*: 3.1) participated in the study.

Stimuli and Procedure

Visual properties of the stimuli did not differ from those in Experiment 1. With respect to the procedure, in half of the trials one or two identical targets could be presented («L», «L L», «X» or «X X», equal proportions), for the other half only distractors were presented («Ts»). When two targets were presented, both were randomly positioned on the stimuli imaginary circle, with at least one distractor between these when the set-size was higher than 2. Set-size, fixation and stimulus durations were identical to those of Experiment 1. Participants were asked about the presence or absence of at least one target («X» or «L»). Then, on 70% of the target present trials, participants were instructed to estimate the number of targets in the scene (or Identification of the Number of Targets, INT). Participants used the "U", "I" and "O" keys with the index, middle and ring fingers of the right hand to report 0, 1 and 2 targets. The experiment consisted of 10 blocks of 80 trials with a 60 second pause

between each block, totaling 400 target present trials, and among them 280 trials with a forced choice estimate of the number of targets. A similar training to the one of Experiment 1 was administered before the main experimental blocks.

Results

As in the previous experiment, median RTs and mean error rate presented a positive and significant correlation across present target trials ($r^2(270) = .05$, $\beta = .24$, $t = 3.98$, $p < .001$), therefore, they were transformed into inverse efficiency scores (IES). Before this transformation, we excluded trials with response times below 200 *ms* and trials with response times 3 *SD* above the median (2.4%) and RTs were log-transformed to approximate normal distribution.

On the detection response, we found a pattern similar to the one of Experiment 1. Namely, search efficiency decreased as a function of set-size in CS but not in FS (see Figure 4A). However, this interaction seemed modulated by the number of targets presented, to the effect that search efficiency for difficult target was less impacted by the set-size when there were two targets. To assess this pattern statistically, we ran an LMM on IES with fixed factors of set-size, number of targets and search type, as well as all possible interactions between these. We tested this model on target present trials. The triple interaction was significant ($F(1,248) = 6.27$, $p < .05$). We also found that the interactions between set-size and search type were significant with one and two targets (one target: $F(1,116) = 18.45$, $p < .001$; two targets: $F(1,116) = 5.96$, $p < .05$). Furthermore, with one as well as two targets, we found a non-significant slope in the FS search condition (one target: $p > .64$; two targets: $p > .92$), while it was significantly positive in the CS condition with one target ($F(1,17.5) = 7.00$, $\beta = 1.19$, $p < .05$) and marginally significant with two targets ($F(1,25.5) = 3.42$, $\beta = .37$, $p = .07$). Finally, we found a main effect of the number of targets in the CS condition so that performance was higher with two targets than with one ($F(1,118) = 11.11$, $p < .001$). In contrast the number of targets had no impact on performance in FS ($p > .10$). In sum, the number of targets facilitated search in the CS condition, but not in the FS condition.

Regarding the number of targets identification (INT), the pattern of results (see Figure 4B) exhibited a triple interaction to the effect that, in the one target condition, increased set-size lead to reports of illusory targets in

both conditions, while in the two targets conditions, we observed a sharp opposition of search types: in FS participants did report seeing both targets, but not in CS.

When we applied a LMM on correct trials with mean INT as dependent variable, we found that the triple interaction between the number of targets, the search type and the set-size factors was significant ($F(1,214.3) = 3.65$, $p < .05$). In one target trials, we only found a main effect of set-size ($F(1,25.8) = 11.90$, $p < .01$), corresponding to the illusory increase of perceived targets, without a significant difference between the search types ($p > .35$), and no interaction ($p > .85$). In contrast, in two target trials, we found a significant main effect of search type ($F(1,94.2) = 17.52$, $p < .001$), and no effect of the set-size factor ($p > .31$). The interaction was also significant, ($F(1,94.2) = 9.61$, $p < .01$): in the CS condition, the number of reported targets decreased with set-size ($F(1,77.8) = 8.43$, $\beta = -.02$, $p < .01$), while the slope was not significant in FS ($p > .88$).

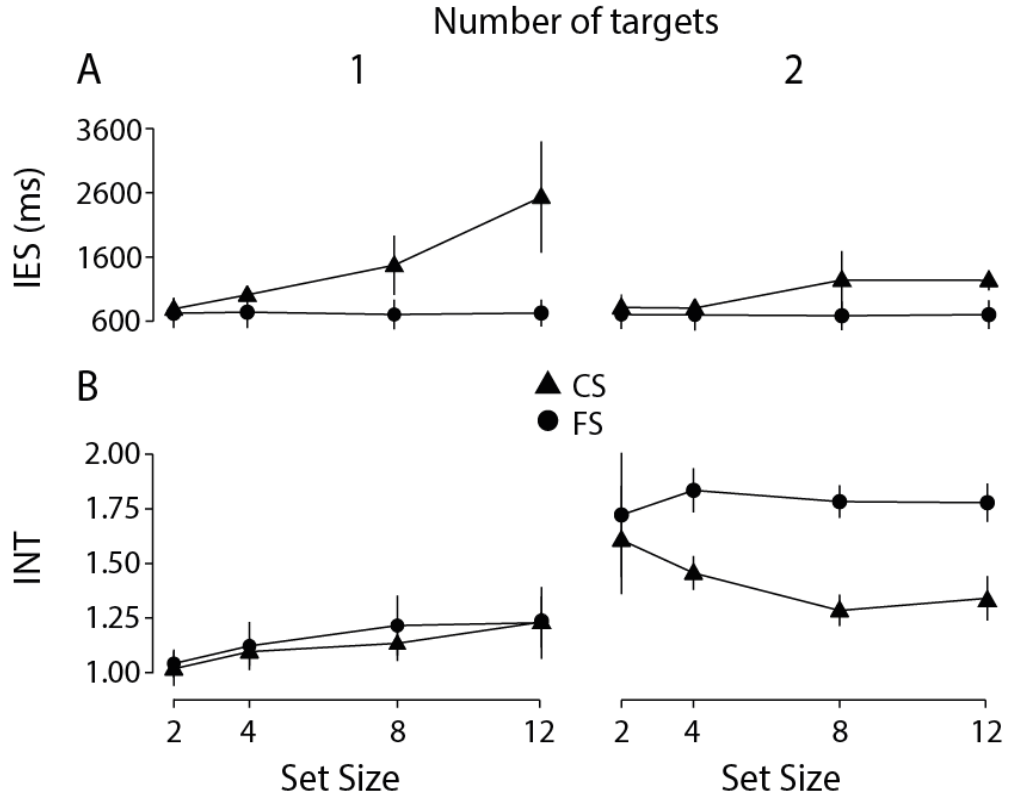


Figure 4. A) Inverse efficiency scores (IES) and B) Identification of the Number of Targets (INT) as a function of set-size, and for each number of targets and search condition, in Experiment 2.

Discussion

We confirmed that our FS induces an efficient search process, while our CS induces an inefficient process. Search is so efficient in FS that a second target does not improve performance, while it does in the inefficient CS condition.

Our main interest was in the secondary task, which was a 3 alternative forced choice decision on the number of targets perceived. The logic of the two targets trials was that if search in CS is capacity limited, then participants should miss more second targets in CS, as an optimal strategy should be to stop the search as soon as they have detected the first one. Results clearly confirmed this prediction: while IES considerably improved with a second target in CS, participants missed the second target in this condition, and did so increasingly as set-size increased. One plausible interpretation is that performance with 2 targets improves in CS because the probability of identifying the first target on the scene increases, thus, the visibility of a second target decreases.

The illusory increase of perceived targets with set-size when only one target is presented mirrors the increase of SNSI with set-size in the first experiment. Set-size might be a variable that is accessed very early in the search process, with increasing set-sizes, perceptual uncertainty on individual items will increase. Thus, identification of individual items might depend more on expectations (de Gardelle, Sackur, & Kouider, 2009). This would translate, in this Experiment into the increase of hallucinated second targets, and in Experiment 1, into the introspective increase in perceptual load.

In conclusion, Experiment 2 shows that our tasks generate capacity limits to which behavioral measures are sensitive. Thus, the question of whether these limits are analogously accessible to introspection is meaningful.

We now discuss the possibility that the lack of introspection for capacity limitations that we found in Experiment 1, should be specific to the implementation of the task. The aspect that might have had a decisive impact on our participants' subjective reports is the short presentation time (200 *ms*). Indeed, Bergen and Julesz (1983) suggest that a short presentation time only favors feature searches. Time pressure in Experiment 1 may have created a bias on the first stages of the visual search process, *before* attentional guidance could comes into play. Current integrated models of visual search (Wolfe, 2003, 2007; Wolfe & Horowitz, 2004) distinguish a pre-attentional

stage during which a target-like signal is extracted in parallel over the scene, and a second stage of target selection, during which attention is guided, according to the signal extracted during the first stage. In fast searches, an optimal strategy, minimizing time spent in the experiment while maintaining high performance, would be to respond quickly, on the basis of the signal extracted during the first stage. This would explain both the impact of the perceptual load in introspection and the absence of introspection of capacity limitations, as the favored strategy would be biased towards the perceptual parallel stage.

We reasoned that in order to render capacity limitation accessible, we needed to allow more time for the search and to force completion of the search. Participants should be forced not only to decide on the target presence, something they can do on average with some reliability on the basis of the information extracted during the first stage. Participants should be required to identify the target, something they cannot do until it has been put under attentional focus. To this end, we required that participants report a feature of the target orthogonal to its defining feature.

2.1.4.3 Experiment 3

In this experiment, we tested whether more favorable conditions would enable participants to introspect on capacity limitations. Compared to Experiment 1, we introduced the following modifications: 1 - the first order task was to report the target's color in an array of randomly colored items; 2 - the stimulus array was presented until participants responded, so as to discourage fast guesses based on incomplete processing; 3 - we added categories on top of the continuous quantitative SNSI scale, a procedure inspired by the Perceptual Awareness Scale (RamsØy & Overgaard, 2004).

Material and methods

Participants

Twenty-one normal adults, French speakers (18 women), aged between 19 and 28 (mean age: 21.3 years, *SD*: 2.1) participated in the study.

Stimuli and Procedure

The stimuli (see Figure 5) consisted of a set of red (luminance: 0.44 cd/m²) and green letters (luminance: 0.33 cd/m²) presented on an imaginary circle around a central fixation (radius: 6.2°). All the trials presented targets ("X" or "L"). Set-size (4, 8 or 16 items) and the target and distractors ("Ts") orientation (0, 90, 180, 270°) were randomized across trials. Stimuli were equally spaced on the imaginary circle.

Participants were instructed to decide whether the target presented was red ("Z" key) or green ("A" key). Stimuli were presented until participants responded. The SNSI scale was presented immediately after response. Under the scale four qualitative categories were specified (in French): "no item", "some items", "many items" and "all items". Each participant performed 480 trials (8 blocks of 60 trials) with a 60 second pause between blocks. A raining phase similar to Experiment 1 was included. Participants were instructed to avoid fast or automatic responses, and they were told that the categories on the scale were to be used as anchors for their subjective estimations, but that they should use all positions on the scale to report their best subjective estimate.

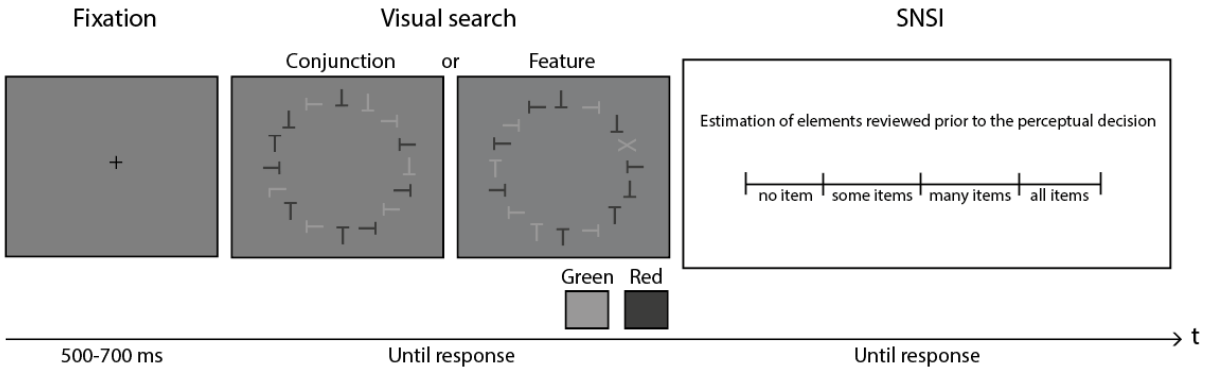


Figure 5. General structure of the task in Experiment 3. After presentation of the fixation cross, participants had to identify, without the time pressure, the color (red or green) of the target. All the trials contained one target, either an X or an L. Immediately after the perceptual decision, participants were requested to estimate the number of items scanned on a qualitatively labeled scale (SNSI).

Results

First order task

We excluded trials with response times below 200 *ms* and trials with response times 3 *SD* above the median (4%). Given the low (3.8%) percentage of errors in this experiment we restricted our analyses to correct trials.

As in Experiment 1, we observed the expected interaction, reflecting the opposition of the capacity limited searches for CS and non capacity limited for FS (see Figure 6A). Indeed, an LMM on median correct RTs (log-transformed) revealed a significant main effect for search type ($F(1,88.7) = 269.6$, $p < .001$) and set-size ($F(1,124.2) = 95.2$, $p < .001$), and a significant interaction between these factors ($F(1,88.7) = 69.71$, $p < .001$). The CS condition lead to a significant increase in response time as a function of set-size ($F(1,41.0) = 307.6$, $\beta = .03$, $p < .001$), and a marginal one in the FS condition ($F(1,20.2) = 5.71$, $\beta = .004$, $p = .051$).

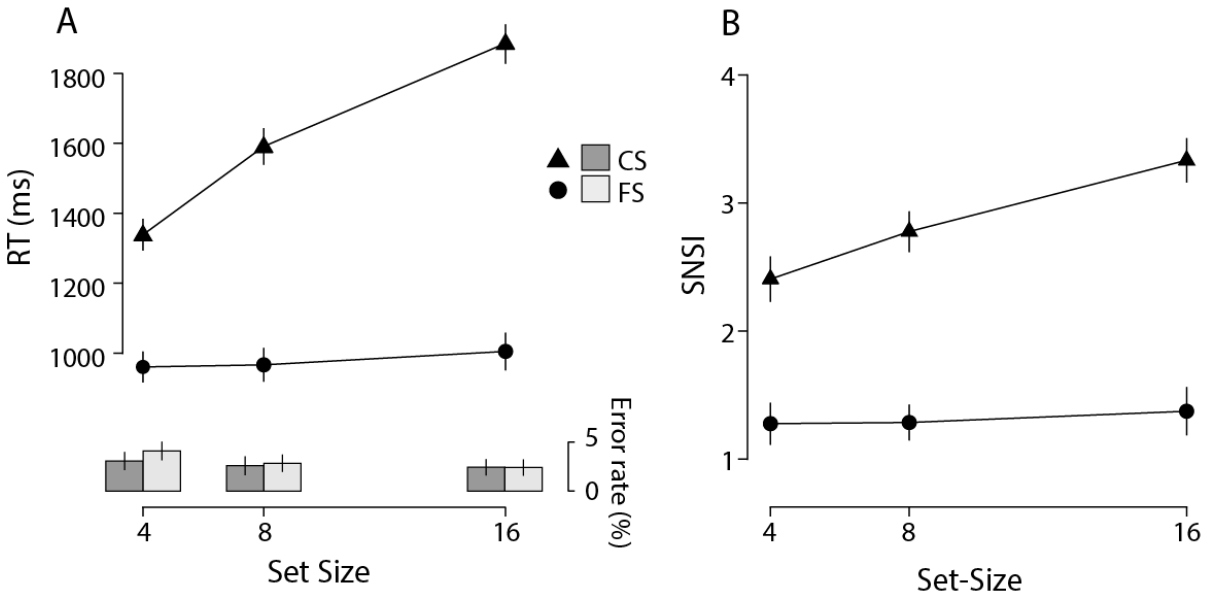


Figure 6. A) Response time and Error rate as a function of set-size in both search conditions, in Experiment 3. B) SNSI and as a function of the same search conditions, in Experiment 3.

Second order task

SNSI responses parallel response times (see Figure 6B). A similar LMM performed on mean SNSI revealed a significant main effect of search type ($F(1,99.9) = 125.5$, $p < .001$) and set-size ($F(1,115.0) = 33.03$, $p < .001$). The interaction between these factors was also significant ($F(1,99.9) = 22.63$, $p < .001$). Critically, set-size significantly impacted participants SNSI in CS ($F(1,20.0) = 92.98$, $\beta = .08$, $p < .001$), but not in FS ($p > .31$).

Contrary to what we had found in Experiment 1, here, participants were able to introspect on the difference in search process generated by the FS and CS stimuli, as SNSI results parallel first order results. If we interpret first-order results as evidence for capacity limitations in CS, it seems that participants are subjectively aware of such limitations. However, these results may very well derive from inferences based on non introspective sources of information. As discussed in the Introduction, participants might have modulated, on a trial-by-trial basis, their introspective estimation on the basis of their RTs, producing confabulatory introspective reports. To assess to what extent SNSI was contaminated by self-observation of response times, we used mediation analyses (Bauer, Preacher & Gil, 2006). The strategy consists in testing whether the impact of search type and set-size on SNSI disappears when RTs are controlled. We focus our analysis on the interaction term between search type

and set-size, given that this effect is diagnostic of capacity limitations. Disappearance of this effect would suggest that introspected capacity limitation is in fact due to self-observation of RT. Along these lines, we estimated the *total effect* (the impact of the interaction term on SNSI), the *indirect effect* (the size of the interaction effect explained by RT, i.e., the mediator variable) and the *direct effect* (the difference between the total and indirect effect, which denotes the impact of the independent variables on SNSI, not mediated by changes in RT). As in previous analyses, we considered that all these effects can vary randomly between participants. Our mediation model has thus two levels: both the outcome (SNSI), the predictor (interaction between set-size and search type) as well as the potentially mediating variable (RT) constitute the first level, which are nested within each participant (i.e., second level). To test the significance of the indirect effect, we used a Monte Carlo confidence interval method (Selig & Preacher, 2008; Preacher & Selig, 2012). We refer the reader to the Appendix for the details of the model.

In agreement with the previous analysis, we found a significant *total effect* of the interaction term on SNSI ($F(1,19.9) = 194.0$, $c = .02$, $p < .001$). We also found that the interaction term impacted RTs ($F(1,19.8) = 489.5$, $\alpha = .04$, $p < .001$). In addition, controlling the interaction effect on SNSI, RT presented a significant relationship with SNSI ($F(1,20.1) = 171.3$, $\beta = .29$, $p < .001$), which is required for RTs to be considered as potential mediator. Finally, after controlling the RT effect on SNSI, we found that the impact of the interaction on SNSI was reduced (*direct effect*: $F(1,19.9) = 39.34$, $c' = .007$, $p < .001$). Thus, we found a partial mediation of SNSI by RTs: the size of the *indirect effect* was .013 (C.I. [.011, .016]). In other terms, 65% of the effect of the interaction term on SNSI was mediated by RTs.

Discussion

In this experiment, we again observed the contrast between feature searches (FS) and conjunction searches (CS): increasing set-sizes gave rise to a significantly steeper RT slope in CS than in FS. Errors were not informative, which is a consequence of the search array being presented until participants' responses. Results on the SNSI scale suggest that participants' introspection not only showed a global difference between the search conditions. Importantly, SNSI increased as a function of the number of distractors in the array only in CS. Furthermore, the impact of the interaction between set-size and search type on the subjective number of scanned items was only

partially mediated by response times. Thus, capacity limits are introspectively accessible by pure mental monitoring, provided that the context of the task makes them sufficiently salient.

Even though the introspective task was identical in Experiments 1 and 3, the context influenced how it was performed: in Experiment 1, high speed demands favored introspection based on the first perceptual stage, and we found only an effect of perceptual load in introspection. Here, the demands of the task shifted towards the second, target identification stage, and participants' introspection followed. We suggest that when task contexts vary, introspection flexibly adapts to different aspects of the same cognitive processes.

Our mediation analyses showed that variability in response times have a major impact on introspection. We can speculate as to how this comes about: first, it may happen through a contaminating bias, i.e., introspective response times were automatically computed, in parallel with the pure SNSI, and then biased responses. Second, response times may impact the estimates on the SNSI scale through confabulation, i.e., if participants have a theory about the link between response duration and the number of scanned items during the decision process. Third, there might be a real introspective process link; without any explicit theory, participants' spontaneous introspective task setting might use diverse sources of information, including decision duration and self-observation of response times.

In the next experiments, in order to control and possibly factor out the use of self-observation in introspection, we used a fixed stimulus presentation time and a late response window (Experiment 4a and 4b), and we recorded (Exp. 4a) or controlled (Exp. 4b) eye movements.

2.1.4.4 Experiment 4a and 4b

In these experiments we seek to better understand the nature of the information used by participants' introspection. For this purpose, we kept the same first-order stimuli and second order task as in Experiment 3. We only introduced a fixed stimulus duration and a response window: responses could only be produced during a 1000 *ms* response window that began immediately after a fixed 3000 *ms* stimulus presentation. In addition, we recorded gaze position during stimulus presentation so as to include eye-movements as possible mediators in the analysis of introspective responses. The only difference between Experiment 4a and b, was that in the former eye movements were allowed but not in the latter.

Material and methods

Participants, Stimuli and Procedure

In Experiment 4a, eighteen normal adults, French speakers (9 women), aged between 18 and 28 (mean age: 22.8 years, *SD*: 2.7) participated. Each participant performed 288 trials (8 blocks of 36 trials) with a 60 second pause between each block. A similar training to the one of Experiment 1 was administered before the main experimental blocks. Eye movements were recorded monocularly with an eye tracker (EyeLink 1000 system, SR Research, Canada), with a sampling rate of 500 Hz and a spatial accuracy better than 0.5° (Camera-Eye distance: ~55 cm). Saccades were determined using a conservative algorithm (velocity threshold: 30°/s, acceleration threshold: 8000°/s2, motion threshold: 0.15°). For all participants the right eye was recorded. Stimuli were identical to those of Experiment 3, except that their duration was fixed at 3000 *ms*. Participants could only respond during a 1000 *ms* window beginning at stimulus offset. A recalibration procedure for the eye-tracker was conducted before each block.

In Experiment 4b, fifteen normal adults, French speakers (11 women), aged between 19 and 29 (mean age: 22.7 years, *SD*: 2.5) participated. The stimuli and procedure did not differ from those of Experiment 4a, except that in this study, participants were requested to fixate on the cross at the center of the stimuli during the entire 3000 *ms* presentation time (blinking was allowed). An invisible circle (radius: 3.0°) around fixation determined the degrees of freedom of eye movement: participants were told that if their gaze moved away from the fixation

cross, the trial would be considered incorrect, and the next trial would begin immediately. During the training period, participants were trained to suppress eye movement during the presentation of the stimuli.

Results

Experiment 4a: First order task

Given the low percentage of errors in this experiment (2%), they were not analyzed. Although the response window greatly sped-up motor responses (see Figure 7A, dashed lines), they still showed a pattern analogous to the one of Experiment 3. An LMM was run on median log-transformed RTs. All main effects and the interaction were significant (search type: $F(1,69.8) = 15.59$, $p < .001$; set-size: $F(1,17.2) = 33.56$, $p < .001$; interaction $F(1,69.8) = 14.88$, $p < .001$). As in Experiment 3, we found a significant and steeper slope for CS ($F(1,17.7) = 38.40$, $\beta = .013$, $p < .001$) than for FS ($F(1,17.3) = 10.28$, $\beta = .005$, $p < .01$).

Next, we analyzed the latency of the first fixation on the target (i.e., First Target Fixation Latency, FTFL). We defined a square window around the target (0.8° x 0.8°), and measured the latency with respect to the first fixation of at least 50 *ms* within this window. As shown in Figure 7A, the latency of the first fixation on the target mirrors the interaction pattern previously found on response times. We ran an LMM on median FTFL (log-transformed) within correct trials. Again, the main effects and the interaction were significant (search type: $F(1,75.4) = 103.7$, $p < .001$; set size factor: $F(1,24.0) = 131.8$, $p < .001$; interaction: $F(1,75.4) = 44.22$, $p < .001$). The CS condition showed a steeper slope as a function of set-size ($F(1,28.1) = 115.8$, $\beta = .05$, $p < .001$) than for FS ($F(1,19.2) = 31.88$, $\beta = .01$, $p < .001$). In addition, as shown in Figure 7B (light grey bars), the number of saccades (log-transformed) follows a similar patter; the two main effects and the interaction term were significant (search type: $F(1,72.3) = 56.63$, $p < .001$; set-size: $F(1,13.8) = 114.9$, $p < .001$; interaction: $F(1,72.3) = 23.27$, $p < .001$; CS: $F(1,35.0) = 135.3$, $\beta = .05$, $p < .001$; FS: $F(1,35.0) = 23.45$, $\beta = .02$, $p < .001$). Finally, we also analyzed mean saccade amplitudes (log-transformed), during the same time window, with a similar LMM. We observed again that the two main effects and the interaction term were significant (search type: $F(1,67.8) = 10.14$, $p < .01$; set-size: $F(1,20.8) = 109.8$, $p < .001$; interaction: $F(1,67.8) = 4.00$, $p = .051$). A more detailed examination indicated that both in CS ($F(1,22.2) = 116.5$, $\beta = -.02$, $p < .001$), as well as in FS ($F(1,35.0) = 39.71$, $\beta = -.01$, $p < .001$), the saccade amplitude decreases as a function of set-size (see light grey bars in Figure 7C). This decrease may be due to the fact that the radius of the imaginary circle for stimuli is constant. Consequently,

with small set-sizes, participants' search will involve greater amplitude eye movements, because stimuli are farther apart.

Experiment 4b: First order task

One participant was excluded from the analyses because he had unusually high error rates (>50%). In this experiment 8% of the trials were excluded from the analysis because eye movements exceeded the acceptable fixation zone.

As previously, RTs exhibited the typical visual search interaction (see Figure 7D) and this was confirmed by an LMM on median correct (log-transformed) RTs (set-size: $F(1,14.5) = 29.35$, $p < .001$; search type: $p > .25$; interaction $F(1,57.6) = 24.99$, $p < .001$). We also found a significant increase of RTs as a function of set-size in CS ($F(1,14.5) = 28.64$, $\beta = .02$, $p < .001$), but not in FS ($p > .10$). In the similar LMM on error rate (arcsine transformed) we found the same significant effects (set-size: $F(1,24.0) = 24.24$, $p < .001$; search type: $p > .64$; interaction: $F(1,72.0) = 33.34$, $p < .001$), which was characterized by a higher increase in CS ($F(1,18.3) = 24.66$, $\beta = .01$, $p < .001$) than for FS ($F(1,18.2) = 5.88$, $\beta = .001$, $p = .05$).

As expected, the instruction to fixate introduced a drastic change in eye-movements (see Figure 7B & 7C). No significant effects of the experimental variables were found on the number of saccades nor on the saccade amplitude (all ps > .10).

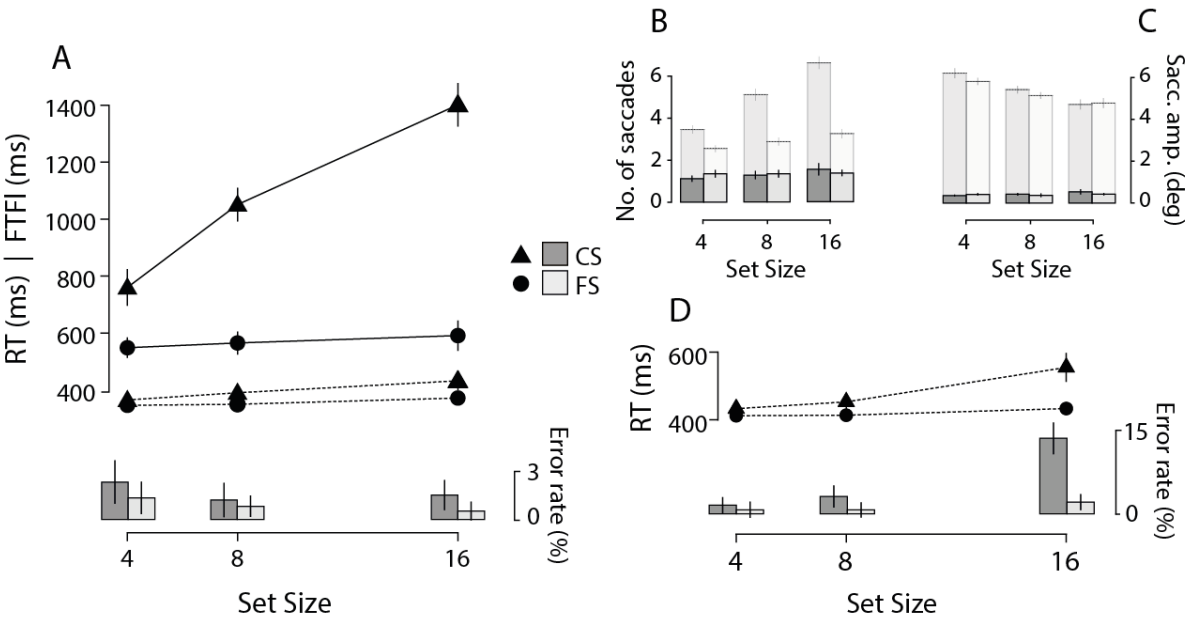


Figure 7. A) Plain lines indicate First Target Fixation Latency (FTFL) and Error rate as a function of both search conditions. Dashed lines indicate RT as a function of the same conditions, in Experiment 4a. B) Number of saccades and C) saccade amplitude as a function of the same conditions, in Experiment 4a (light grey bars behind each graph) and in Experiment 4b (solid grey bars). D) RT and Error rate as a function of set-size in both search conditions, in Experiment 4b.

Experiment 4a: Second order task

As shown in Figure 8, participants' introspection depended on the two factors of set-size and search type. In an LMM performed on mean SNSI, we found that both main effects, as well as the interaction were significant (search type: $F(1,70.0) = 89.17$, $p < .001$; set-size: $F(1.17.0) = 21.13$, $p < .001$; interaction: $F(1,70.0) = 22.54$, $p < .001$). A more detailed look at the interaction revealed that SNSI increased as a function of set-size in CS ($F(1,35.0) = 53.92$, $\beta = .10$, $p < .001$), but not in FS ($p > .60$).

Then, we investigated the possible contamination of this interaction effect by self-observation. First, we ran a multilevel mediation model with RTs as mediator. The analysis showed a significant interaction between set-size and search type on SNSI ($F(1,16.9) = 260.5$, $c = .06$, $p < .001$) and on RTs ($F(1,17.3) = 128.3$, $\alpha = .01$, $p < .001$). Second, controlling the interaction effect on SNSI, we found a significant RTs impact on SNSI ($F(1,13.8)$

= 54.35, β = .32, p < .001). Finally, when the RT impact on SNSI was controlled, the interaction effect was marginally reduced (F(1,381.4) = 520.7, c' = .05, p < .001, indirect effect: .005, C.I. [.003, .006]; 8% contaminated). Following this logic, we found that the interaction between set-size and search type had a significant impact on first target fixation latency (FTFL) (F(1,17.3) = 387.3, α = .06, p < .001). At the same time, FTFL presented a significant relationship with SNSI, after controlling the interaction term (F(1,17.1) = 85.89, β = .67, p < .001). Then, after controlling FTFL, we observed that the interaction impact on SNSI was only partially mediated by the latency of the first fixation on the target (F(1,379.9) = 127.6, c' = .01, p < .001, indirect effect: .04, C.I. [.037, .045]; 66% contaminated).

Moreover, we found a significant interaction effect between set-size and search type on the number of saccades (F(1,17.1) = 542.5, α = .06, p < .001), a significant impact of the number of saccades on SNSI, after controlling the interaction term (F(1,17.2) = 93.26, β = .17, p < .001) and a significant interaction effect on SNSI, after controlling the number of saccades effect on SNSI (F(1,17.8) = 35.01, c' = .05, p < .001). These results suggest a marginal mediation effect of the number of saccades (indirect effect: .01, C.I. [.009, .014]; 16% contaminated). Finally, we found a significant interaction effect between set-size and search type on the saccades amplitude (F(1,16.6) = 37.91, α = -.01, p < .001). Then, after controlling this effect, we observed a significant impact of the saccades amplitude on SNSI (F(1,17.8) = 55.40, β = .16, p < .001) and a significant interaction effect on SNSI, after controlling the mediator (F(1,16.4) = 261.5, c' = .06, p < .001). However, the indirect effect was not significant: -.002, C.I. [-.008, .008]).

Experiment 4b: Second order task

Participants' introspection presented the same pattern as in the previous experiment. A similar LMM on mean SNSI showed that both main effects, as well as the interaction, were significant (search type: F(1,55.3) = 18.05, p < .001; set-size: F(1,14.8) = 33.19, p < .001; interaction: F(1,55.3) = 32.81, p < .001).This interaction was characterized by a significant SNSI increase as a function of set-size in CS (F(1,14.2) = 47.21, β = .19, p < .001), but not in FS (p > .10, see Figure 8). Then, we evaluated whether this interaction effect was mediated by RTs or eye-movements, even though both were restricted in this experiment. Multilevel mediation models showed that the interaction between set-size and search type presented a significant effect on SNSI (F(1,13.2) = 67.74, c =

.03, $p < .001$) and on RT ($F(1,15.0) = 49.72$, $\alpha = .01$, $p < .001$). After controlling the interaction effect, we found a significant RT / SNSI relationship ($F(1,14.1) = 31.38$, $\beta = .10$, $p < .001$). Finally, controlling this RT effect, the interaction effect on SNSI was only marginally reduced ($F(1,13.1) = 63.24$, $c' = .03$, $p < .001$, indirect effect: .0018, C.I. [.001, .002]; 6% mediated). The same model ran on the number of saccades ($\alpha = .007$, $p > .53$) and on the saccade amplitude ($\alpha = .006$, $p > .13$), confirmed that none of these variable presented a significant relationship with the interaction term.

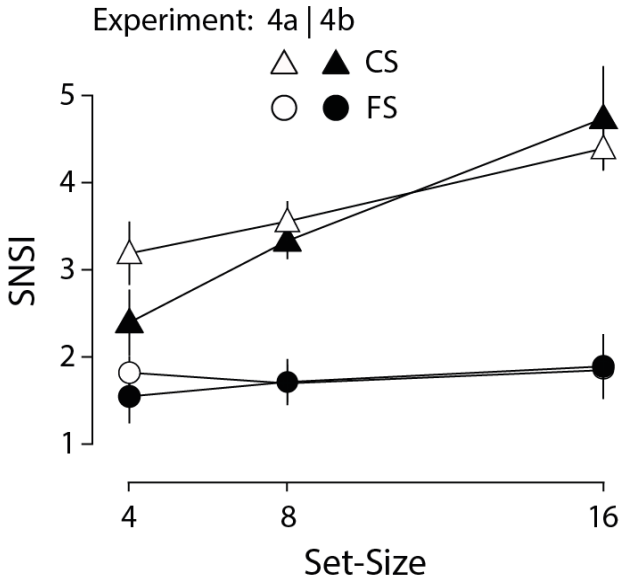


Figure 8. SNSI as a function of set-size in both search conditions in Experiment 4a (white symbols) and in Experiment 4b (black symbols).

Discussion

In Experiments 4a, we tried to factor out response times, so as to assay whether participants could still do the introspective task without access to this behavioral information. Our use of a response window had a drastic influence on response times, without totally eliminating the information they carry about the search processes, as evidenced by the fact that we still find a pattern of response times characteristic of capacity limited and unlimited searches. Eye-movement results agree with the literature: the number of saccades was higher in CS

than in FS, increasing with greater intensity in CS as a function of the number of distractors in the scene. Most importantly, the latency of the first fixation on the target exhibited the same interaction pattern as response times.

Our second order measure also showed the interaction between search type and set-size factor, that we interpret as a sign of introspective access to capacity limitations. Thus, even though we managed to greatly diminish the saliency of behavioral responses with the response window, this manipulation did not drastically modify participants' subjective estimate, confirming the robustness of our results and suggesting that introspection can diagnose differences between the search conditions. In line with this result, we found that the mediating role of response times with respect to SNSI is now greatly reduced although not absent. This suggests that self-observation of overt response behavior is not a necessary source of information for the form of introspection that we elicit from participants.

In Experiment 4b, we controlled eye movement during the first order task, because Experiment 4a demonstrated that eye movements were a potential source of self-observation as they acted as mediating variables. We hypothesized that if participants' introspection, as reported in Experiment 3 and 4a, is not solely due to eye movement, we should observe the same SNSI pattern if we controlled them. This was indeed the case. Thus, it seems that introspection is still possible when self-observation both of manual response times and eye-movements are not available.

2.1.4.5 Conclusion

In this article, we investigated whether one could have introspective access to cognitive processes, as opposed to introspective access to the cognitive states or behavioral consequences that they generate. To do so, we used two visual searches as our test bed: a difficult, capacity limited search, and an easy, non-capacity limited search. We thus endeavored to test Nisbett and Wilson's (1977) thesis about the limits of introspection, within the context of elementary cognitive processes. We devised an introspective task such that participants had to report how many items they felt they had scanned during the search process (the Subjective Number of Scanned Items, SNSI), which we took as an index of the subjective access to capacity limitation, or an inverse index of the strength of attentional guidance.

Two broad classes of results emerge from the series of four experiments we present, and they all point to flexibility and contextual modulation of introspection. The first class of results stems from the contrast between Experiment 1 and the four other experiments. In Exp. 1, we discovered that participants' reports of SNSI were sensitive only to the overall number of items in the search array, without much hint at sensitivity to the difference between capacity limited and unlimited processes. This negative result was all the more notable than in the same experiment, the two other second order tasks we used (i.e., confidence and introspective response times) did show good metacognitive sensitivity. As opposed to this null result, in Exp. 3, 4a and 4b we found clear evidence of behavioral and introspective access to the difference in capacity limitations. The critical distinctions between the two sets of experiments were the short presentation time of stimuli in Exp. 1 (200 *ms*) compared to the long presentation (unlimited / 3000 *ms*) in the others, on the one hand, and the fact that the search task demanded *identification* of the target in the last group of experiments, as opposed to simple *detection* in the first experiment, on the other hand.

The null result of Exp. 1 with respect to introspection of capacity limitation has, first, an important methodological import: it means that the SNSI task we rely on is not trivially contaminated by confabulation. It is not the case that participants report that they scanned more items when the task is more difficult, or even, when they search for an L among Ts as opposed to an X among Ts, because they rely on a theory that the former must require extensive scanning. In fact, as results on confidence ratings and subjective duration of the task showed in Exp. 1, participants clearly introspected at the trial level that capacity limited searches were more

difficult and took longer to perform than capacity unlimited searches. But that did not translate into an increase of the number of items they subjectively felt they had scanned. As further experiments showed, on the contrary, that the SNSI task can be sensitive to capacity limitations, the null result of Exp. 1 must be interpreted as a sign of introspective inaccessibility. To make sure that there was indeed a genuine difference between our two searches, in other words, that the pattern of behavioral first-order results was not simply mimicking differences in capacity limitation, we used a non-introspective procedure in Exp. 2, which demonstrated that our two searches did create two qualitatively different search processes. In brief, something was available to introspection in Exp. 1, but it was not accessed.

We now interpret this result in the light of recent models of visual search (Wolfe, 1994, 2003, 2007; Wolfe, et al., 2011) that distinguish two stages: a first parallel stage consisting of extraction of visual features, and a second, possibly guided, stage of target selection. We hypothesized that when the stimuli are presented for a brief duration, the search process is imbalanced in favor of the first stage; so that responses are generated mostly on the basis of the information extracted during the first parallel pass. A rational cost / benefit analysis of optimal behavior, in the sense of maximizing correct responses while minimizing time in the experimental booth, might show that in such circumstances it is best not to commit too much resources in guided target selection, as this would lengthened each decision without much benefit to performance. In this situation, the decision variable simply integrates the information available after feature extraction over the entire display. On the contrary, in the latter group of experiments, both modifications concur to shifting the optimal behavior towards slower, possibly guided searches; as the display is shown for a longer time, it is beneficial to spend more time in the search. In addition, as the task requires target identification, the cost of not finishing the search would be disproportionate.

According to this speculative hypothesis, the controversy between pure signal detection models of visual search (Carrasco & Yeshurun, 1998; Carrasco & McElree, 2001; Cameron, Tai, Eckstein, & Carrasco, 2004) and guidance models might be more a matter of relative weighting of search subprocesses according to task context, than a question about the essence of visual search. Furthermore, we suggest that introspection tracks the imbalance of these sub-processes: when the first pass dominates, the subjective number of scanned items corresponds to the complexity of the scene. When the context of the task renders the second, selection stage critical to optimal performance, then it contributes to introspection, and participants are subjectively aware of the

presence or absence of guidance in the search. This takes precedence on whatever subjective salience the complexity of the scene could have had.

Critically, what we hypothesized as an imbalance in the search processes is not marked in the pattern of response times which, in each and every of our 4 experiments, exhibited the traditional interaction of set size and search type factors. However, we suggest that this surface similarity across the first and latter experiments hides processing differences that introspection is able to reveal. The notion of "model mimicking", familiar from the literature on serial *versus* parallel processes (Townsend & Wenger, 2004), is precisely meant to capture the fact that this interaction, which could easily be taken as diagnostic of the opposition of limited and unlimited processes, is in fact a *non sequitur*. Here we argue that in our Exp. 3 the interaction is indeed a sign of the opposition of two types of processes, whereas in Exp. 1 it is not. We base this conclusion on the absence of any introspective difference between the search processes in Exp. 1, as opposed to clear differences in the last experiments.

If the above reasoning is correct, the increase in SNSI reports with set-size in Experiment 1 corresponds to the increasing perceptual load, caused by information accrual during the first stage of feature extraction, while in the latter experiments it corresponds to an introspective access to capacity limitation. Of course, this interpretation raises the difficulty that the very same instruction to introspect is in fact ambiguous and corresponds to different internal targets of introspection. It is important to note that we did not change the wording of the instruction for the SNSI across experiments.

We should note here that this ambivalence of instructions with respect to introspection is in fact not an accident but an essential feature of introspection (Jack & Roepstorff, 2002), as there is by definition no external fact of the matter to which performance can be aligned. However, our study demonstrates that this ambivalence can be tamed, so that introspective data can be used both, on the one hand, with a view to complementing basic behavioral responses in order to better understand cognitive processes, and on the other hand, in order to understand the process of introspection itself.

The second class of results concerns the purity of the introspective judgments. In Experiments 4a and 4b, we successfully isolated reports on the number of internally scanned items from two major potential contaminants,

namely response times and eye-movements. We showed, using both experimental controls and statistical analyses, that introspection on the target selection stage is not a construction based on other informational sources that are already known to be accessible to self-observation, e.g., response times (Corallo, et al., 2008; Marti, et al., 2010).

It seems that we succeeded in eliciting pure mental process monitoring from our participants. Miller et al. (2010) used a simple go / no-go tasks to pure mental monitoring of decision time, and concluded that "decision time reports are not very accurate but they may be usable for some purposes". To reach this conclusion they relied on manipulations of difficulty in the primary task, which is supposed to influence internal decision time, and on manipulations of the complexity of the response, which by contrast is not supposed to impact decision time. However, note that the authors did not use a direct manipulation of the response time itself. Therefore, the purity of the subjective decision time reports is not beyond doubt, and reports may well be in part contaminated by self-observation of behavior. By contrast, we made sure that introspection of the cognitive processes in visual search derives from direct access to them, and does not build on inferences based on overt or covert behavior. The SNSI task is thus, to our knowledge one of the first clear instance of pure mental monitoring, as opposed to behavioral self-observation. Of course we must be cautious with respect to the selectivity of our introspective measure: rather than reporting the number of scanned items, participants may have reported their internal decision time. These two variables are of course highly correlated, and it is difficult to decide between them, as the only diagnostic feature might be whether the distribution of responses is discrete or continuous. Be that as it may, both cases are clear cases of pure mental monitoring, the possibility of which was the main question of our study.

Thus, contrary to the claims of Nisbett & Wilson (1977), we should state that introspection of mental processes, and not only of mental states, is possible. Of course, we should keep in mind the very specific conditions under which this is true: First, the first order task that is the target of introspection is elementary and short. Our results may not extend to more complex and longer tasks, where the latitude for confabulations may be higher. However, the results on the rehearsal literature (Montague, Hillix, Kiess & Harris, 1970; Kroll, Kellicutt & Parks, 1975) suggest that in some cases introspection could be reliable at longer time scales than ours. Second, we designed the first order task so as to maximize the salience of the attentional guidance in the target selection stage of our visual search. Third, introspection was performed systematically and immediately after the first order task, so that participants were trained to focus on the processes of interest, and could report their

introspection while it was still present in working memory. It is notable that all these conditions correspond to the recommendations of previous and contemporaries researchers on introspection (Titchener, 1899; Schooler, 2002).

We can tentatively synthesize our results in a descriptive model of introspection (see Figure 9). In this model we represent two dimensions that define the space within which introspection can flexibly be focused: first, the dimension that opposes mental monitoring and self-observation, and second the timing with respect to task processes. Previous results (Corallo et al., 2008; Marti et al., 2010) and the present ones suggest that participants can focus their introspection on data that are more or less objective. We suggest that there is a gradation with respect to the purity of introspection, with pure mental monitoring at one extreme and pure self-observation at the other. On the other dimension, we suggest that participants are able to introspectively focus on different stages of a given task process. This is evidence in our study by the contrast between Experiment 1 and the last two. We must also mention here the literature on error monitoring (Yeung & Summerfield, 2012) that opposes conflict monitoring (van Veen & Carter, 2002; Yeung, Botvinick, & Cohen, 2004) and post-decision processing (Petrusic & Baranski, 2003; Resulaj, Kiani, Wolpert, & Shadlen, 2009) as potential sources of error detection. This opposition can also be understood in terms of whether introspection is focused on early or late task processes.

This model can serve as a framework for further research on the mechanisms of introspection: if it is true that introspection can flexibly move within this task processes space, it is not self-evident that it could be divided, so that different portion of this space could be simultaneously monitored.

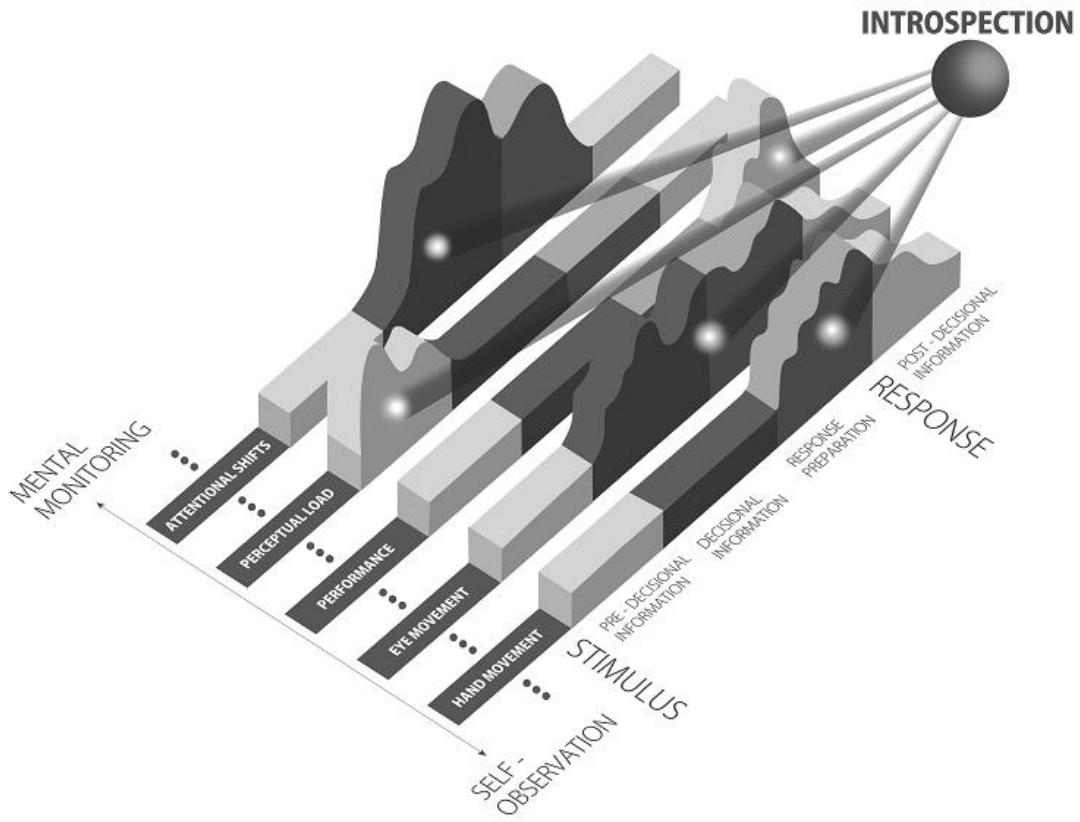


Figure 9. Schematic representation of the flexibility of introspection. The model represents the flexibility of introspection in two dimensions. The first dimension (mental monitoring - self observation) organizes sources of information on which introspection is focused. When an introspective index primarily uses behavioral information sources, introspection is conceptualized as a process of self-observation. By contrast, when the introspective source of information comes from the cognitive process, the index is properly conceptualized as mental monitoring. The second dimension (early *versus* late processes) specifies different stages, during the development of the first order task, at which mental states are available to introspection. Here, it is possible to distinguish an introspective *composite* index, which combines information on all the stages and a *pure* introspective index (Sternberg, 2001), where participants' introspection is focused on a specific stage.

These questions are critical with respect to the possibility of compound introspective tasks, an issue which is particularly important with respect to confidence. The bases of confidence judgments, which one may think as the most important introspective task, from a behavioral perspective, are so far unclear. Indeed, recent models of confidence (e.g., Ratcliff & Starns, 2009; Pleskac & Busemeyer, 2010) suppose that confidence in simple choices is driven by the rate of information accrual during the decision. Transposed within the present

framework, this would mean that confidence is the resultant of mental monitoring of the speed to reach the decision. An alternative hypothesis, which admittedly is so far purely speculative, would be that confidence is a compound metacognitive judgment, which might integrate various sources accessible to introspection.

Acknowledgments

We thank Sid Kouider, Nathan Faivre, Jaime R. Silva and Hielke Prins for useful discussions. We thank also Anne-Caroline Fiévet, Isabelle Brunet, Cécile Girard and Claire Mégnin for their help in data acquisition.

This work was supported by a doctoral fellowship from the National Commission for Scientific and Technological Research (CONICYT 72090838, Chile) to G.R., a grant from the Agence National de la Recherche (DYNAMIND ANR-10-BLAN-1902-01, France) to J.S. and by a grant from the Agence Nationale de la Recherche (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL) to G.R. and J.S.

References

Austen, E., & Enns, J. T. (2003). Change detection in an attended face depends on the expectation of the observer. *Journal of Vision, 3*, 64-74.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological Methods, 11*, 142-163.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience, 33*, 16657-16665.

Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature, 303*, 696-698.

Boring, E. G. (1953). A history of introspection. *Psychological Bulletin, 3*, 169-189.

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the Inverse Efficiency Score (IES) a better dependent variable than the mean Reaction Time (RT) and the Percentage of Errors (PE)? *Psychologica Belgica, 51*, 5-13.

Cameron, E. L., Tai, J. C., Eckstein, M. P., & Carrasco, M. (2004). Signal detection theory applied to three visual search tasks: Identification, Yes/No detection and localization. *Spatial Vision, 17*, 295-325.

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception & Performance, 24*, 673-692.

Carrasco, M., & McElree, B. (2001). Covert attention speeds the accrual of visual information. *Proceedings of the National Academy of Sciences, 98*, 5341-5436.

Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science, 19*, 1110-1117.

Costall, A. (2006). 'Introspectionism' and the mythical origins of scientific psychology. *Consciousness and Cognition, 15*, 634–654.

Cousineau, D. (2005).Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*, 42-45.

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences, 16*, 241-262.

de Gardelle, V., Sackur, J., & Kouider, S. (2009). Perceptual illusions in brief visual presentations. *Consciousness and Cognition, 18*, 569-577.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*, 1-36.

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

Feest, U. (2012). Introspection as a method and introspection as a feature of consciousness. *Inquiry, 55*, 1-16.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*, 1541-1543.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London – Series B, 367*, 1338-1349.

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

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

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus-response to script-report. *Trends in Cognition Sciences, 6*, 333-338.

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

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a "preattentive" feature search task. *Nature, 387,* 805-808.

Kroll, E., Kellicutt, M. H., & Parks, T. (1975). Rehearsal of visual and auditory stimuli while shadowing. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 215-222.

Lyons, W. (1986). *The Disappearance of Introspection*, Cambridge, MA: MIT Press.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition, 115*, 303-313.

McElree, B., & Carrasco, M. (1999). Temporal dynamics of visual search: A speed-accuracy analysis of feature and conjunction searches. *Journal of Experimental Psychology: Human Perception & Performance, 25*, 1517-1539.

Miller, J., Vieweg, P., Kruize, N., & McLea, B. (2010). Subjective reports of stimulus, response, and decision times in speeded tasks: how accurate are decision time reports? *Consciousness and Cognition, 19*, 1013-1036.

Montague, W., Hillix, W., Kiess, H., & Harris, R. (1970). Variation in reports of covert rehearsal and in STM produced by differential payoff. *Journal of Experimental Psychology*, *83,* 249-254.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

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

Overgaard, M. (2006). Introspection in science. *Consciousness and Cognition, 15*, 629-633.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access, *Philosophical Transactions of the Royal Society of London – Series B, 367*, 1287-1296.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10*, 177-183.

Piccinini, G. (2003). Data from introspective reports: upgrading from common sense to science. *Journal of Consciousness Studies, 10*, 141-156.

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864-901.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*, 77-98.

Prinz, J. (2004). The fractionation of introspection. *Journal of Consciousness Studies, 11*, 40-57.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences, 3*, 1-23.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59-83.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461*, 263-266.

Sackur, J. (2009). L'introspection en psychologie expérimentale [Introspection in experimental psychology]. *Revue d'histoire des sciences, 62*, 5-28.

Schooler, J. (2002). Re-representing consciousness: dissociations between consciousness and meta-consciousness. *Trends in Cognitive Sciences, 6*, 339-344.

Schooler, J., & Schreiber, C. A. (2004). Experience, meta-consciousness, and the paradox of introspection. *Journal of Consciousness Studies, 11*, 17-39.

Selig, J. P., & Preacher, K. J. (2008). Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects [Computer software]. Retrieved from: http://quantpsy.org/.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review, 85*, 355-362.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 4*, 1787-1792.

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science, 153*, 652-654.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*, 147-246.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and Serial Processes in Visual Search. *Psychological Review, 114*, 71-103.

Titchener, E. B. (1899). *An outline of psychology*, New York: Macmillan.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Townsend, J. T., & Ashby, F. G. (1983). *The Stochastic Modeling of Elementary Psychological Processes*. Cambridge: Cambridge University Press.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1*, 46-54.

Townsend, J. T., & Wenger, M. J. (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review, 11*, 391-418.

van Veen, V., & Carter, C. S. (2002). The anterior cingulated as a conflict monitor: fMRI and ERP studies. *Physiology & Behavior, 77*, 477-482.

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

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

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

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

Wilson, T. D. (2003). Knowing when to ask. Introspection and the adaptive unconscious. *Journal of Consciousness Studies, 10*, 131-140.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419-433.

Wolfe, J. M. (1994). Guided Search 2.0: a revised model of visual search. *Psychonomic Bulletin & Review. 1*, 202-238.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences, 7*, 70-76.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 495-501.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems*, (pp. 99-119). New York: Oxford.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Sciences, 15*, 77-84.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error-detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111*, 931-959.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, *367*, 1310-21.

Young, A. H., & Hulleman, J. (2012). Eye movements reveal how task difficulty moulds visual search. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 168-190.

Appendix

The mediation model considered two levels: the outcome (SNSI), the potentially mediating variable (e.g., RT) and the predictor (the interaction between set-size and search type) are the within-participant variables (*i*, Level 1), which are nested within each participant (*j*, Level 2). Thus, we estimated a simple within-participant mediation analysis model, i.e., how much SNSI changes as a function of the interaction term (*total effect*), then, how much this change is contaminated by a mediator (*indirect effect*), and finally, how much SNSI changes as a function of the interaction term, after controlling the mediator effect (*direct effect*). We considered that all these effects can vary randomly between Level-2 units; therefore, the main aim of this analysis was to estimate the average of these effects across participants, considering the covariance between the random effects. According to Bauer et al. (2006) the coefficients for a multilevel mediation model were calculated as follows:

(1) $$Mediator_{ij} = \beta_{Mediator\ j} + \alpha_j\ Set\text{-}size * Search\ type_{ij} + e_{Mediator\ ij}$$

(2) $$SNSI_{ij} = \beta_{SNSI\ j} + \beta_j\ Mediator_{ij} + c'_j\ Set\text{-}size * Search\ type_{ij} + e_{SNSI\ ij}$$

(1.1) $$\alpha_j = \alpha + \upsilon_{\alpha_j}$$

(2.1) $$\beta_j = \beta + \upsilon_{1\,\beta_j}$$

(2.2) $$c'_j = c' + \upsilon_{2\,c'_j}$$

(3) Total effect = $\alpha\beta + \sigma_{\alpha_j\beta_j} + c'$

(4) Indirect effect = $\alpha\beta + \sigma_{\alpha_j\beta_j}$

Where (j) refers to participants and (i) to the observation. With the equation (1) we estimated the within interaction effect on the mediator (α_j). With the equation (2), we estimated the within interaction (*direct*) effect on the SNSI (c'_j), after controlling the within mediator effect on the SNSI (β_j), and vice-versa. In these equations, $e_{Mediator\ ij}$, $e_{SNSI\ ij}$ and $\beta_{Mediator\ j}$, $\beta_{SNSI\ j}$, denote residuals and random participant intercepts for the mediator and the SNSI, respectively. The coefficients α_j , β_j and c'_j , in equation (1) and (2), denote

random slopes for each participant (level-2 units). Note that the averages of these random effects across participants, as (1.1), (2.1) and (2.2) indicate, are the fixed effects of each model, where, υ_{α_j} , $\upsilon_{1\,\beta_j}$ and $\upsilon_{2\,c'_j}$, represent residuals for each of them. Thus, we estimated (1.1) the average effect of the interaction on the mediator variable (α); (2.1) the average effect of the mediator variable on the SNSI, after controlling the interaction effect (β); and (2.2) the average effect of the interaction on the SNSI, after controlling the mediator variable (c'). The total effect (c) and the indirect effect ($\alpha\beta$) can be obtained by the formulas (3) and (4), respectively; where, again, $\alpha = \bar{\alpha}_j; \beta = \bar{\beta}_j; c' = \bar{c}'_j$ and $\sigma_{\alpha_j\beta_j}$ is the covariance between the two random effects: α_j and β_j

2.2. Introspective access to an implicit shift of attention

Reyes, G., & Sackur, J. (*Unpublished*). Introspective access to an implicit shift of attention.

2.2.1. Introduction

Our recent study (Reyes & Sackur, 2014) proposed that introspection is not limited to cognitive states, as the literature has maintained over the decades (Nisbett & Wilson, 1977). In particular, our previous study shows that introspection is capable of distinguishing a serial processing of attention from an automatic or "parallel" processing. However, we could not pinpoint the relevant cognitive routine that was used by participants to increment the SNSI counter. We hypothesized that participants were tracking the number of attention shifts during the search. To evaluate this particular matter, using a similar experimental design to Experiment 4 (Paper 1), the attentional shift was modulated through a pre-conscious cue. The strategy was to maintain the stimuli conditions constant (equal set-size: 16) to prevent noisy information. The aim was to force the introspective mechanism into selecting information that takes place between the presentation of the stimuli and the decision. Three experimental conditions were presented: trials without attentional cues, trials with a congruent cue (at the position of the target) and trials with an incongruent cue (at the position of a distractor). Our hypothesis is that the estimation of the number of scanned items before participants take the perceptual decision (SNSI) will congruently vary with the subliminal modulation of the attentional shift. This variability should not be explained by changes in the first-order measures.

2.2.2 Results

Modulation of attention during serial visual search task was shown by eye movement. At the same time, two visibility tests confirm that the cue was not consciously perceived by the participants. In addition, introspection correctly captured the attentional shift in the serial condition, confirming the reliability of our metacognitive scale and our previous results.

2.2.3 Discussion

This study confirms our previous findings: under precise experimental controls it is possible to direct the introspection focus towards cognitive processes. Crucially, this study enables us to claim that part of the relevant inner information that introspection uses in our tasks is the number of attentional shifts effected by participants.

2.2.4 Paper

Introspective access to an implicit shift of attention

Gabriel Reyes [1, 2, 3] and Jérôme Sackur [1]

[1] Laboratoire de sciences cognitives et psycholinguistique, CNRS/EHESS/ENS, Paris, France

[2] Centro de Apego y Regulación Emocional (CARE), Universidad del Desarrollo, Santiago, Chile.

[3] Escuela de Psicología, Universidad Austral de Chile, Valdivia, Chile.

Author Notes

Acknowledgments: We thank Sid Kouider for useful discussions.

This work was supported by a doctoral fellowship from the National Commission for Scientific and Technological Research (CONICYT 72090838, Chile) to G.R., a grant from the Agence National de la Recherche (DYNAMIND ANR-10-BLAN-1902-01, France) to J.S. and by a grant from the Agence Nationale de la Recherche (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL) to G.R. and J.S.

Corresponding author: G. R. and J. S., Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure – PSL Research University, 29 rue d'Ulm, 75005, Paris, France. E-mail: jerome.sackur@gmail.com. Tel: + 33 (0) 1 44 32 26 25.

Abstract

Literature in metacognition has systematically rejected the possibility of introspective access to complex cognitive processes preceding a decision. This situation derives in part from the difficulty of experimentally manipulating cognitive processes while abiding by the two seemingly contradictory constraints: on the one hand participants must not be aware of the experimental manipulation, otherwise they run the risk of incorporating their knowledge of the experimental manipulation in some rational elaboration; thus eschewing the condition for real introspection and falling into confabulation. On the other hand, we need an external, third person perspective evidence that the experimental manipulation did impact some relevant cognitive processes. Here, we study introspection during visual searches, and we try to overcome the above dilemma, by presenting a barely visible, "pre-conscious" cue just before the search array. We aim at influencing the attentional guidance of the search processes, while participants would not notice that fact. Concurrently, we record eye movements, so as to collect behavioral evidence of the impact of our experimental manipulation. Results conclusively show that introspection of the complexity of a search process is driven in part by subjective access to its attentional guidance.

Keywords: introspection, consciousness, visual search, pre-conscious processing, attention

Introduction

The term introspection refers to the cognitive mechanism through which individuals can access their own mental states (Flavell, 1979; Lyons, 1986). Like many other cognitive processes, introspection also presents certain functional determinants (Jack & Shallice, 2001; Overgaard, 2006), that we can study in experimental psychology (Jack & Roepstorff, 2002, 2003, 2004; Schooler, 2002) and neuroscience (Fleming & Frith, 2014). In this line, theoretical models (Carruthers, 2010; Schwitzgebel, 2011) have suggested that individuals would introspectively access only mental states with low cognitive complexity (e.g., perceptual states). On the contrary, information of a high cognitive complexity (e.g., judgments and decisions) would be inaccessible to the individuals' consciousness. In particular, literature converges on the idea that reports about cognitive processes preceding a decision would not be truly introspective, but rather interpretative (Nisbett & Wilson, 1977; Overgaard & Sandberg, 2012). In effect, recent evidence in social psychology (Johansson, Hall, Silkström & Olsson, 2005; Johansson, Hall, Silkström, Tärning, & Lind, 2006; Petitmengin, Remillieux, Cahour, & Carter-Thomas, 2013) suggest that when participants are asked to describe the causes guiding their behavior, they systematically engage in confabulatory explanations; confirming Nisbett and Wilson's pessimistic views on introspection of 1977.

However, we (Reyes & Sackur, 2014) and others (Marti, Bayet & Dehaene, 2015) recently evidenced that under precise experimental conditions, participants' introspection was able to accurately access the basic cognitive process in visual search (that is, the mechanisms of a decision). Notice that, so as to reach this conclusion, it was necessary to resort to much simpler and controlled experimental conditions than what is generally the case in the literature supporting the inaccessibility of cognitive processes. In effect both studies (Reyes & Sackur, 2014; Marti, Bayet & Dehaene, 2015) relied on visual search as first order task on which participants were asked to apply introspection. This task has the convenient property of being experimentally well understood and controllable (Wolfe, 1994), while at the same time generating complex cognition, in the sense that searching for difficult visual targets leads to multi-step processes. With this in mind, recent results not only contradict what has been traditionally supposed in literature, but also opens the possibility that the controversy about the differential access to mental content (i.e., access to processes *vs*. access to perceptual states) does not depend – at least not completely – on the nature of such contents (Carruthers, 2010), but more on the functional aspects of the introspective mechanism itself (Goldman, 2006; Prinz, 2007).

In Reyes and Sackur (2014), we investigated the introspective access to the cognitive process underlying two visual search processes. Two experimental conditions were studied: difficult, effortful searches (search L among Ts) and easy, pop-out searches (search X among Ts). It has been suggested that these two visual searches rely on different forms of attentions (Treisman & Gelade, 1980; Wolfe, 1994; Wolfe, Võ, Evans & Greene, 2011): difficult searches lead to a partial sequential scanning of the search array, while in pop-out searches, attention is directly driven to the target, as if the whole array had been processed in parallel. As it has been commented above, the interest of the study was to evaluate whether the participants' introspection was able to describe the change in the underlying cognitive attentional processing. The introspective task consisted on estimating the number of scanned items *before* the decision. We argued that this subjective report (Subjective Number of Scanned Items, SNSI) is an index of the complexity of the cognitive *processes* leading to the decision. We found evidence that participants had an adequate introspective access to internal cognitive processes in the sense that subjective reports generally tracked experimental manipulations while they were not simply reflections of obvious contaminating sources (notably eye-movements and decision time). Thus, we suggested that participants' introspection was, through the Subjective Number of Scanned Items index, tracking the attentional guidance in effect during the search process. Here, we attempt to provide a direct proof of this by controlling attention during the search process while keeping all other properties of the search stimulus exactly identical.

To achieve this, we presented a barely visible cue before 50% of the search arrays. The cue could be congruent (at the position were the target was to appear) or incongruent (at the position of a future distractor), and was meant to exogenously drive participant's attention. The strategy was to maintain all other stimuli conditions as constant as possible (notably we used a fixed set-size of 16), so that we could observe the biasing effect of the cue on introspection of the processes preceding the decision. So as to gauge the visibility of the cue we applied a questionnaire about its visibility and a behavioral visibility test. Critically, we used an innovative procedure: we presented the visibility questionnaire and test after the first half of the experiment: participants thus had no information about potential cues during the first part of the experiment. This enabled us to assess the effect that knowing that a cue was sometimes present could have on introspection.

2.2.4.1 Experiment

Material and Method

Participants. 20 adults, French speakers (15 women), aged between 18 and 37 (mean age: 23 years, *SD*: 3.9) participated in the study. Informed consent was obtained before the experimental session. Participants received a compensation of €10 for 1-hour session. All the participants had no knowledge regarding the task and all had normal or corrected to normal vision.

Stimuli. Stimuli (Figure 1) consisted of a set of 16 green and red letters (T, L or X, size: 0.8° x 0.6°, luminance: 0.09 cd/m²) on a uniform grey background (luminance: 3.11 cd/m²). Stimuli were presented on an imaginary circle (radius: 6.2°) preceded by a black central fixation point at the center of the screen. Stimuli were equally spaced on the imaginary circle. On one half of the trials, a black transient (26 ms) cue was presented just before the stimulus array (*, size: 0.3° x 0.3°). When present, the cue was always at the position of a letter. Stimuli were presented on a CRT screen (size 17", resolution of 1024 x 768 pixels, refresh rate of 75 Hz, viewing distance of ≈ 55 cm). The experiment took place in a dark cabin with the monitor as the only source of light. Eye movements were recorded monocularly with an eye tracker (EyeLink 1000 system, SR Research, Canada), with a sampling rate of 500 Hz and a spatial accuracy better than 0.5° (Camera-Eye distance: ~55 *cm*). Saccades were determined using a conservative algorithm (velocity threshold: 30°/s, acceleration threshold: 8000°/s2, motion threshold: 0.15°). For all participants the right eye was recorded.

Procedure. After the fixation (500 *ms*), stimuli were presented for 3000 *ms*, either preceded or not by the cue (26 *ms*). The task consisted in deciding the color of the target (L or X) within the set of distractors Ts. In line with our previous study (Reyes & Sackur, 2014), participants were required to identify the color of the target, something they cannot do until it has been put under attentional focus. After presentation of the stimuli, there was a response window of 1500 *ms*, during which participants were requested to decide if the target presented was red ("Z" key) or green ("A" key) on a standard French keyboard. On half on the total trials, between the fixation point and the stimuli, a cue was presented, followed by an inter-stimulus interval (13 *ms*). In 50% of the cue present trials, the cue anticipated the position of the target. In the other 50%, the cue was presented at the position of a distractor. In this case, the cue kept a minimum (randomized) distance in the circle of stimuli of 4 or

6 letters with respect to the position of the target. The color (red or green) and the individual orientation of each stimuli (0, 90, 180, 270°) were randomized, as well as the type of target (feature condition: X or conjunction condition: L) and the type of cue (no-cue, congruent, incongruent). Immediately after the response of the participants, a visual analog introspective scale was presented: *How many elements do you think you scanned before reaching your decision?* (Subjective Number of Scanned Items, SNSI). Two labels were presented (in French) at each end of the scale: "*no item*" and "*all items*". Participants used their right hand, without time constraints, to move the cursor and click on the scale to give their introspective estimate. Meaning and use of the introspective scale was explained before the main experiment, while during the experiment instructions were presented in an abbreviated manner under the scale. Importantly, participants were instructed to avoid fast or automated responses and to use the whole scale.

With the objective of evaluating the visibility of the cue, after the first half of the experiment (i.e., after the fourth block), a questionnaire and an intermediary visibility test regarding the cue were presented. The aim of presenting these intermediary tests was to determine whether after they became aware of the experimental manipulation, individuals would modify their introspective performance. The questionnaire consisted in three yes-no questions. (1) *Did you see anything irregular or special during the experiment?* (2) *Did you see a stimulus between the fixation cross and the circle of stimuli?* (3) *Did you see this little black point presented just before the stimuli* (at that moment the researcher shows to participant the cue (*) on a piece of paper)*?* Whatever their responses, participants were then informed about the presence of the cue in 50% of the trials, while nothing about its relation to the target was said. Then, in a separate experimental block participants were presented with the same stimuli as in the main experiment, and their only task was to respond whether they saw the cue or not by means of a visibility scale, with two responses (in French): "*seen*" & "*not seen"*. Participants responded without time constraints. Immediately after the visibility questionnaire and test, participants continued with 4 similar experimental blocks.

Before beginning the experiment, participants received two-stage training. During the first (12 trials) the visual search task was presented without the introspective scale and with an audio feedback on correct and incorrect responses. This phase was repeated until participants reached a performance of 90% correct. The second training (12 trials) had the objective of exercising the use of the introspective scale. During the second stage participants proceeded to the main experimental block without performance criterion. The experimental session comprised

288 trials (144 repetitions per target type and 72 repetitions per congruent cue, 72 per incongruent cue, and 144 per no-cue) in 8 blocks (4 blocks before / after intermediary-test) with a 60 second pause between each block. The intermediary-test comprised 48 trials. The experimental session lasted 1 hour.

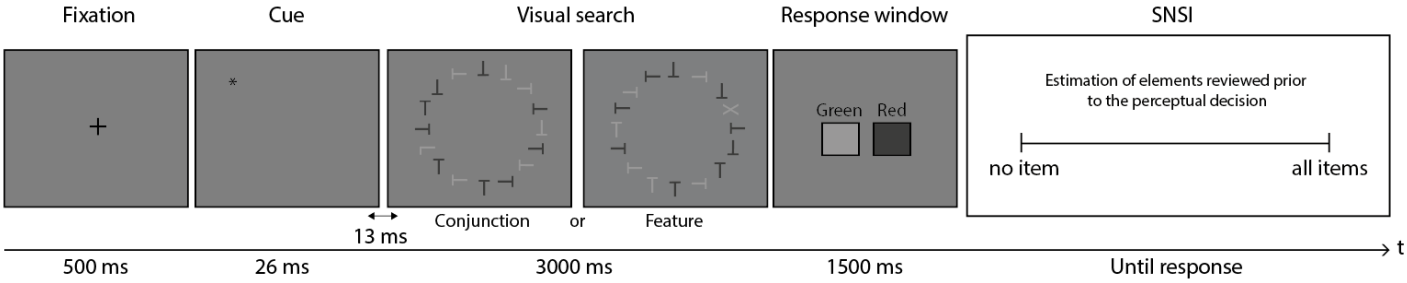


Figure 1. General structure of the experimental task. After presentation of the fixation point, a total of 25% of the trials presented a cue that anticipated the position of the target, another 25% presented a peripheral cue in respect to the position of the target. The stimuli were presented during 3000 *ms*, without the possibility of a response. During this stage the subjects were asked to identify the color of the target (red or green). All trials presented a target. Once the presentation of the stimuli concluded, a response window was presented to register the perceptual decision. Immediately after, the participants were requested to estimate the number of elements reviewed prior to having identified the color of the target on a qualitative scale (SNSI).

Results

Visibility tests. The visibility questionnaire and the visibility test in the middle of the experiment were meant to gauge how *noticeable* and how *detectable* the cue was. All participants reported not having noticed the cue during the first half of the experiment: all three questions of the questionnaire were negatively answered by *all* participants, precluding the application of statistical tests. However and critically, when we directed participants' attention on the cue for the detection test, it was clearly detected: mean accuracy across participants was 59%. We computed a visibility *d*' for each participant, which was strictly positive (median *d*' = .47, *SD* = .71, *t*-test against 0 (19) = 4.34, $p < .001$). Response bias (median *c* = .65, *SD* = .59, *t*-test against 0 (19) = 5.61, $p < .001$) was also positive, corresponding to a conservative bias. In sum, these results indicate that the cue was not subliminal in the sense that it could not be reported. However, it was clearly not spontaneously registered by

participants, who all failed to notice it. We suggest that this intermediary visibility state corresponds to the pre-conscious state proposed by (Dehaene, Changeux, Naccache, Sackur & Sergent, 2006). Thus any effect of the cue on introspection will be difficult to explain away by the evocation of an inferential process, at least in the first half of the experiment. It cannot be the case that participants change their introspection because they think, according to some causal theory, that the presence of the cue must have an impact on their second order reports, as, precisely, they are not spontaneously aware of the presence of the cue. Furthermore, as the visibility questionnaire and test did draw attention to the presence of the cue, if inferential processes were to intervene, this could be seen as a change in the effect of the cue between the first and second part of the experiment.

First order results. Response times (RT, computed from the onset of the response window) differed between L (497.0 *ms, SE* = 29.3) and X (459.9 *ms*, *SE* = 30.8) conditions but there was no difference depending on cue type or between the first and second half of the experiment. This was confirmed through a repeated measure ANOVA run on median RTs with three factors: target type (L and X) cue type (no-cue, incongruent and congruent) and pre / post (pre: first half, before the visibility test and questionnaire; post: second half), and all possible interactions. Here, as in the next analysis, participants were considered as a random factor. Results indicate only a significant main effect of target type ($F(1,19) = 12.02$, $p < .01$, $\eta^2_p = .39$) and an interaction between pre / post X target type ($F(1,19) = 8.93$, $p < .01$, $\eta^2_p = .32$). No other significant effects were found (all ps > .07). A deeper inspection to the interaction revealed that for L condition RTs significantly decrease from the first to the second half of the experiment (paired *t*-test pre (*L*) *vs*. post (*L*): $\Delta M = 58.8$, $SD = 88.9$, $t(19) = 2.96$, $p < .01$), but not for the X condition ($p > .70$, Figure 2A). Accuracy was high, with a mean error rate of 2% without any difference between experimental conditions (all ps > .15).

For eye movements, we calculated the latency of the first fixation on the target (the First Target Fixation Latency, FTFL, Figure 2B). We defined a square window around the target (0.8° x 0.8°), and measured FTFL with respect to the first fixation of at least 50 *ms* within this window. An ANOVA was run on median FTFL, with the same factors as for RTs, and we found that all main effects were significant: target type (L : $M = 1064.1$ *ms*, $SE = 34.7$, X : $M = 524.8$ *ms*, $SE = 22.0$, $F(1,19) = 410.1$, $p < .001$, $\eta^2_p = .96$); cue type ($F(2,38) = 20.6$, $p < .001$, $\eta^2_p = .52$) and pre / post ($F(1,19) = 10.1$, $p < .01$, $\eta^2_p = .35$). Importantly, a significant interaction between target and cue type was observed ($F(2,38) = 8.93$, $p < .01$, $\eta^2_p = .32$). A deeper inspection revealed that for the L condition congruent trials were faster than no-cue trials ($\Delta M = 199.4$, $SD = 209.6$, $t(19) = 4.25$, $p < .001$) and

congruent trials were faster than incongruent trials (ΔM = 225.2, SD = 282.4, $t(19) = 3.56$, $p < .01$), without differences between no-cue and incongruent trials ($p >$. 52). Regarding X condition, we found the same differences with smaller sizes (congruent *vs*. no-cue: ΔM = 49.1, SD = 72.6, $t(19) = 3.02$, $p < .01$; congruent *vs*. incongruent: ΔM = 90.6, SD = 80.8, $t(19) = 5.016$, $p < .001$). In addition no-cue trials were slightly faster than incongruent trials (ΔM = 41.5, SD = 68.5, $t(19) = 2.70$, $p < .05$). These results suggest that attention is in part driven by the cue, even though it is not spontaneously noticed when a congruent cue is presented. The differential effect of the cue in the X and L conditions demonstrates that the cue and the pop-out features of the X target concurrently affect attention. We found strictly similar results on the number of saccades during stimulus presentation (see Figure 2C).

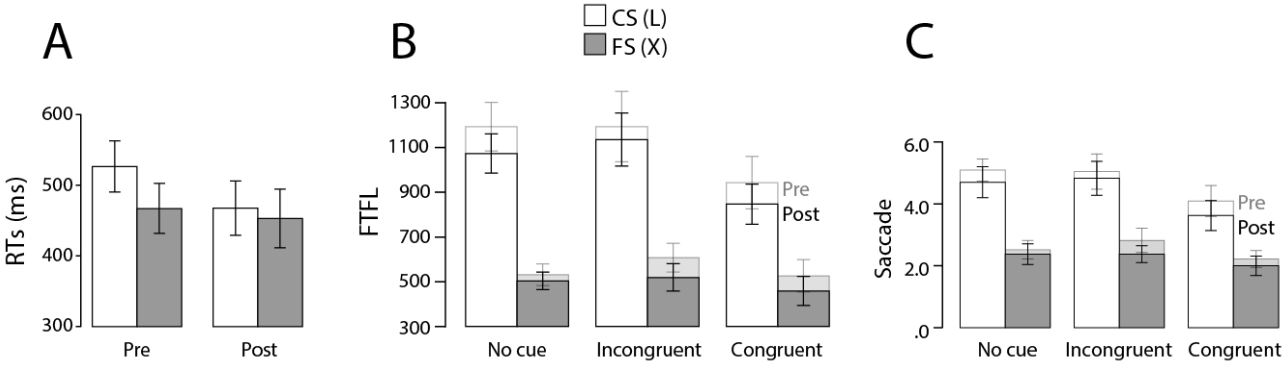


Figure 2. A) Response time B) First Target Fixation Latency (light grey and black bars denote pre and post factor, respectively), and C) Number of Saccade, as a function of the type of cue and for each visual search condition. Error bar represents ± 2 *SD*.

Second order result. We analyzed the subjective number of scanned items (SNSI) as an index of participants' introspection with similar ANOVAs. Critically, SNSI was reliably modulated by the cue, but only in the L condition (Figure 3). A repeated measure ANOVA over median SNSI, with the same showed that the three main effects were significant: SNSI was smaller for the X than for the L condition (L : M = 2.26, SE = .26; X : M = .69, SE = .19, $F(1,19) = 50.0$, $p < .001$, $\eta^2_p = .72$); it was driven by the type of the cue ($F(2,38) = 14.6$, $p < .001$, $\eta^2_p = .44$), and smaller in the second half of the experiment ($F(1,19) = 11.45$, $p < .01$, $\eta^2_p = .38$). Two interactions were evidenced: pre / post factor X target type ($F(1,19) = 6.95$, $p < .05$, $\eta^2_p = .27$) and target type X cue type ($F(2,38) = 7.68$, $p < .01$, $\eta^2_p = .29$). No other effects presented statistical significance ($p > .37$). A deeper inspection to the second interaction revealed that SNSI significantly differed within the L condition ($L_{\text{no-cue}}$ =

2.50, L $_{incongruent}$ = 2.62, L $_{congruent}$ = 1.65, $F(2,38) = 11.41$, $p < .001$, $\eta^2_p = .38$) but not within the X condition ($p > .63$). Multiple paired *t*-test comparisons indicate that for the L condition SNSI significantly decreased from no-cue to congruent cue trials ($\Delta M = .85$, $SD = 1.00$, $t(19) = 3.79$, $p < .01$), and from incongruent cue trials to congruent cue trials ($\Delta M = .97$, $SD = 1.07$, $t(19) = 4.07$, $p < .01$), however, no difference was found between no-cue and incongruent trials ($p > .54$).

Additionally, we investigated differences on the median response time for the SNSI task, that is, the time taken for the participants to respond on the SNSI scale. It is important to assess this point to rule out the possibility that the observed effects are not due to differences in how participants behaviorally performed the introspective task. A similar repeated measure ANOVA evidenced only a significant pre / post main effect ($F(1,19) = 12.70$, $p < .01$, $\eta^2_p = .40$) reflecting a learning effect across blocks. No other main effect or interactions were found.

Now, as the pattern of influence of the cue on SNSI is the same as the one found for the latency of the first fixation on the target, we investigated whether eye movements could explain them as a mediating variable. We restricted this analysis to the L condition, as results above showed that there was no impact of the cue on SNSI in the X condition. In fact, we found a significant linear regression between SNSI and FTFL ($r^2(119) = .38$, $\beta_{non\text{-}standardized} = .003$, $t = 8.51$, $p < .001$). Thus we tested whether the main effect between cue type and SNSI it is still significant after controlling FTFL, using a multilevel mediation model (Bauer, Preacher, & Gil, 2006). The model, run over individual trials, indicated that the cue type has a significant and negative relationship with FTFL ($r^2(2508) = .02$, $\beta_{non\text{-}standardized} = -98.18$, $t = -6.34$, $p < .001$). At the same time, FTFL and SNSI were, independently of cue type, significantly and positive related ($r^2(2508) = .23$, $\beta_{non\text{-}standardized} = .002$, $t = 27.20$, $p < .001$). However and importantly, after controlling FTFL, the relationship between SNSI and cue type was not significant anymore ($p >. 14$). This result suggests that participants' introspection tracks the attentional guidance during the visual search, as it is here manipulated by means of the cue.

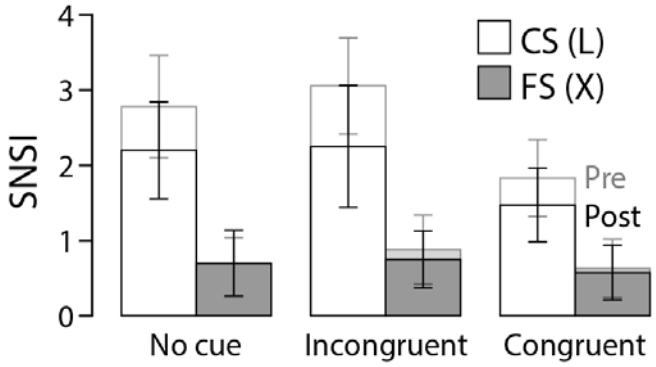


Figure 3. Subjective Number of Scanned Items as a function of type of cue and for each visual search condition. Light grey and black bars denote pre and post condition, respectively. Error bar represents ± 2 *SD*.

2.2.4.2 Conclusion

In this experiment we found that a non perceived cue presented before the target in a visual search paradigm influenced the subjective complexity of the search. This cue had no impact on the response times, as the responses were made during a response window after stimulus presentation, but it did impact eye movements: a cue presented at the position of the target shortened the time for the first fixation on the target when this was a hard to find one (L). In order to understand the import of this result on the mechanisms of introspection, it is important first to discuss the status of the cue with respect to visibility.

We introduced a visibility questionnaire and a visibility test in between the first and the second half of the experiment, so as to both assess the visibility of the cue and the impact of the knowledge of its presence on participants' introspection. Remember that participants were not apprised about the presence of the cue until the end of the visibility questionnaire. Results on this questionnaire conclusively show that no participant had noticed that a cue had been presented in half of the trials. However, when informed about this fact, they could detect it with better than chance performance. Thus we conclude that while not subliminal in a strict sense, the cue was not consciously (here in the sense of spontaneously) perceived. It was thus presumably pre-conscious in the sense of Dehaene *et al.* (2006). Importantly, with reference to theories of introspection, it could not in any way have been used by participants in rational, explicit inferences in the introspective judgments (the Subjective Number of Scanned Items response, SNSI), as they simply did not know about its presence. Moreover, since this fact was revealed after the questionnaire, we could test whether knowing about the cue (formally, the epistemic status of the cue) would impact introspection. As we did not find any significant interaction between the pre and post questionnaire phases of the experiment and cue effects on SNSI, we tentatively conclude that the cue impacts introspection through a non-conscious effect on the source of information that introspection draws upon, and not through a modification of some ex-post rational explanation that participants build. Indeed, within each search condition (X and L), congruent and incongruent cue trials were perceptually equivalent to participants, and lead to the same motor response. Nevertheless, they lead to different introspective reports because the non-perceived cue modified the mechanisms of the search.

Now, in what way did the cue impact the search? In fact, as it is shown by the eye movement results, the cue exogenously drew attention to its own location. Of course we do not have direct evidence for the attentional

effect of the cue, but links between eye movements and attention are now solidly established (Donk & van Zoest, 2011). Notice, in addition, that it is generally admitted (Haggard, 2005) that saccades themselves are not consciously experienced. Thus, in the case of the X condition, the incongruent (resp. congruent) cue competed with (resp. reinforced) the pop-out features of the target and slowed (resp. sped up) the first fixation on the target. Notice that this had no effect on introspection, as in all cue conditions, the subjective number of scanned items before the decision was already at floor (median SNSI below 1): the target pops-out, meaning that it is nearly always the first item perceived. In the L condition, the cue does have a facilitatory effect in case it is congruent, because it directs attention to the correct location. It does not have any detrimental effect in case it is incongruent because it is randomly positioned, as, we may presume, the spontaneous location of participants' attention when there is no cue. Thus, we may presume that unbeknownst to participants, the cue modified the trajectory of their search, by controlling its attentional guidance (Wolfe, 1994). Thus we successfully controlled the search process outside participants' awareness, so that they could not rationally elaborate on the role that the manipulation could have on their subjective reports. Given this, our results show that participants are able to report on the complexity of the search process as it is in part driven by its attentional guidance. We do not claim that attentional guidance is the sole contributor to the SNSI index. There may be additional factors, some of which may be rational elaborations akin to confabulations (for instance participants may generally think that "it is not possible that I scanned *all* items"). But the full mediation of the cue effect on SNSI by its effect on eye movements suggests that at least some of the variability in the SNSI response is directly caused by the number of "attentional fixations" during the search.

These results complement our previous results (Reyes and Sackur, 2014) that introspection can be flexibly attuned to first order processes so that, under some conditions pure mental monitoring of the search process was possible. However, in our previous study, we could not conclusively establish *what* internal source of information was used by participants. Here, we can do so, to the extent that at least some information in the SNSI index must come from an active tracking of attentional guidance during the search. Our results converge with recent results by (Marti, Bayet & Dehaene, 2015) showing that participants were able, to some extent, to report the trajectory of their fixations during a serial search. The authors could even present evidence that the divergences between real and subjectively reconstructed eye movements were due to confusions between eye and attentional fixations. However, since the authors did not have external control on the eye movements themselves, they could not conclusively assess whether divergences between introspection and observed eye

movements were due to errors of introspection (be they confabulatory or simple mistakes), or to introspection that does not match behavior. Our methodology enables us to conclude that, while certainly not infallible, introspection does access internal attentional mechanisms that are instrumental to the search process.

In their famous 1977 paper, Nisbett and Wilson used instances where participants introspection was driven by a cause that they did not acknowledge to show how prone to confabulation they were: according to their often quoted example, participants elaborate on the quality of the socks they chose to justify their choice, while in fact it appears that they most probably simply chose the ones on the right. Ironically, we turn the argument around: it is precisely because participants are not aware of a cause (of which we have implicit behavioral evidences through eye movements) that influences their search process, that we can assert that their introspection is, at least in part, veridical. Now, if introspection of inner processes is in principle veridical and based on identifiable sources of information, it means that it might be assimilated to a decision process, in the same way that perception is classically understood as a decision process. This opens exciting avenue for further research about the causes and consequences of errors or introspection (introspective false alarms and misses) as well as of correct introspections (hits and correct rejections).

References

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychol. Methods, 11*, 142-163.

Carruthers, P. (2010). Introspection: Divided and Partly Eliminated. *Philos. Phenomenol. Res., 80,* 76-111.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci.,* 10, 204-211.

Donk, M. & van Zoest, W. (2011). No control in orientation search: The effects of instruction on oculomotor selection in visual search. *Vision Res., 51,* 2156-2166.

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

Fleming, S. M., & Frith, C. D (Ed.). (2014). *The Cognitive Neuroscience of Metacognition.* Berlin: Springer.

Goldman, A. (2004). Epistemology and the evidential status of introspective reports. *J. Conscious. Stud., 11*, 1-16.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends Cogn. Sci., 9,* 290-295.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus-response to script-report. *Trends Cogn. Sci., 6,* 333-338.

Jack, A. I., & Roepstorff, A (Ed.). (2003). *Trusting the subject? Volume 1*. Exeter: Imprint Academic.

Jack, A. I., & Roepstorff, A (Ed.). (2004). *Trusting the subject? Volume 2*. Exeter: Imprint Academic.

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

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

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

Lyons, W. (1986). *The Disappearance of Introspection*. Cambridge, MA: MIT Press.

Marti, S., Bayet, L., & Dehaene, D. (2015). Subjective report of eye fixations during serial search. *Conscious Cogn, 33*, 1-15.

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

Overgaard, M. (2006). Introspection in science. *Conscious. Cogn., 15*, 629-633.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Phil. Trans. R. Soc. B, 367,* 1287-1296.

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Conscious. Cogn., 22,* 654-669.

Prinz, J. (2007). Mental pointing: Phenomenal knowledge without concepts. *J. Conscious. Stud., 14,* 184–211.

Reyes, G., & Sackur, J. (2014). Introspection during visual Search. *Conscious Cogn., 29,* 212-229.

Schooler, J. (2002). Re-representing consciousness: dissociations between consciousness and meta-consciousness. *Trends Cogn. Sci., 6,* 339-344.

Schwitzgebel, E. (2011). *Perplexities of consciousness.* Cambridge, MA: MIT Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol., 12,* 97-136.

Wolfe, J. M. (1994). Guided Search 2.0: a revised model of visual search. *Psychon. Bull. Rev., 1,* 202-238.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends Cogn. Sci., 15,* 77-84.

2.3 Introspection during working memory scanning

Reyes, G., & Sackur, J. (*Unpublished*). Introspection during working memory scanning.

2.3.1 Introduction

The introspection of cognitive processes is possible in the context of a visual search task (Reyes & Sackur, 2014). The following study attempts to extend these findings to another cognitive domain: memory search. The strategy consists in investigating whether introspection is able to describe the cognitive process displayed during the memory information recovery process. The first experiment investigated participants' introspection during a Judgment of Recency task (JOR, Muter, 1979; Hacker, 1980). In the second experiment, introspection was investigated during an Item Recognition task (IR, Sternberg 1966, 1969).

It is important to point out that both tasks do not differ with respect to stimuli conditions, but only in the instruction or cognitive demand. As a consequence, stimuli itself cannot be considered a source of confabulatory information. Similar to our previous studies, on a trial-by-trial basis, participants were requested to do an estimation of the number of scanned items before the decision (SNSI). Previous studies based on purely behavioral data converge on the idea that an IR task generates a cognitive parallel process. On the other hand JOR tasks would generate serial cognitive processes. We predicted that introspection will successfully describe the nature of the information recovery process, and thus yield different pattern of quantified introspection for the IR and JOR tasks.

2.3.2 Results

In both experiments we replicate the classical first order patterns for JOR and IR tasks. The second-order results confirm our predictions: In Experiment 1 (JOR task), the SNSI significantly increases as a function of the serial pattern of information recovery. In Experiment 2 (IR task), introspection of the cognitive processes is constant, which is compatible with a parallel process. To summarize, introspective reports and first order measures indicate that individuals performed a serial mental scan in Experiment 1 but not in Experiment 2.

2.3.3 Discussion

The following paper demonstrates that introspection, under the presented experimental conditions, is sensible to precise changes in the cognitive process, here defined as the mechanism of information recovery from working memory. At the same time, these results extend our previous findings to another cognitive domain, suggesting the presence of a common introspective mechanism, independent from cognitive task.

2.3.4 Paper

Introspection during Working Memory Scanning

Gabriel Reyes [1, 2, 3] and Jérôme Sackur [1]

[1] Laboratoire de sciences cognitives et psycholinguistique, CNRS/EHESS/ENS, Paris, France

[2] Centro de Apego y Regulación Emocional (CARE), Universidad del Desarrollo, Santiago, Chile.

[3] Escuela de Psicología, Universidad Austral de Chile, Valdivia, Chile.

Author notes

Acknowledgments: We thank Sid Kouider for useful discussions. We thank also Isabelle Brunet and Cécile Girard for their help in data acquisition.

This work was supported by a doctoral fellowship from the National Commission for Scientific and Technological Research (CONICYT 72090838, Chile) to G.R., a grant from the Agence National de la Recherche (DYNAMIND ANR-10-BLAN-1902-01, France) to J.S. and by a grant from the Agence Nationale de la Recherche (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL) to G.R. and J.S.

Corresponding author: G. R. and J. S., Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure – PSL Research University, 29 rue d'Ulm, 75005, Paris, France. E-mails: gureyes@uc.cl; jerome.sackur@gmail.com. Tel: + 33 (0) 1 44 32 26 25.

Abstract

The literature in metacognition has argued, for many years, that introspective access to our own mental contents is restricted to the cognitive *states* associated with the response to a task, such as the level of confidence in a decision or the estimation of the response time, however, the cognitive *process* that underlies such states would be inaccessible to the participants' consciousness. Here, we set out to expand the range of introspective information submitted to experimental scrutiny by asking whether participants could introspectively distinguish the cognitive process that underlies two working memory tasks. For this purpose, we asked participants, on a trial-by-trial basis, to report the number of items that they scanned during their working memory retrieval, which we have named "Subjective Number of Scanned Items" (SNSI). The SNSI index was evaluated, in Experiment 1, immediately after a Judgment of Recency (JOR) task and, in Experiment 2, after an Item Recognition (IR) task. The results showed that participants' introspection successfully accessed the complexity of the decisional processes.

Keywords: introspection, memory scanning, cognitive processes, metacognition, working memory.

Introduction

The reliability of introspection, that is the ability to monitor the content of our own mental activity, has been much discussed within experimental psychology (Lyons, 1986; Costall, 2006). For instance, according to an influential theory (Nisbett & Wilson, 1977), we only have introspective access to information relative to the consequences of our cognitive processes: With respect to perceptual decisions, participants might access their confidence (Pleskac & Busemeyer, 2010; Fleming, Weil, Nagy, Dolan, & Rees, 2010) and the overall decision duration (Corallo, Sackur, Dehaene, & Sigman, 2008; Marti, Sackur, Sigman, & Dehaene, 2010; Miller, Vieweg, Kruize, & McLea, 2010), but not the properties of the process itself, i.e., set of cognitive operations that precedes the decision. According to Nisbett and Wilson theory, while a cognitive state may in some context be introspectively accessed, a cognitive process is, by and large inaccessible, producing only confabulatory introspective reports (Kozuch & Nichols, 2011). Despite substantial and early objections (Smith & Miller, 1978; White, 1980, 1987, 1988; Ericsson & Simon, 1980), and recent reformulations (Wilson, 2002, 2003; Wilson & Dunn, 2004), this idea is often viewed as a canon in the literature (Johansson, Hall, Sikström, & Olsson, 2005; Overgaard, 2006; Overgaard & Sandberg, 2012). In opposition to this, we recently showed (Reyes & Sackur, 2014) that introspection is highly flexible, in the sense that it can combine *self-observation* and *mental monitoring*, so that it can ultimately target cognitive processes. We reached this conclusion based on evidence that, in a visual search paradigm, participants introspectively differentiate attentionally guided and non-guided searches, which are features of the cognitive search process. Here, we tested whether this conclusion could generalize to other kinds of processes. We ask whether short-term memory recovery processes were accessible through introspection.

To assess if individuals can introspectively access memory search processes, we designed two short term memory tasks that differed in the complexity of the information retrieval process, but were nearly identical with respect to stimuli and experimental design. Literature converges on the idea that the nature of the information requested in a memory task determines the complexity of the recovery operation deployed (Dosher, 2003; McElree, 2006; Jonides, Lewis, Nee, Lustig, Berman, & Moore, 2008; Chan, Ross, Earle, & Caplan, 2009). For instance, when short lists of items are memorized, tasks that require participants to recover the simple identity of an item (Sternberg, 1966, 1969) evidence a direct access to this information (McElree & Dosher, 1989; McElree, 2001): participants have a parallel access to all items stored in working memory. In contrast, with the same

material, memory recovery tasks of relational information (the order of presentation, Muter, 1979; Hacker, 1980) require a serial and ordered access to the set of representations in memory (McElree & Dosher, 1993). In sum, depending on the task instruction two types of cognitive processes would take place. Now, knowledge of these differences come from the classical, objective, third person perspective of experimental cognitive psychology. Here, we ask whether from a first person perspective, participants have access to the difference of the processes involved.

In Experiment 1 we engaged participants in a judgment of recency (JOR) task. Immediately after the presentation of a list of six consonants, participants were required to determine (in a two-alternative forced-choice procedure, 2AFC) which of two letters was closer to the end of the list. The letter closer to the end of the list was considered the target and the other the distractor. Earlier studies with the JOR task (Muter, 1979; Hacker, 1980) showed that response times (RTs) increase linearly as a function of the position of the target: the closer to the end of the list, the faster the response, independent of the distractor position. This effect suggests a serial and self-terminating recovery mechanism (McElree & Dosher, 1993). In Experiment 2, we asked participants to perform an item recognition task (IR), similar to the Sternberg rapid scanning task (Sternberg, 1966). Immediately after the presentation of a list of consonants, participants were asked to determine with a 2AFC which of the two letters was present in the list presented beforehand. In contrast to the JOR task, here only one consonant was previously presented in the list. Sternberg (1966, 1969) showed that the time it takes participants to decide on the presence of the target varies linearly as a function of the number of items presented (list length or set-size). In addition, Sternberg observed that this slope was identical in target present and absent trials. This pattern was at first (Sternberg, 1966) interpreted as evidence for a serial exhaustive scanning process, however, McElree & Dosher (1989) showed that a direct access mechanism better explains the recovery mechanism. The authors showed that if we examine each set size separately, accessibility of the information (i.e., both the amount of information necessary to correctly identify the target, as well as the speed of recovery) remained constant for each target positions in the list of items. The notable exception to this finding was that the last position presented significantly higher recovery availability than the rest (McElree, 2006; Wickelgren, 1980). With this exception in mind, the consensus is now that in item recognition tasks, the recovery process accesses information in parallel, in a way that mimics (Townsend & Wenger, 2004) a serial exhaustive process. Overall, the literature reviewed establishes that information recovery mechanisms are qualitatively different in the judgment of recency and item recognition tasks.

Our aim here was to investigate participants' introspection of these mechanisms. To this end, in each task and on a trial-by-trial basis, we asked participants to report the number of items scanned during the memory recovery process. We called this index "Subjective Number of Scanned Items" (henceforth, SNSI, Reyes & Sackur, 2014). Our prediction is that if participants were able introspectively to access their cognitive memory processes, then the two tasks should lead to dissimilar results on the SNSI: in the JOR task we expected that the SNSI would vary as a function of the target position, as participants would consciously and subjectively access the serial memory scanning process. In the IR task, we expected that SNSI could only vary as a function of the global memory load. In other words, we expect SNSI to vary with respect to set-size but not as a function of the target position within each list, as no effective scanning process does take place.

2.3.4.1 Experiment 1

Material and methods

Participants

Eighteen normal adults, French speakers (11 women), aged between 19 and 30 (mean age: 23 years, *SD*: 3.06) participated in the study. In both experiments informed consent was obtained before the experimental session and participants received compensation of €10 for a 1-hour session.

Stimuli and Procedure

Stimuli consisted of black consonants (size: 0.8º X 0.6º; luminance: 0.5 cd/m^2) on a uniform grey background (luminance: 44.1 cd/m^2). Stimuli were presented on a CRT screen (size 17", resolution of 1024 X 768 pixels, refresh rate of 100 Hz, viewing distance of ~55 cm). The experiment took place in a dark room with the monitor as the only source of light.

A trial (Figure 1) consisted in a fixation cross (500 *ms*) followed by a list of six letters each of which was presented for 500 *ms*, with an inter-stimulus interval of 50 *ms*. Stimuli were drawn randomly without replacement from the set of consonants except "X". After the last letter, a mask (random string of five symbols) was presented for 500 *ms*. The aim was to avoid a simple match between the last letter presented and the two probe letters, which were presented left and right of fixation (separated by 2°), until participants' response. The two letters were drawn from the list, and participants were instructed to decide, which was closer to the list's end. All fifteen targets and the distractor positions combinations were tested, and the position of the target in the 2AFC task was counterbalanced across trials. To respond, participants used the "A" and "Z" keys on a standard French AZERTY keyboard, corresponding respectively to the left and right probe letter.

Immediately after this first order decision, a visual analog scale appeared on screen with the introspective question: *how many consonants did you scan before you reached your decision?* This subjective number of scanned items (SNSI) scale had four labels (in French): "no item", "some items", "many items" and "all items",

but we recorded the exact position of the cursor on the scale, and participants were instructed to use all the scale. Participants were instructed to avoid fast and automatic responses.

Two trainings preceded the main experimental blocks. During the first training of 30 trials, the JOR task was presented without the SNSI scale, but with audio feedback on correct and incorrect responses. This phase was repeated until participants reached a criterion of 90% correct. The second training of 30 trials introduced the SNSI scale and participants proceeded to the main experimental block without performance criterion. The experimental session, without audio feedback, contained 240 trials in 8 blocks (16 repetitions per target / distractor combination) with a 60 second pause between each block.

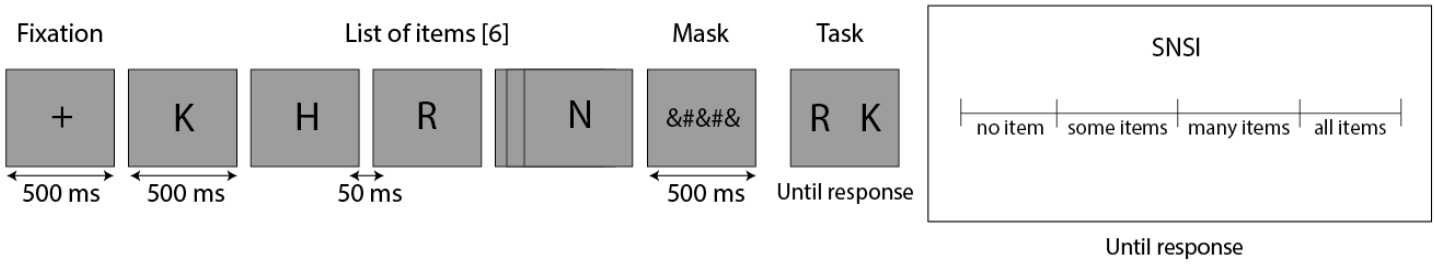


Figure 1. General structure of the JOR task. After the fixation cross, a list of 6 consonants and a concluding mask were presented. The task consisted of determining which of the two letters probe letters, was the last in the list (e.g. in the figure, the target is the letter "R", presented in 3rd position and accompanied by the distractor "K", presented in the 1st position). After this decision, participants gave, without time constraints, the subjective estimation of the number of items scanned in the memory *before* identifying the target (SNSI). Labels were presented on the scale, but participants were instructed to use all the positions.

Results

First order task

We excluded trials with response times below 200 *ms* and trials with response times 3 *SD* above the median (4%). RTs were log-transformed to normalize the distribution. In order to provide a concise summary of the first order results, and given that response times and error rate were correlated across conditions ($r(269) = .18$, $\beta = 960$, $SE = 322$, $p < .01$; mean Error rate for each target position: 6th: 5%; 5th: 11%; 4th: 18%; 3th: 28%; 2th: 30%;

Figure 2A), we combined these variables in the form of Inverse Efficiency Scores (henceforth: IES, Townsend & Ashby, 1983; see also Austen & Enns, 2003; Bruyer & Brysbaert, 2011), which is the ratio of median RTs over proportion of correct responses for each participant and experimental condition; thus, lower values indicate better performance. We performed a linear mixed model (LMM) on IES, with fixed effects of target position (from 2 to 6), distractor position (from 1 to 5) and their interaction. As random effects, we considered intercepts and a random slope for distractor position levels for each participant.

First order results agree with the literature: performances were higher for targets near the end of the list, independent of distractor position. Accordingly, the LMM showed a significant main effect for target position ($F(3, 244) = 8.20$, $p < .001$), without main effect for distractor position ($p > .27$) nor interaction between these ($p > .76$). IES did not vary within each target position condition (all ps > .08).

Second order task

Second order results paralleled first order results (Figure 2B): participants reported that they subjectively scanned fewer items when the target was closer to end of the list. This introspection was not impacted by distractor position. We statistically tested this effect with an LMM on SNSI, with the same fixed and random factors as before. We found a significant main effect for target position ($F(3,231.2) = 17.27$, $p < .001$), with no effect of distractor position ($p > .56$), and no interaction ($p > .78$). Visual inspection of the results showed that introspection for the last target position presented a large decrease compared to other positions. However, when we repeated the above analysis excluding the 6th position, we still found a main effect of target position ($F(2,141) = 3.10$, $p < .05$). The same analysis, separately conducted over each target position level, showed that participants' estimates did not vary as a function of distractor position (all ps > .44).

A possible interpretation of these results is that participants responded to the SNSI by mentally rehearsing the sequence of the letters and then simply counted the number of letters between the target and the end of the list. Participant might essentially do the same task twice. We reasoned that if this was the case, the time taken by participants to do the SNSI task would present a similar pattern as first order response time. However, the same LMM on the median response time for the SNSI task (RT_{SNSI}, Figure 2C) showed that none of these main effects were significant (target position: $p > .12$; distractor position: $p > .82$) nor the interaction term ($p > .77$). In

addition, median RT_{SNSI} and median first-order RTs were not significantly correlated ($p > .22$). In sum, participants did not appear to perform the first-order task a second time when asked to report their introspection: they seem to report a value that they read-out from the first-order process.

However, it is still possible that participants did not base their introspective reports only on internal information generated during the memory scanning process. They might have used also self-observation of their own behavior, or other internal dimension of the process, such as felt difficulty (Bryce & Bratzke, 2014). In this experiment, we only had access to RTs as an aggregated measure of these other subjective parameters. Thus, so as to form a conservative estimate of the weight of mental monitoring in SNSI reports, we tested a multilevel mediation model on correct trials (Bauer, Preacher & Gil, 2006, Figure 2D)[17], testing the potential mediating role of RTs in the formation of the SNSI.

In agreement with the previous analysis, we found a significant main effect, of the target position on SNSI ($F(1,17.0) = 21.69$, $c = -.08$, $p < .001$). We also confirmed that the same factor impacted RTs ($F(1,16.9) = 66.60$, $\alpha = -.22$, $p < .001$). In addition, controlling the target position main effect on SNSI, RTs presented a significant relationship with SNSI ($F(1,16.9) = 49.67$, $\beta = .24$, $p < .001$), which is required for RTs to be considered as potential mediator. Finally, we estimated the mediation model: after controlling the RT effect on SNSI, we found that the impact of the target position factor on SNSI was significantly reduced but not eliminated ($F(1,16.4) = 9.09$, $c' = -.02$, $p < .01$), suggesting a partial mediation of SNSI by RTs. The size of the indirect effect ($\alpha\beta$) was -.05 (C.I. [-.07, -.03]), which indicate that 63% of the target position effect on SNSI is mediated by RTs.

[17] This model allows us to factor out the mediating role of RTs in the target position effect on SNSI. The mediation model has two levels. The outcome (SNSI), the mediating variable (RTs) and the predictor (target position factor: TP) are the within-participant variables (*i*, Level 1), which are nested within each participant (*j*, Level 2). Thus, we estimated a within-participant mediation analysis model, i.e., how much SNSI changes as a function of the predictor (*total effect* = *c*). Then, how much this change is contaminated by a mediator (*indirect effect* = *αβ*), and finally, how much SNSI changes as a function of the predictor, after controlling the mediator effect (*direct effect* = *c'*). We considered that all these effects can vary randomly between Level-2 units. The main aim of this analysis was to estimate the average of these effects across participants (α, β and *c'*), considering the covariance between the random effects ($\sigma_{\alpha_j\beta_j}$). To test the significance of the indirect effect, we used a Monte Carlo confidence interval method (Selig & Preacher, 2008; Preacher & Selig, 2012).

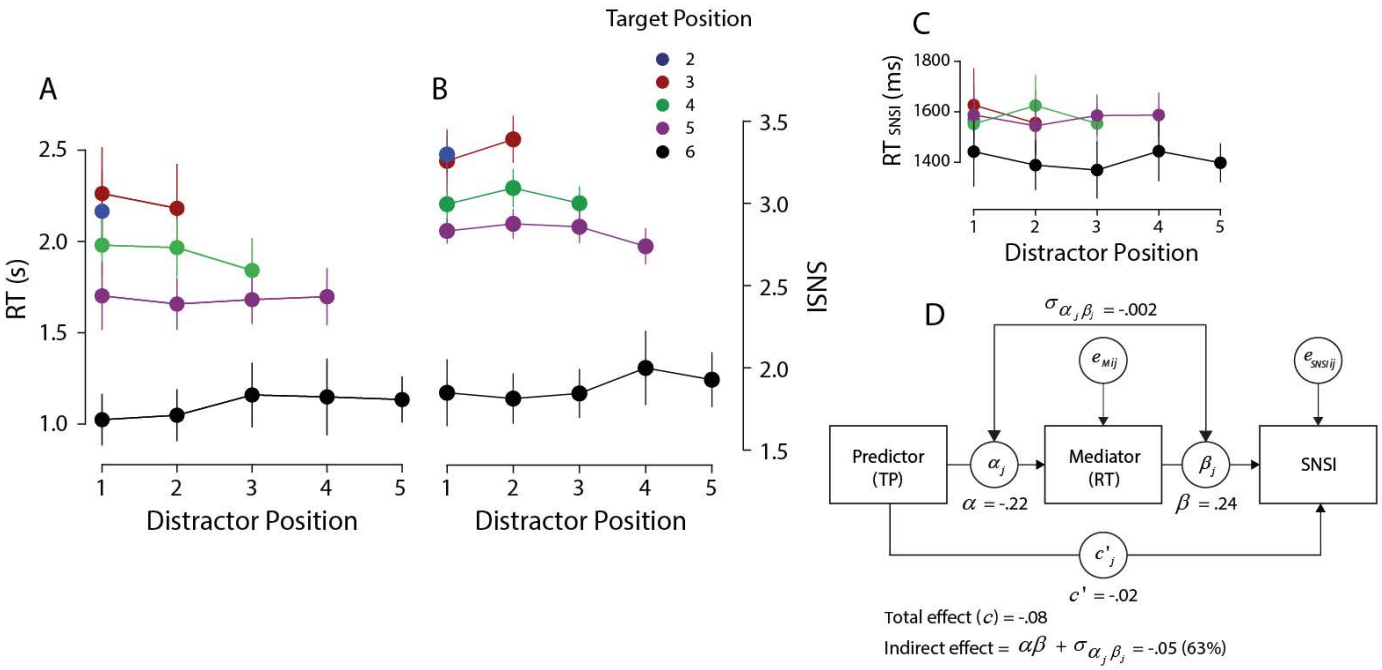


Figure 2. A) Response Time (RT) as a function of distractor and target positions. Error bars, here and in the next figures, represent Cousineau-Morey within-subjects 95% confidence intervals (Cousineau, 2005; Morey, 2008). B) Subjective Number of Scanned Items (SNSI) and (C) RT_{SNSI} as a function of the same factors. D) Mediation model: In the figure, boxes indicate variables tested (Target Position (TP), Response Time (RT) and Subjective Number of Scanned Items (SNSI)), circles represent random effects for each regression (α_j , β_j and c'_j) and under these, the averaged effects associated. Every path shows a circle for the residuals of each regression.

Discussion

In Experiment 1, we used a judgment of recency (JOR) task as a test bed. It has been shown that this task generates a serial and self-terminating recovery processes (Muter, 1979; Hacker, 1980; Chan, Ross, Earle, & Caplan, 2009). Immediately afterward, we asked participants, on a trial-by-trial basis, to report how many items they had scanned during the recovery process. We considered this introspective variable (Subjective Number of Scanned Items, SNSI) as an index of subjective access to the complexity of the memory recovery process. Results on the first order task agree with the literature: performance increased as targets were closer to the end of the list, without any effect of distractor position. This pattern is consistent with a serial self-terminating process (McElree & Dosher, 1993; McElree, 2006). Results on the second order task were strikingly similar, as SNSI decreased the later the targets appeared in the lists. Thus, it seems that participants are subjectively aware of the

serial scanning process involved in the judgment of recency task. We took a number of steps to strengthen the interpretation of the SNSI index. First, we wanted to make sure that, when asked to tell how many items they scanned, participants were not simply rehearsing the list. We reasoned that if participants had been implementing this strategy, RTs for the second order task would positively correlate with the RTs in the first-order task. Even more stringently, we reasoned that if participants had used that, or any other, rehearsing strategy it should show as an effect of target position on response times for the second order task. Both analyses yielded null results, suggesting that when participants perform the SNSI task they do not operate on the first order stimulus or its re-representation. It seems more parsimonious to suppose that the SNSI is generated from a read-out of the internal signal produced during the first-order process.

Second, we made sure, by means of a mediation model that results on the SNSI are not trivially explained by self-observation of behavior: the relation of target position on SNSI is only partially mediated by response times (RTs), meaning that participants' introspection is not fully explained by self-observation of their own motor behavior during the first order task. We insist that the results of the mediation model should essentially be interpreted negatively: We do not claim that response times do actually mediate the effect of target position on SNSI, and indeed we do not have any evidence for this claim. The presence of a significant direct effect (37%) suggests nevertheless that part of the SNSI derives from mental monitoring rather than self-observation of RTs.

In conclusion, analyzes of first order results and introspection concur in suggesting that recovery of relational information in the JOR task is achieved through an introspectively accessible scanning mechanism. Now, in Experiment 2, we tried to shift the memory process from a serial scanning mode to a direct parallel access mode, while we kept the stimuli as close as possible to those of the JOR task.

2.3.4.2 Experiment 2

Material and methods

Participants

Twenty-four normal adults, French speakers (14 women), aged between 20 and 28 (mean age: 22 years, *SD*: 2.01) participated in the study.

Stimuli and Procedure

The stimuli do not differ from those of Experiment 1, except that we used three list length (4, 6 and 8). Here, participants were requested to decide which of the two letters was present in a list of consonants. Only one of these two letters was randomly selected from the list, the other letter was always different. The experimental session contained 240 trials in 8 blocks. Ten repetitions of three list lengths (4, 6 or 8 consonants) were randomly intermixed in each block. Immediately after the first order decision, participants responded on the SNSI scale, with exactly the same instructions as in Experiment 1. Other aspects of the stimuli, procedure and training phase did not differ from those of Experiment 1.

Results

First order task

We excluded trials with response times below 200 *ms* and trials with response times 3 *SD* above the median (3.2%). RTs were log-transformed. As in the previous experiment, median RTs and mean error rate presented a positive and significant correlation across conditions ($r(431) = .28$, $\beta = 1334.9$, $SE = 220$, $p < .001$; mean Error rate for each Set-Size: 8 consonants: 14%; 6 consonants: 9%; 4 consonants: 7%), therefore, they were transformed into inverse efficiency scores (IES). Results showed that performances increased with shorter list length and with targets closer to the end of the lists (Figure 3A). So as to statistically test these effects, we ran a LMM on IES with fixed effects of list length (4, 6 and 8 items), target position within the list, and their interaction. As random effect we considered a random target position slope for each participant and random

participant intercepts. We found a significant main effect for list length ($F(2,403.2) = 17.3$, $p < .001$) and for target position ($F(1,403.5) = 14.7$, $p < .001$), without interaction ($p > .12$).

Second order task

Results on the SNSI task do not mirror first order performance: the number of scanned items increased with list length (Figure 3B), but within each list length, the only subjective difference seemed to be between the last item in the list and all previous ones. Again, we quantified these results by means of an LMM on mean SNSI, with fixed factors of list length (4, 6 and 8), target position within the list and their interaction. We included a random participant intercept and a random target position slope for each participant. First, we found a significant main effect for list length ($F(2,380.5) = 63.2$, $p < .001$), a main effect for target position ($F(1,46.1) = 7.72$, $p < .01$), with an interaction ($F(2,380.5) = 3.08$, $p < .05$: 4 consonants ($p > .90$), 6 consonants ($F(1,34.8) = 12.7$, $\beta = -.12$, $p < .01$), 8 consonants ($F(1,30.6) = 20.9$, $\beta = -.16$, $p < .001$)). None of these main effects (list length: $F(1,19.5) = 60.1$, $p < .001$; target position: $F(1,28.7) = 13.0$, $p < .01$) nor the interaction effect ($F(1,367.6) = 60.8$, $p < .001$) disappeared after controlling RTs.

Second, we repeated this analysis without the last target position at each set-size level. We still found that the list length factor had a significant impact on SNSI ($F(2,310.9) = 55.9$, $p < .001$). This effect was not simply explained by self-observation of RTs (Figure 3C)[18]. Interestingly, in line with our hypothesis, the target position factor was no longer significant ($p > .99$). The interaction between these factors was also significant ($F(2,310.9) = 3.47$, $p < .05$). More precisely, SNSI did not vary when 4 ($p > .08$) or 6 ($p > .32$) consonants were presented, but it did for 8 consonants ($F(1,33.3) = 7.17$, $\beta = -.09$, $p < .05$).

[18] We applied a multilevel mediation model, with RTs as potential mediator variable for the persistent effect we found on SNSI. We investigated the source of the list length main effect on SNSI (i.e., without considering last position). The multilevel mediation model, run on a trial-by-trial basis and only on correct trials, confirmed a significant main effect of list length on SNSI ($F(1,27.2) = 214.6$, $c = .37$, $p < .001$). At the same time, we confirmed that this factor has also an impact on RTs ($F(1,25.0) = 78.9$, $\alpha = .06$, $p < .001$). Then, we observed that after controlling the list length effect, SNSI and RTs presented a significant relationship ($F(1,22.6) = 87.5$, $\beta = 1.28$, $p < .001$). Finally, after controlling the RTs effect on SNSI, we found that the impact of the set-size factor on SNSI was significantly reduced, but not eliminated ($F(1,22.9) = 106.2$, $c' = .28$, $p < .001$). The size of the indirect effect ($\alpha\beta$) was .08 (C.I. [.05, .10]), indicating that 22% of the set-size effect on SNSI is mediated by RTs.

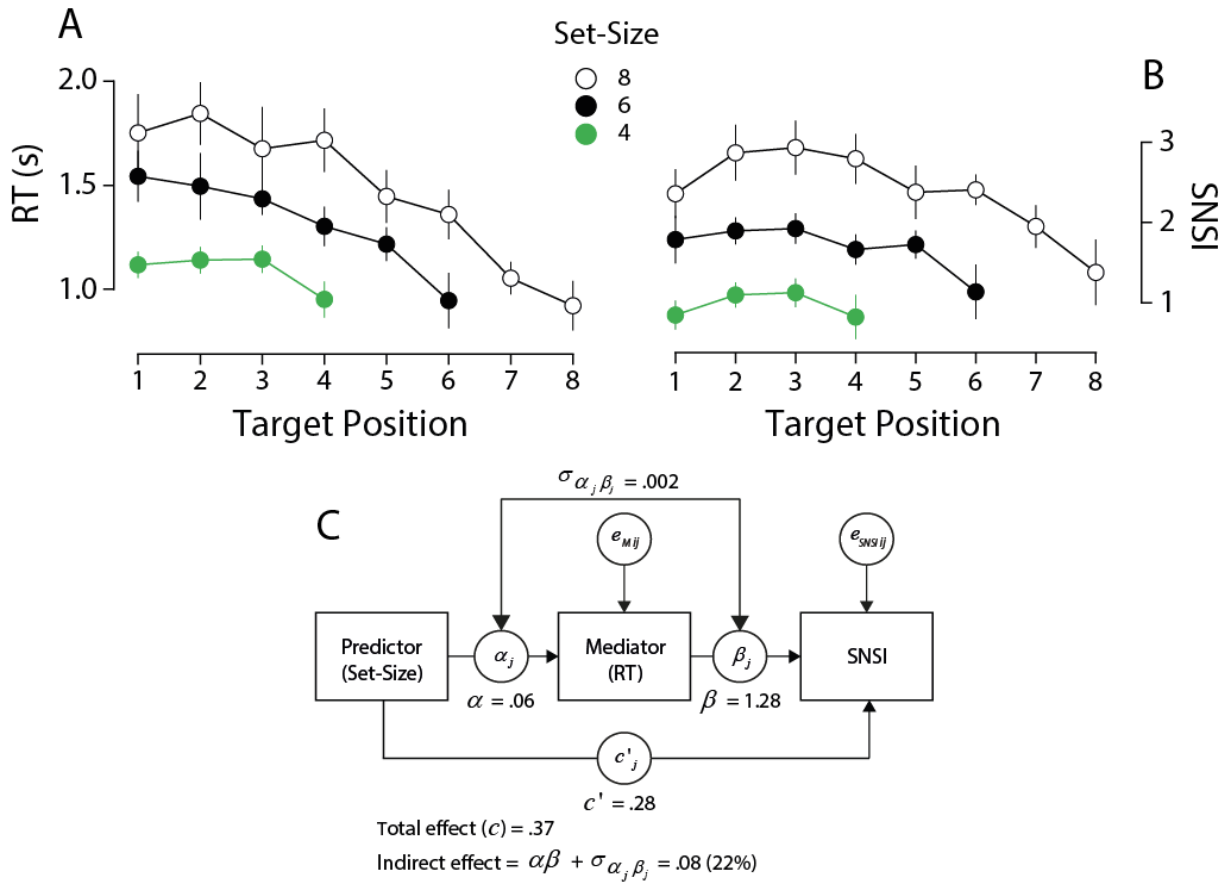


Figure 3. A) Response Time (RT) as a function of set-size and target position and B) SNSI as a function of the same factors. C) Multilevel mediation model, with set-size as predictor and RTs as mediator.

Discussion

In this second experiment, we switched to an item identification task, with a view to modify the cognitive processes while keeping the stimuli as close as possible to the ones used for the judgment of recency task. The results we obtained on first order performance are consistent with the literature (Sternberg, 1966, 1969); we found a decrease of performance with increasing list lengths, and similarly, a decrease with target positions at the beginning of the lists. Thus, from a strict behavioral perspective, these results are similar to those obtained on the JOR task, but previous studies suggest that they originate from very distinct underlying processes (McElree, 2006). As commented above, McElree and Dosher (1989) suggest a direct access mechanism to the information, which is supported by the evidence that both the amount of information necessary to correctly identify the target in this task, as well as the speed of recovery for each target position in each set-size, is statistically the same.

Unfortunately, from our experimental design it is not possible to confirm this pattern: In order to maintain as much similarity as possible between the experimental design of the Experiment 1 and 2, we did not incorporate a *speed - accuracy tradeoff* (SAT) procedure to evaluate this aspect, contrary to McElree and Dosher (1989) did.

Regarding introspection, participants reported (SNSI) not only fewer scanned items when the list of items was smaller, but also when the target was present closer to the end of the list. Interestingly, this pattern disappears when we excluded the last position. In line with the evidence discussed, the primming effect of the last item presented seem to generate an automatic recovery mechanism, probably not related to an introspective recovery information process from working memory, but an automatic recognition process of information still present in the focus of attention (Wickelgren, 1980) or in iconic memory (Sperling, 1960). Coherently, after controlling the last position for each set-size a completely different SNSI pattern was observed. Although there is still a clear difference in participants' introspection related to the perceptual load (set-size) - not totally explained by the participants' response time - the introspective description of the cognitive process was different from that reported in Experiment 1: the number of items scanned within list of 6 and 4 items was the same, regardless of the target position.

Finally, to confirm our interpretation we applied a direct comparison to the two experiments, on the subset of trials that use the same, 6 items, list length. We thus ran an LMM on participant responses with factor of task (First order: IES | Second order: SNSI), type of experiment (JOR | IR) and target position (from 1^{st}, to 5^{th} position, see Figure 4). The analyses considered all double and the triple interactions. This analysis did not take into account the last position on the list, for the reasons discussed above. Crucially, we found that the triple interaction was significant ($F(2,339.0) = 5.39$, $p < .01$). To make sure that this interaction could be interpreted as meaning that the difference between first and second order responses differs between the two tasks, we investigated the double interaction for each type of task (First order | Second order). Regarding the first order tasks (IES), we found a significant main effect of target position ($F(1,152.2) = 80.5$, $p < .001$) and type of experiment ($F(1,164.1) = 72.2$, $p < .001$). The interaction between these factors was also significant ($F(1,152.2) = 23.8$, $p < .001$), which was characterized by a steeper decrease of IES as a function of target position in JOR task ($F(1,53.0) = 48.9$, $\beta = -.47$, $p < .001$) than in IR task ($F(1,93.5) = 19.7$, $\beta = -.13$, $p < .001$). Regarding the second order task (SNSI), the interaction between type of experiment and target position ($F(1,113.5) = 4.14$, $p < .05$) presented a different pattern: SNSI significantly decreased in a JOR task as a function of target position

($F(1,45.0) = 10.05$, $\beta = -.17$, $p < .01$), but not in an IR task ($p > .40$). Both main effects were also significant (target position: $F(1,36.1) = 5.99$, $p < .05$; type of experiment: $F(1,132.0) = 53.7$, $p < .001$). In sum, our results suggest that participants' introspection was sensitive to a specific shift in the cognitive process generated by a simple change in the instruction in a memory recovery task.

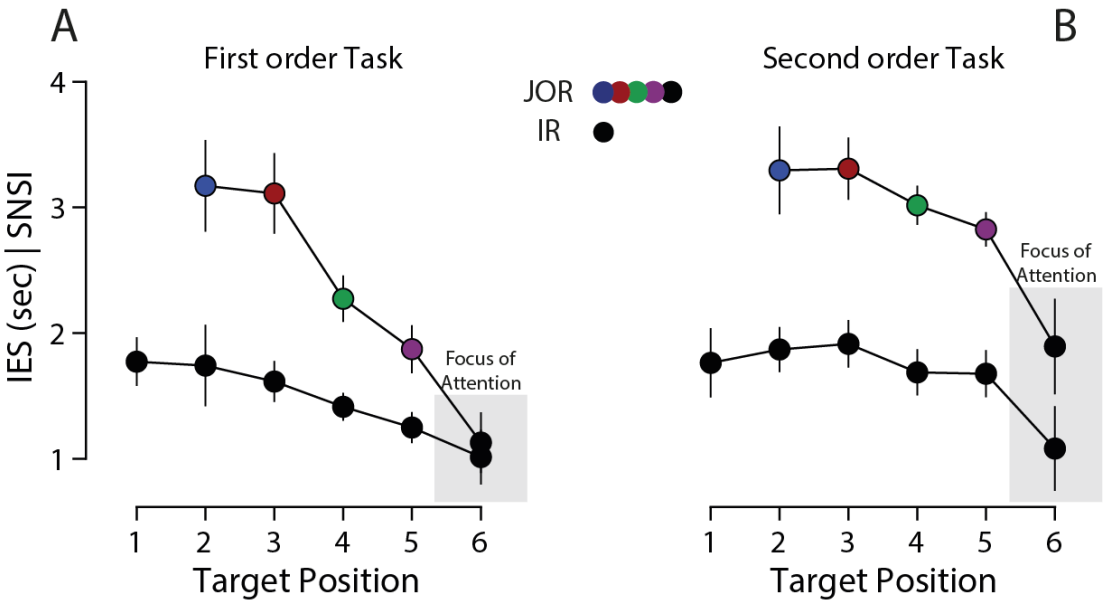


Figure 4. A) IES (First order Task) as a function of target position in Experiment 1 (JOR) and 2 (IR): Color dots indicate target position in JOR task, collapsing all distractor positions. Black dots indicate target position in IR task when the set-size was equal to 6 consonants. B) SNSI (Second order Task) as a function of target position in both experiments. The analyses did not take into account the last position on the list, illustrated in the figure as "Focus of Attention".

2.3.4.3 Conclusion

For the last 40 years the literature on metacognition has suggested that the introspective capacity has no access to higher order processes that underlie decision making. Nisbett and Wilson (1977) vindicated this idea from an extensive review of social psychology studies that showed that individuals often confabulate regarding the causes of their behavior. Even though this idea has been established as a canon in the literature (Overgaard & Sandberg, 2012) – which has received certain re-conceptualizations (Wilson, 2002, 2003; Wilson & Dunn, 2004) and certain recent experimental support (Johansson, Hall, Sikström, & Olsson, 2005; Johansson, Hall, Sikström, Tärning, & Lind, 2006) – as far as we know there is no evidence in experimental psychology that go in further details on the limits of introspection, with respect to the nature (process / state) of the targeted mental content. Above all, the question that arises is whether the limit of introspection is the result of a functional determinant of the cognitive system or the result of an – until now – inadequate context of experimentation. Our intuition was that, by focusing participants' attention on the mental content of interest, an idea present in the literature for many years (Flavell, 1979; Hurlburt & Heavey, 2001), and under the appropriate methodological safeguards (Piccinini, 2003; Goldman, 2004), introspection should be able to describe with some precision the nature of a cognitive process. Along these lines, we evaluated if introspection is capable to distinguish the cognitive process that underlies two working memory tasks. Previous experimental evidence (Sternberg, 1966, 1969; Muter, 1979; Hacker, 1980; McElree, 2006) suggests that by a simple instruction modification (item identity task or relative position task), it is possible to shift in the nature of the cognitive process deployed. Our aim was to assess if introspection was able to detect such changes.

Our experimental design is motivated by the *script-report* procedure (Jack & Roepstorff, 2002), where it is possible to contrast the objective first-order information with the second-order subjective report. To evaluate this, we used the Subjective Number of Scanned Items (SNSI), which we previously used in the context of visual search tasks (Reyes & Sackur, 2014). In Experiment 1 (JOR task), the SNSI was consistent with access to a serial mechanism of information retrieval: SNSI scores increased as a function of the target position, independent of the distractor position. This effect was not due to confounding factors. In Experiment 2 (IR task), the SNSI was consistent with a direct access mechanism: SNSI did not vary as a function of the target position (when excluding the last position, and for equal list length as in the first experiment). These results confirm that mental monitoring of cognitive processes is possible in these elementary memory tasks. These results are

consistent with our previous study (Reyes & Sackur, 2014), where we observed that the participants' introspection was successful in describing the nature of the cognitive process underlying a visual search task. In this line, the present study confirms that such idea might extend to the more abstract domain of short term memory.

One strength of the present study is that we could use identical stimuli, and change the nature of the processes by a redefinition of the task, whereas in our previous study we used perceptually different stimuli to generate either serial or parallel search processes. This had the unfortunate consequence that participants would know the nature of the experimental manipulation of interest, the one that we expected would cause distinct cognitive processes. Thus, they may have fallen into using this knowledge as an ingredient to their subjective reports, yielding in effect more interpretative (confabulatory) than really introspective reports. However, the correlated limitation of the present study is that the experimental conditions were studied in two different experiments, as it seemed difficult to change instructions on a trial-b-trial basis.

In sum, claims about intrinsic limits of introspection should be considered premature until introspection is given the proper focus. Recently, Marti, Bayet and Dehaene (2015) showed that participants could to some extent reproduce the trajectory of their attentional shifts in a serial visual search, which should be contrasted to the notion that saccades themselves seem largely non-conscious. We previously (Reyes & Sackur, 2014) showed that whether participants are able to subjectively access the nature of the process in a visual search depends on the time pressure imposed on the first order task: stronger time pressure shifts the introspective focus on the perceptual load, while low time pressure enables access to the search process itself. Here, we showed that differences of mental processes that were inferred indirectly from behavioral evidence are in fact directly accessible through introspection. Thus, recent empirical results should be taken into account in the full re-conceptualization of introspection that is underway in cognitive science.

References

Austen, E., & Enns, J. T. (2003). Change detection in an attended face depends on the expectation of the observer. *J. Vis., 3*, 64-74.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychol. Methods, 11,* 142-163.

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the Inverse Efficiency Score (IES) a better dependent variable than the mean Reaction Time (RT) and the Percentage of Errors (PE)? *Psychol. Belg., 51*, 5-13.

Bryce, D., & Bratzke, D. (2014). Introspective reports of reaction times in dual-tasks reflect experienced difficulty rather than timing of cognitive processes. *Conscious. Cogn.*, *27*, 254-267.

Chan, M., Ross, B., Earle, G., & Caplan, J. (2009). Precise instructions determine participants' memory search strategy in judgments of relative order in short lists. *Psychon. Bull. Rev., 16,* 945-951.

Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychol. Sci., 19,* 1110-1117.

Costall, A. (2006). 'Introspectionism' and the mythical origins of scientific psychology. *Conscious. Cogn., 15*, 634–654.

Cousineau, D. (2005).Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Quant. Meth. Psych., 1,* 42-45.

Dosher, B. (2003). Working memory. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science Vol. 4* (pp. 569-577). London: Nature Publishing Group.

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychol. Rev., 87,* 215-251.

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

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541-1543.

Goldman, A. (2004). Epistemology and the evidential status of introspective reports. *J. Conscious. Stud., 11,* 1-16.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *J. Exp. Psychol.- Learn. Mem. Cogn., 6,* 651-675.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: describing inner experience. *Trends Cogn. Sci., 5,* 400-403.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus-response to script-report. *Trends Cogn. Sci., 6,* 333-338.

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

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

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol., 59,* 193-224.

Kozuch, B., Nichols, S. (2011). Awareness of Unawareness: Folk Psychology and Introspective Awareness. *J. Conscious. Stud., 18,* 135-160.

Lyons, W. (1986). *The Disappearance of Introspection*, Cambridge, MA: MIT Press.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition, 115,* 303-313.

Marti, S., Bayet, L., & Dehaene, D. (2015). Subjective report of eye fixations during serial search. *Conscious Cogn, 33*, 1-15.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *J. Exp. Psychol.- Gen., 118,* 346-373.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *J. Exp. Psychol.- Gen., 122,* 291-315.

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol.- Learn. Mem. Cogn., 27,* 817-835.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 155-200). San Diego: Academic Press.

Miller, J., Vieweg, P., Kruize, N., & McLea, B. (2010). Subjective reports of stimulus, response, and decision times in speeded tasks: how accurate are decision time reports? *Conscious. Cogn., 19,* 1013-1036.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Quant. Meth. Psych., 4,* 61-64.

Muter, P. (1979). Response latencies in discriminations of recency. *J. Exp. Psychol.- Learn. Mem. Cogn., 5,* 160-169.

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

Overgaard, M. (2006). Introspection in science. *Conscious. Cogn., 15,* 629-633.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access, *Phil. Trans. R. Soc. B, 367,* 1287-1296.

Piccinini, G. (2003). Data from introspective reports: upgrading from common sense to science. *J. Conscious. Stud., 10,* 141-156.

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol. Rev., 117,* 864-901.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Commun. Methods Meas., 6*, 77-98.

Reyes, G., & Sackur, J. (2014). Introspection during visual Search. *Conscious Cogn., 29,* 212-229.

Selig, J. P., & Preacher, K. J. (2008). Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects [Computer software]. Retrieved from: http://quantpsy.org/.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychol. Rev., 85,* 355-362.

Sperling, G. (1960). The information available in brief visual presentations. *Psychol. Monogr., 74*, 1-29.

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science, 153*, 652-654.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *Am. Sci.*, *57*, 421-457.

Townsend, J. T., & Ashby, F. G. (1983). *The Stochastic Modeling of Elementary Psychological Processes*. Cambridge: Cambridge University Press.

Townsend, J. T., & Wenger, M. J. (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychon. Bull. Rev. 11*, 391-418.

White, P. A. (1980). Limitations on verbal reports of internal events: A Refutation of Nisbett and Wilson and of Bem. *Psychol. Rev., 87,* 105-112.

White, P. A. (1987). Causal report accuracy: retrospect and prospect. *J. Exp. Soc. Psychol., 23,* 311-315.

White, P. A. (1988). Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later. *Br. J. Psychol., 79,* 13-45.

Wickelgren, W.A., Corbett, A.T., & Dosher, B.A. (1980). Priming and retrieval from short-term memory: A speed-accuracy tradeoff analysis. *J. Verb. Learn. Verb. Behav., 19,* 387-404.

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

Wilson, T. D. (2003). Knowing when to ask. Introspection and the adaptive unconscious. *J. Conscious. Stud., 10,* 131-140.

Wilson, T. D., & Dunn, E. (2004). Self-Knowledge: its limits, value, and potential for improvement. *Annu. Rev. Psychol., 55*, 493-518.

2.4 Self-knowledge dim-out: Stress impairs metacognitive accuracy

Reyes, G., Silva, J., Jaramillo, K., Rehbein, L., Sackur, J. (2015). Self-knowledge dim-out: Stress impairs metacognitive accuracy. *PLoS ONE* 10(8): e0132320.

2.4.1 Introduction

The relation between stress and an alteration of the cognitive functions is well established in literature (Lupien et al, 2007; Starcke & Brand, 2012). In particular, it has been observed that the presence of stress alters the decision-making process and also the processing of associated feedback (Porcelli & Delgado, 2009; Schwabe & Wolf, 2011; Otto et al., 2013). In addition, recent studies (Qin et al., 2010; Hermans et al., 2014) suggest that acute stress dampens activity in regions sub-serving endogenous attention (dorsolateral and medial PFC), in favor of orienting resources to vigilance and action, that is, to exogenous attention. Interestingly, the areas affected by stress are precisely those related to meta-cognitive sensibility (Fleming et al., 2010; Rounis et al., 2010; Yokoyama et al., 2010; Baird et al., 2013; de Martino et al., 2013; Fleming, et al., 2012). As a consequence, this study investigates whether biological reactivity to stress is related to the degree of accuracy with which individuals access their mental states. To this aim, three groups of high, medium and low stress responders were constituted based on the cortisol concentration in saliva in response to a standardized psycho-social stress challenge (the Trier Social Stress Test, TSST). We then assessed the accuracy of participants' confidence judgments in a perceptual decision task. The prediction is that participants under stress know less when they know and when they do not know.

2.4.2 Results

We found a decrease in metacognitive sensibility in the group with high biological reactivity to stress. At the same time neither the (almost null) variability in the first-order measures (RTs, accuracy, contrast) nor demographic factors (gender, age) explain the effect of stress in metacognition.

2.4.3 Discussion

Recent studies suggest the presence of inter-individual differences in the metacognitive capacity (Fleming et al., 2010). Along these lines, we propose that a relevant aspect in the origin and consolidation of such differences is the biological reactivity to stress. Our results suggest the presence of a new structural factor that determines the accuracy of the participants' introspection.

2.4.4 Paper

Self-knowledge dim-out: Stress impairs metacognitive accuracy

Gabriel Reyes [1,2,3*], Jaime R. Silva [4], Karina Jaramillo [4], Lucio Rehbein [5], Jérôme Sackur [1*]

1. Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), PSL Research University, Paris, France
2. Université Pierre et Marie Curie, Paris, France.
3. Escuela de Psicología, Universidad Austral de Chile, Valdivia, Chile.
4. Centro de Apego y Regulación Emocional, Universidad del Desarrollo, Santiago, Chile.
5. Departamento de Psicología, Universidad de La Frontera, Temuco, Chile.

[*]Corresponding author: gureyes@uc.cl (GR); jerome.sackur@gmail.com (JS).

Abstract

Modulation of frontal lobes activity is believed to be an important pathway trough which the hypothalamic-pituitary-adrenal (HPA) axis stress response impacts cognitive and emotional functioning. Here, we investigate the effects of stress on metacognition, which is the ability to monitor and control one's own cognition. As the frontal lobes have been shown to play a critical role in metacognition, we predicted that under activation of the HPA axis, participants should be less accurate in the assessment of their own performances in a perceptual decision task, irrespective of the effect of stress on the first order perceptual decision itself. To test this prediction, we constituted three groups of high, medium and low stress responders based on cortisol concentration in saliva in response to a standardized psycho-social stress challenge (the Trier Social Stress Test). We then assessed the accuracy of participants' confidence judgments in a visual discrimination task. As predicted, we found that high biological reactivity to stress correlates with lower sensitivity in metacognition. In sum, participants under stress know less when they know and when they do not know.

Keywords: Metacognition, Stress reactivity, Trier Social Stress Test (TSST), Cortisol.

Introduction

Acute stress is associated with altered cognitive functioning, in particular with respect to decision making [1,2]. Under stress, individuals exhibit less flexible cognitive processing [3] together with altered risk and feedback processing [4,5]. Collectively, these effects suggest that stress taxes executive functions [2,6], and thus they point to the potential impact of stress on the regulation and monitoring of decision processes. Indeed, decision is not only about selecting the right option; it is also about the assessing its appropriateness relative to the circumstances, that is, whether one can be confident or not about one's action. Sound confidence judgments are essential for the decision maker both to make behavioral adjustments and to efficiently cooperate with others [7,8]. Optimal confidence judgments should be calibrated (reflect decision performance in the long run), and sensitive (discriminate correct from incorrect decisions). In this study, we focus on the impact of stress on the sensitivity of confidence judgments, also termed metacognitive accuracy.

The stress response consists in a cascade of mechanisms governed by the Hypothalamic-Pituitary-Adrenocortical (HPA) axis, leading notably to the release of cortisol and catecholamines. Within the brain, these hormones are known to target specifically the prefrontal cortex (PFC) [6,9], thereby altering higher cognitive functions. Of considerable interest regarding metacognition, it has been proposed [10,11] that one of the early effects of acute stress is to dampen activity in regions subserving endogenous attention (dorsolateral and medial PFC), in favor of orienting resources to vigilance and action, that is, to exogenous attention.

While the cognitive mechanisms behind confidence judgments are not fully understood, they certainly involve orienting attention inwards and require flexible processing. Most importantly, it has repeatedly been shown that the rostral and dorsal aspect of the lateral PFC is the neural basis for metacognitive accuracy, where the neuroendocrine impact of stress on the brain is thought to be maximal [12,13,14,15,16,17].

Furthermore, it has recently been shown [18] that cortisol release under acute stress reduces the depth of strategizing in the beauty contest game. In this behavioral economics game, participants are tested on their ability to reason about other participants' reasoning. Under stress, participants seem to have shallower strategic mind reading, and we expected that the same might be true for metacognition, as a form of intra-individual mind

reading. Thus, converging neurophysiological and cognitive evidence suggest that stress reactivity, and more precisely cortisol release, should lead to alterations in metacognition.

To test this hypothesis, we first designed a first session during which three groups of participants (high, medium and low responders) were constituted, according to the concentration of cortisol in saliva at the peak of hormonal response to interpersonal stress. In a later second session (12 month after session 1), in order to avoid direct contextual effects of stressors, participants performed a perceptual decision task with confidence judgments. We operationalized metacognitive sensitivity as the extent to which confidence judgments discriminate correct and incorrect responses. We predicted that high responders should have lower scores on this measure. Between session 1 and 2 a pilot exploratory study was run (8 months after session 1 and 4 month before session 2) to evaluate general differences in metacognition according to individual cortisol reactivity to stress. Results and data are available at https://osf.io/6zsap/?view_only=39fb68df24094788b9daffcb0fe5a683

2.4.4.1 Experiment

Materials and Methods

Session 1: stress screening

120 students (mean age: 20.6, *SD*: 1.5; 58 women) from the Universidad de La Frontera (Chile) participated in the study. These 120 participants are common both for the pilot study, as well as for the main experiment. Participants gave their written informed consent (Declaration of Helsinki). The study at all stages was approved by the ethics committee of Universidad de La Frontera. Participants were compensated the equivalent of about $20 for the 2 hour session. Experimental sessions were scheduled from 2:30 to 6:30 pm. Exclusion criteria were: a body mass index < 18 or > 30 kg/m²; receiving medical treatment known to affect the HPA axis; a history of psychiatric or neurological disorders; abnormal vision; smoking; pregnant or lactating women, and women taking oral contraceptives. Participants were asked not to eat or brush their teeth one hour before the session, and not to drink alcohol or play sports the day before.

We applied the Trier Social Stress Test (TSST) [19], with seven interspersed saliva samples to assay cortisol concentration (Fig. 1A). The TSST asks that participants perform a ten minutes, videotaped oral presentation and do mental arithmetic in front of a non-supportive panel of three judges. The first saliva sample was taken after a ten minutes rest, immediately followed by the State-Trait Anxiety Inventory (STAI) [20]. Participants then performed the main presentation of the TSST, after which the second saliva sample was taken and participants filled out a second STAI with only the state subscale. Six more saliva samples were taken during a post-exposure rest phase to control normal decrease in the stress reactivity curves. Heart rate (HR) was monitored throughout. Here as well as in the second session, we also recorded electrodermal activity, but technical failures precluded our using the data. Electrodes were located on the medial side of both ankles and the distal anterior aspect of the right forearm (BIOPAC MP150, Goleta, CA). Saliva samples were sent to the Molecular Biology Laboratory of the Universidad de La Frontera, for quantitative determination of cortisol concentration. Salivary concentrations of cortisol were obtained using an enzyme immunoassay commercial kit following the manufacturer's instructions (DRG Salivary Control ELISA Kit, DRG Instruments GmbH, Germany). Concentrations were obtained by interpolating from a standard curve plotted using the software GraphPad Prism version 5.0 (GraphPad Software, San Diego CA, USA).

Session 2: confidence accuracy experiment

Based on the results of session 1, we formed three groups of 9 participants with high, medium and low stress reactivity (mean age: 20.1, *SD*: 1.2; 16 women: 5 in the low, 6 in the medium and 5 in the high group), which took part in the experiment one year after session 1. 10 participants who participated in the pilot study were included (3-Low, 3-Medium, 4-High). Participants gave their informed consent and were compensated the equivalent of about $10 for a 1 hour session. Stimuli were arrays of six vertical gabor patches (2.8° in diameter, spatial frequency of 2.2 cycles per visual degree, Fig. 1B) on a uniform grey background (luminance: 44.1 cd/m²), presented on an imaginary circle (6.2°) at the center of a CRT screen (size 17", resolution of 1024 X 768 pixels, refresh rate of 100 Hz, viewing distance ~55 cm). Participants were tested in a darkened room.

After a fixation spot (500 *ms*) participants viewed two arrays for 200 *ms* each, separated by an interval of 300 *m*s. In one of the arrays, one random gabor patch had a higher contrast. Participants had to decide in which interval the contrasted gabor was presented (two intervals forced choice task, 2IFC) during a 2000 *ms* response window, by pressing the "Q" (first interval) or "W" (second interval) key on a standard QWERTY keyboard. Participants first went through a calibration stage (120 trials), during which we adapted the target contrast with 1-up 2-down interleaved staircases [21], so as to reach similar accuracy levels for all participants (accuracy at the end of the calibration: 75%, which did not significantly differ from the targeted 71% accuracy). The threshold contrast for each participant was calculated as the average of the last six reversals of both the initially ascending and descending staircases.

During the main experiment, contrast was fixed at the participant's adapted level. Immediately after the 2IFC response, participants were instructed to give the best estimate of their confidence about their decision, by means of a visual analog scale for confidence. We used a half range scale with the labels "*guess*" and "*absolutely certain*" at the left and right ends, which we scored as a probability scale ranging from .5 to 1. The experimental session comprised 320 trials in 8 blocks with a 60 second pause between each block. Before and after the main task participants filled out the STAI-S (state subscale). Heart rate was monitored throughout. As a way to reactivate participants' stress response, a mild interpersonal stressor in the middle of the task was introduced: Immediately after the fourth block, an error message appeared on screen. The experimenter came into the

experimental room with a neutral emotional expression and stated that "*there was a problem with your responses*", without saying anything about participants' performance.

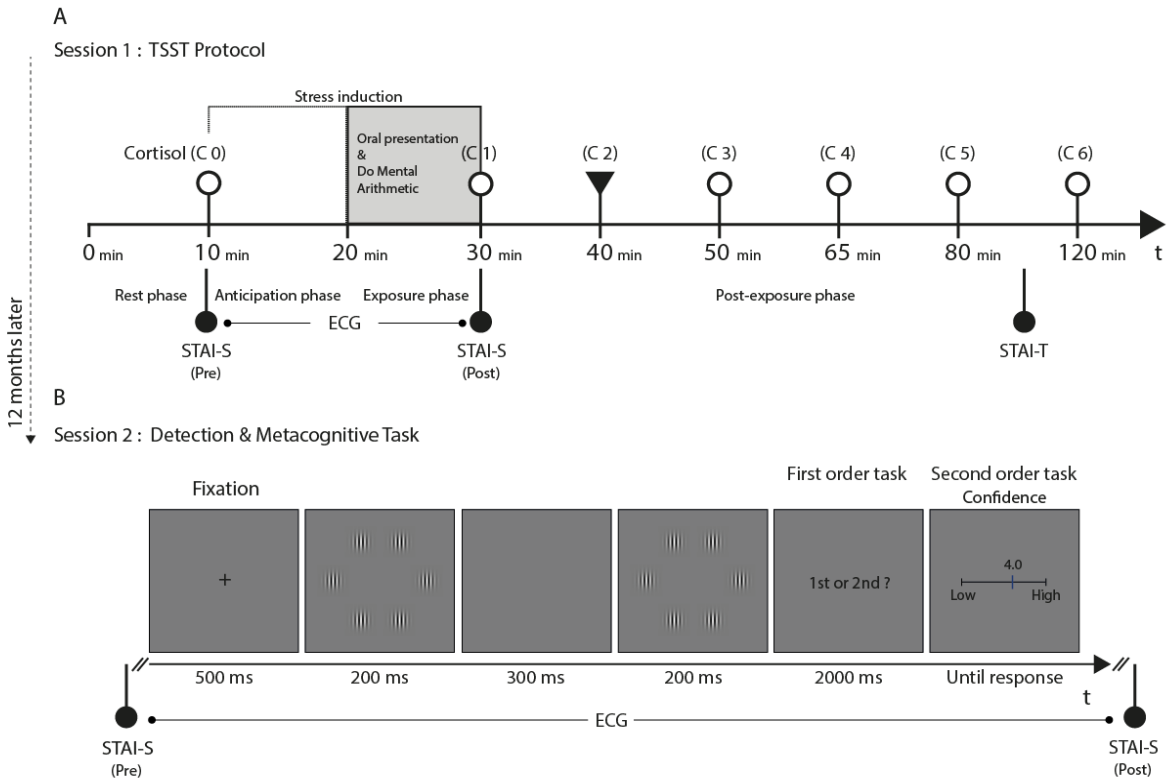


Fig. 1. (A) TSST protocol (*session 1*). (B) Detection and Metacognitive tasks (*session 2*).

Results

Session 1

Cortisol level at C2 (10 minutes after stress induction) was taken as indicative of participants' reactivity to stress [22]. Participants above the 75th percentile were categorized as having high reactivity (mean: 10.7 nmol/l; *SD*: 1.4; range: 9.6, 14.2; Cortisol increase (C2-C0 or Baseline) = 2.14), participants below the 25th percentile as low reactivity (mean: 4.5 nmol/l; *SD*: .79; range: 2.9, 5.7; Cortisol increase = .21) and participants around the 50th percentile as medium reactivity (mean: 7.3 nmol/l; *SD*: .95; range: 6.3, 9.3; Cortisol increase = 1.45; Fig. 2A-B). Classifications reached the criteria for distinguishing cortisol responders from non responders (threshold for cortisol responder: 1.1 nmol/l) [23]. We applied the conversion provided by Miller et al. [24], available at http://psylux.psych.tu-dresden.de/i1/biopsych/ac.html for the DRG immunoassay kit.

We randomly selected 9 participants out of each the three above percentile groups. Studies interested in cortisol responders use a similar or smaller numbers of participants (for instance, [25] (20 participants); [26] (20 participants); [27] (20 participants); [28] (26 participants), etc., see many others in [22]). In addition, it is important to note that previous studies select responders and non-responders *ex-post*, which often yield an imbalance between groups. As opposed to that, our selection of responders is done *ex-ante*, ensuring a clean statistical framework for establishing any effect of stress reactivity on the dependent variables.

These groups did not differ with respect to gender ($p > .73$), age ($p > .50$) or stress personality traits (STAI-Trait, $p > .39$). The TSST procedure increased psychological acute stress (Fig. 2C): STAI-State was 11.3 point higher after TSST than before ($F(1,24) = 44.0$, $p < .001$, $\eta^2_p = .65$) without difference between groups ($p > .16$) and no interaction ($p > .46$). Mean heart rate (HR) did not differ across groups ($p > .11$), however hear rate variability (HRV), which is indicative of activation of the stress response [29] as computed by standard deviation, was significantly different ($F(2,24) = 4.33$, $p < .05$, $\eta^2_p = .27$). More precisely, collapsing medium and low stress reactivity groups, we found that high reactivity participants had a higher HRV (16.8 beats/min) (low and medium stress participants: 14.0 beats/min; $F(1,25) = 4.46$, $p < .05$, $\eta^2_p = .15$).

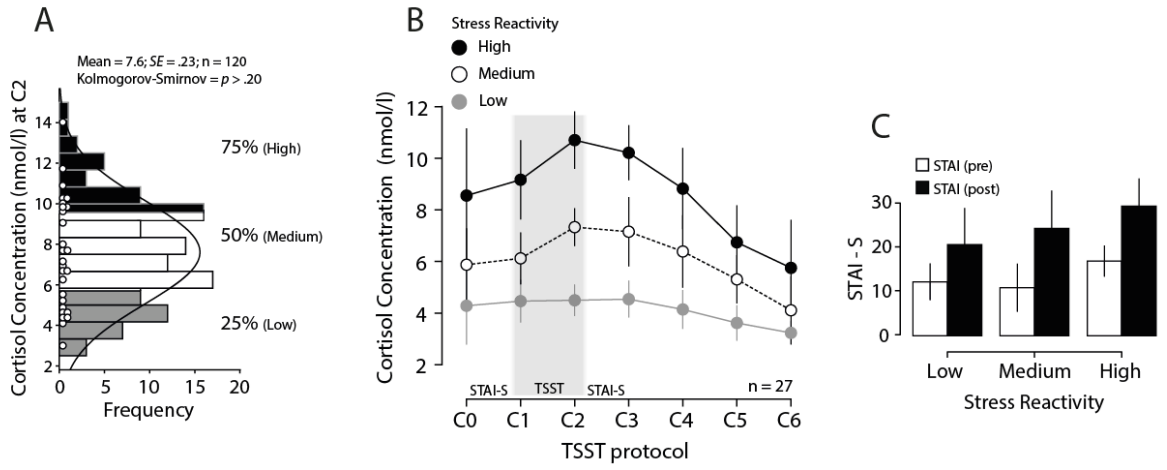


Fig. 2. Session 1: Stress screening. A) Histogram of cortisol concentration in saliva at C2, with a superimposed normal fit. Each dot represents a participant tested in session 2. B) Cortisol concentration in saliva during the TSST protocol for each experimental group. Error bars are 95% confidence intervals. C) State-Trait Anxiety Inventory (STAI-S), before and after stress induction. Error bars are 2+/- *SE*.

Session 2

Stress measures

The experimental context reactivated the stress responses specific to each three groups: Heart rate variability (HRV) was higher during the first half of the experiment for high reactivity participants (7.31 beats/min) than low and medium reactivity participants (6.2 beats/min; $F(1,25) = 5.78$, $p < .05$, $\eta^2_p = .19$). This difference vanished when we took into account the second half of the experiment, but we found a peak of HRV at the time of the interpersonal stressor, only for the high reactivity group (see Supporting Information and S1 Fig.). Psychological stress (STAI-State) increased during the experiment (before / after difference of 7.62 points: $F(1,24) = 19.1$, $p < .001$, $\eta^2_p = .44$; Fig. 3A), without difference between groups ($p > .42$) and no interaction ($p > .65$).

Performance: accuracy and response times

We excluded trials with response times (RTs) below 200 *ms* and above 2000 *ms* (4.3%). Crucially, there were no differences on first order performance between stress groups: First, the contrast needed to achieve the targeted performance at the end the staircase calibration procedure did not differ across stress groups ($p > .40$). During the main experiment, accuracy increased with respect to the end of the calibration (75% correct), but it was stable at 83% (Fig. 3B) and did not differ between groups ($p > .35$) or across blocks ($p > .93$) and these factors did not interact ($p > .39$). RTs were marginally slower for low reactivity participants ($p > .06$; Low: 818 *ms*; Medium: 699 *ms*; High: 692 *ms*) with a significant learning effect ($F(3.6,88.5) = 5.16$, $p < .01$, $\eta^2_p = .17$), but without interaction between these factors ($p > .23$, Fig. 3B).

Confidence

Mean confidence was .84 and did not differ across stress groups ($p > .91$), meaning first that participants were well calibrated in view of their average performance of 83%, and second, that stress reactivity did not translate into under - or over - confidence. Now, our critical construct was metacognitive sensitivity, or resolution on the confidence scale. We quantified metacognitive sensitivity by the area under the curve (AUC) for the type-2

Receiver Operating Characteristic (ROC), calculated separately for each participant (Fig. 3C). This measure consists in plotting one against the other the cumulative proportions of correct and incorrect responses for increasingly liberal confidence percentiles [30]. Type-2 AUC is a model-free, empirical estimate of metacognitive accuracy, which has been used in previous studies [12]. Higher AUC scores denote a stronger link between confidence and accuracy. We relied on this empirical metric for the estimation of metacognitive sensitivity, given that there is currently no consensus on the appropriate signal theoretic models for confidence judgments [31,32].

In order to compute the type-2 ROCs, for each participant we binned the responses on the continuous confidence scale in five equal bins, and we plotted the cumulative proportions of correct against incorrect responses, anchored at 0 and 1 for increasingly liberal bins. We found that higher reactivity was associated with lower AUCs, that is with poorer metacognitive sensitivity (High: .68; Medium: .73; Low: .78). The effect was statistically significant with a one-way ANOVA on AUC with stress reactivity as factor and participants as random effect ($F(2,26) = 5.86$, $p < .01$, $\eta^2 = .33$). Post-hoc comparison only showed a difference between low and high reactivity participants (Δ mean= .11, $SE = .03$, $p < .01$, Bonferroni corrected). We next regressed participants' AUCs on cortisol concentrations at C2, and found a significant negative correlation ($r^2 = .34$, $\beta = -.58$, $t(26) = -3.57$, $p < .001$; Fig. 3D). Importantly, we checked that AUCs were not related to performance ($p > .98$) or contrast ($p > .16$) or RTs ($p>.32$, see Supporting Information for a more detailed account of RTs). Finally, there was no difference in the response times for confidence between stress reactivity groups (Low: 1258 *ms*; Medium: 1376 *ms*; High: 1165 *ms*, $p > .56$).

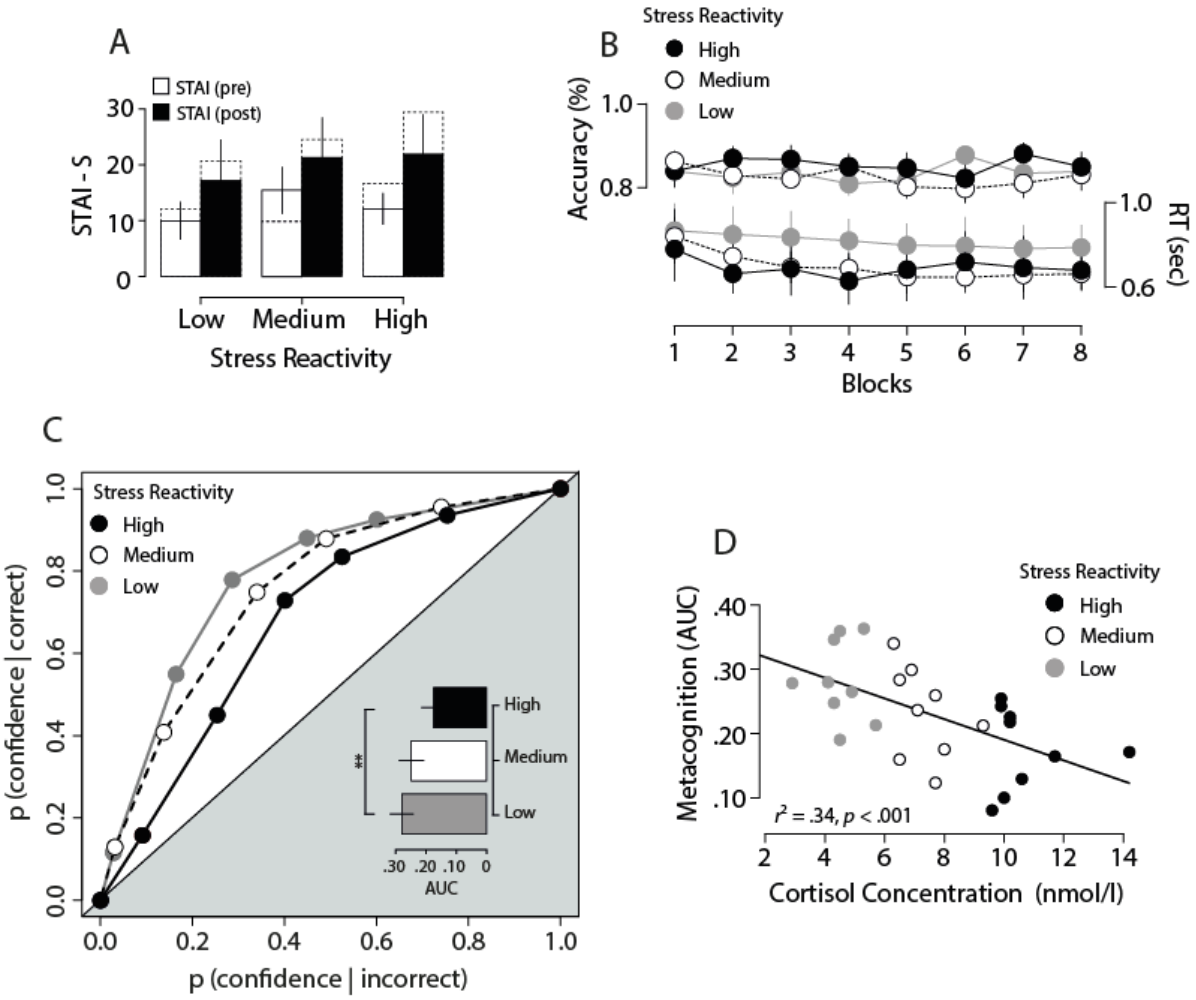


Fig. 3. Session 2: Confidence accuracy. A) State-Trait Anxiety Inventory (STAI-S), before and after detection task. Error bars denote 2+/- *SE*. Dashed lines behind each bar represent STAI before and after stress induction during TSST protocol (Session 1). B) Response Time and Accuracy as a function of blocks for each stress group. Error bars denote 95% confidence intervals. C) Metacognitive sensitivity was quantified by area under the type-2 ROC for each stress group. D) Linear regression of AUC scores on cortisol concentration in saliva at C2. Each dot represents one participant.

2.4.4.2 Conclusion

We investigated the impact of stress on metacognition. We identified participants with high, medium and low biological reactivity to an interpersonal challenge. High responders were subsequently found to have lower metacognitive accuracy in a perceptual decision task, while no difference was observed on perceptual decision performances or confidence calibration. We insist on the fact that first order performances were equated across participants with an adaptive staircase method, so much so that the decrement in metacognitive sensitivity cannot simply be an indirect consequence of the impact of stress on first order cognition [33]. Thus, even though stress had no effect on first order cognition, and did not create under or over-confidence biases, it decreased metacognitive accuracy. Stress induces a relative blindness to one's cognitive performance.

Contrary to many stress studies, we did not include an explicit stressor prior to the main cognitive task: We relied on the experimental context, combined with individual reactivity, to modulate stress. While this design eliminates contextual cues, it opens the possibility that the observed decrease in metacognitive sensitivity be due to concomitant personality trait differences, rather than to acute stress. However, this interpretation is not favored by our data: First, we selected our three groups based on their cortisol reactivity, and this was not associated with traits differences, as assessed by the STAI-T. Second, the high reactivity group, although defined by peak reactivity at the moment of the interpersonal challenge, already shows an elevated cortisol response at the moment of the first cortisol sample, while it went back to baseline after 2 hours (Fig. 2B). Therefore, it seems that for these participants, the mere fact of walking in the laboratory is a stressor, and we may expect that this should be the case at the occasion of the main cognitive task. Indeed, heart rate variability is higher in the high reactivity group during the main task, which gives us a direct sign of stress reactivity. Therefore, it seems more parsimonious to suppose that the impairment in metacognition is due to acute stress activation.

Now, one could argue that the impact of stress reactivity on metacognition is not specific. Indeed, one could argue that the observed effect of stress on metacognition is one among the many effects of stress on executive functioning. While it is clear from the literature that stress does have an impact on executive functions [1,2,3,4,5,6], one should note that in our experiment, metacognitive accuracy is the only outcome variable that shows an effect of stress reactivity. Although we cannot rule out the notion that our perceptual discrimination task is not sensitive enough to elicit such effects, it seems unlikely as first order tasks have consistently

demonstrated their higher sensitivity than second order tasks [30]. Thus, if metacognitive decrements were due to a general impairment of cognitive control, it would most probably first be apparent in a change in performance for the first order task.

Now, we see two broad classes of mechanisms through which acute stress would yield such impairment in metacognitive sensitivity: First, it might be that acute stress is associated with intercurrent ruminative thoughts that could interfere with metacognitive processing ("high level" account). These thoughts might potentially be accessible and controlled by participants. Alternatively, it might be that higher circulating cortisol has a direct impact on prefrontal cortex regions that subserve metacognition for decisions ("low level" account). The detrimental effect of stress on metacognition might then be cognitively impenetrable. Note also that cortisol might not be the unique cause of the modifications of prefrontal cortex functioning, as it has been proposed that the cognitive effects of stress are due to the conjoint release of cortisol and catecholamines [6].

Our study adds to the growing literature on dissociations of first and second order performance [12, 31, 32, 34, 35]. Results suggest that first order performance and its subjective evaluation might rely on distinct cognitive mechanisms and information. One promising line of research in this respect would be to study the differential role of working memory in the formation of confidence judgments [34,35]. Indeed, while the first order task is performed immediately after presentation of the stimulus, the second order confidence judgment is performed later and requires the retention of both stimuli and of the participant's own response for a longer time. Thus it might be that stress reactivity affects metacognition through its impact on working memory. If that were the case, we should find that high working memory capacity would protect against stress induced decrease in metacognitive sensitivity, as was recently found for stress induced decrease in model based learning [3].

Additional studies are needed to fully unravel the mechanisms and the generalization of metacognitive impairment under stress. First, the high and low level accounts sketched above, and the working memory hypotheses are not exclusive, and we need to assay their relative weights. Second, recent studies have highlighted the disunity of metacognition [36]. With this in mind, it is of paramount theoretical, clinical and applied importance to determine whether our results hold beyond perceptual decision, that is, whether, for instance, metacognition in general knowledge or memory tasks would similarly be impaired under stress.

Acknowledgments

We thank Sid Kouider, Vincent de Gardelle, Mark Wexler and Franck Ramus for useful discussions. We thank also the editor and two anonymous reviewers for their constructive comments.

References

1. Lupien SJ, Maheu F, Tu M, Fiocco A, Schramek TE. The effects of Stress and Stress Hormones on human cognition: Implications for the field of brain and cognition. Brain Cogn. 2007;65, 209–237.

2. Starcke K, Brand M. Decision making under stress: a selective review. Neurosci. Biobehav. Rev. 2012;36, 1228–1248.

3. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working memory capacity protects model-based learning from stress. PNAS. 2013;110, 20941– 20946.

4. Porcelli A, Delgado M. Acute stress modulates risk taking in financial decision making. Psychol. Sci. 2009;20, 278–283.

5. Schwabe L, Wolf OT. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. Behav. Brain Res. 2011;219, 321–328.

6. Arnsten AF. Stress signaling pathways that impair prefrontal cortex structure and function. Nat Rev Neurosci. 2009;10, 410–422.

7. Frith CD. The role of metacognition in human social interactions. Phil. Trans. R. Soc. B. 2012;367, 2213–2223.

8. Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD. Supra-personal cognitive control and metacognition. Trends Cogn. Sci. 2014;18, 189–196.

9. Arnsten AF. Catecholamine modulation of prefrontal cortical cognitive function. Trends Cogn. Sci. 1998;2, 436–447.

10. Qin S, Hermans EJ, Van Marle HJ, Luo J, Fernandez G. Acute psychological stress reduces working memory-related activity in the dorsolateral prefrontal cortex. Biol. Psychiatry. 2010;66, 25–32.

11. Hermans EJ, Henckens MJ, Joëls M, Fernández G. Dynamic adaptation of large-scale brain networks in response to acute stressors. Trends Neurosci. 2014;37, 304–314.

12. Fleming SM., Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010;329, 1541–1543.

13. Rounis E, Maniscalco B, Rothwell J, Passingham R, Lau H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn. Neurosci. 2010;1, 165–175.

14. Yokoyama O, Miura N, Watanabe J, Takemoto A, Uchida S, Sugiura M, et al. Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. Neurosci Res. 2010;68, 199–206.

15. Baird B, Smallwood J, Gorgolewski KJ, Margulies DS. Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. J. Neurosci. 2013;33, 16657–16665.

16. de Martino B, Fleming SM, Garrett N, Dolan RJ. Confidence in value-based choice. Nat. Neurosci. 2013;16, 105–110.

17. Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. Phil. Trans. R. Soc. B. 2012;367, 1280–1286.

18. Leder J, Häusser JA, Mojzisch A. Stress and strategic decision-making in the beauty contest game. Psychoneuroendocrinology. 2013;38, 1503–151.

19. Kirschbaum C, Pirke KM, Hellhammer DH. The Trier Social Stress Test: A tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology. 1993;28, 76–81.

20. Spielberger CD, Gorsuch RL, Lushene RE. The State-Trait Anxiety Inventory: Test manual. Palo Alto, CA: Consulting Psychologist Press; 1970

21. García-Pérez MA. Forced-choice staircases with fixed stepsizes: asymptotic and small-sample properties. Vision Res. 1998;38, 1861–1881.

22. Allen AP, Kennedy PJ, Cryan JF, Dinan TG, Clarke G. Biological and psychological markers of stress in humans: focus on the Trier Social Stress Test. Neurosci Biobehav Rev. 2014;38, 94–124.

23. Miller R, Plessow F, Kirschbaum C, Stalder T. Classification criteria for distinguishing cortisol responders from nonresponders to psychosocial stress: evaluation of salivary cortisol pulse detection in panel designs. Psychosom. Med. 2013;75, 832–840.

24. Miller R, Plessow F, Rauh M, Gröschl M, Kirschbaum C. Comparison of salivary cortisol as measured by different immunoassays and tandem mass spectrometry. Psychoneuroendocrinology. 2013;38, 50–57.

25. Kirschbaum C, Prüssner JC, Stone AA, Federenko I, Gaab J, Lintz D, et al. Persistent high cortisol responses to repeated psychological stress in a subpopulation of healthy men. Psychosom. Med. 1995;57, 468–474.

26. Nater UM, Moor C, Okere U, Stallkamp R, Martin M, Ehlert U, et al. Performance on a declarative memory task is better in high than low cortisol responders to psychosocial stress. Psychoneuroendocrinology. 2007;32, 758–763.

27. Roelofs K, Bakvis P, Hermans EJ, van Pelt J, van Honk J. The effects of social stress and cortisol responses on the preconscious selective attention to social threat. Biol. Psychol. 2007;75, 1–7.

28. Engert V, Efanov SI, Duchesne A, Vogel S, Corbo V, Pruessner JC. Differentiating anticipatory from reactive cortisol responses to psychosocial stress. Psychoneuroendocrinology. 2013;38, 1328–1337.

29. Xhyheri B, Manfrini O, Mazzolini M, Pizzi C, Bugiardini R. Heart rate variability today. Prog. Cardiovasc. Dis. 2012;55, 321–331.

30. Fleming SM, Lau H. How to measure metacognition. Front. Hum. Neurosci. 2014;8, 1–9.

31. Scott RB, Dienes Z, Barret AB, Bor D, Seth AK. Blind insight: metacognitive discrimination despite chance task performance. Psychol. Sci. 2014; 25, 2199–2208.

32. Jachs B, Blanco MJ. On the independence of visual awareness and metacognition: a signal detection theoretic analysis. J. Exp. Psychol. 2015;41, 269–276.

33. Galvin SJ, Podd JV, Drga V, Whitmore, J. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychon. Bull. Rev. 2003;10, 843–876.

34. Bona S, Cattaneo Z, Vecchi T, Soto D, Silvanto J. Metacognition of visual short-term memory: dissociation between objective and subjective components of VSTM. Front. Psychol. 2013; 4, 1–6.

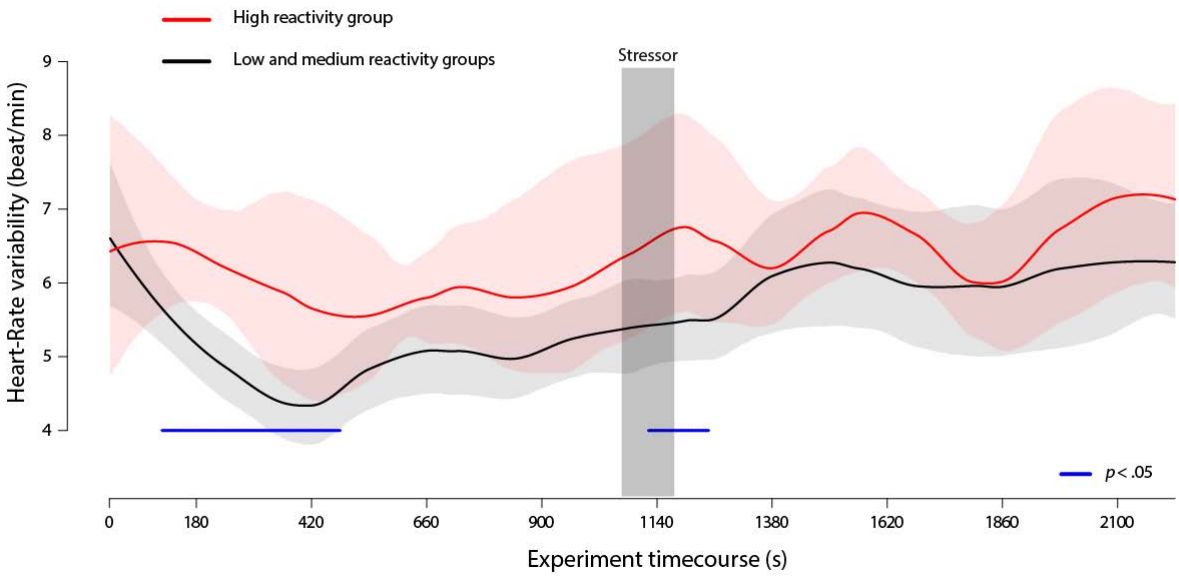
35. Bona S, Silvanto J. Accuracy and confidence of visual short-term memory do not go hand-in-hand: behavioral and neural dissociations. PLoS ONE. 2014;9, 1–10.

36. Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain. 2014;137, 2811–2822.

Supporting Information

Supplementary Figures

For each participant, we computed the standard deviation of heart rate in non-overlapping windows of 1 minute throughout the duration of the experiment, and smoothed them with locally weighted polynomial regression (LOESS, 1). The black and red curves are the means of the smoothed heart rate standard deviations for the low and medium reactivity groups on the one hand and for the high reactivity group on the other (S1 Fig). Shaded areas are bootstrapped 95% confidence bands for the two groups (5000 samples). Dark blue segments indicate bootstrapped $p < .05$ for the null hypothesis of no difference in heart rate standard deviations between the two groups. Two clusters of significant differences emerge: one at the beginning of the experiment (during the first 8 minutes) and one at the start of the second half of the experiment, just after stress reactivation (gray area).



S1 Fig. Heart rate variability during the main experiment.

Supplementary Analyses

Response times on the first order task (2IFC) and mediation analysis

As the only difference on performance on the 2IFC task was that participants in the low stress groups had marginally significant longer response times, we wanted to analyze this trend in greater details with a view to

gauge whether response times could in part explain the observed differences in metacognitive accuracy found between stress groups.

First, we analyzed response time distributions as generated by a sequential sampling process (2), so as to so as to extract, for each participant, decision and non decision components of the response. We used the Fast-DM toolbox (3), to fit the diffusion model of Ratcliff (4), for each participant individually, assuming no bias. Notice that since contrast is fixed for each participant, there is only one experimental condition for the set of 320 trials. This yielded the following six parameters for each participant: the drift rate (v), the separations of the boundaries (a), the non-decision time (t), and the variability for each of these. We submitted each of these parameters to separate one-way ANOVAs with stress group as a factor and participant as random effect. The only significant effect was on non-decision times (all other ps > .19): non-decision time was longer for the low stress group (409 *ms*) than for the medium (281 *ms*) and high (301 *ms*) stress groups ($F(2, 24)=4.94$, $p < .05$, $\eta^2=.29$). Similarly we found that cortisol release at C2 linearly predicted non-decision time ($\beta = -.016$, $t= -2.42$, $p < .05$, $r^2=.15$). This effect on non-decision time was expected, as accuracy was not different across groups. Difference in any other parameter than non-decision time would in general entail difference in accuracy.

One interesting hypothesis would be that the longer non-decision time for low stress participants would allow them to better encode decision parameters, leading to better confidence accuracy thereafter. Thus, we investigated whether individual differences in non-decision times would mediate the relationship between stress and confidence accuracy. Specifically, we tested the model that the impact of cortisol release at C2 on the area under the type 2 ROC curve (AUC) is mediated by non-decision time. However, we found that, controlling for cortisol, non-decision times did not predicted AUC ($p > .60$), which precludes any further investigation of the mediation.

Thus it seems that stress reactivity is associated both with faster response times and to lower metacognitive accuracy, but that these two effects are mainly independent. Notice further that while the main difference in metacognitive accuracy is between the high group on the one hand and the medium and low groups on the other, differences in response times seems to lie between the low group on the one hand and the high and medium on the other.

Response times on the confidence scale

Due to the response mode (mouse movement and then mouse click) response times on the confidence scale are noisier than on the first order task. Nevertheless, we investigated whether the speed at which participants gave the estimate of their confidence could in part explain their variable metacognitive accuracy. First, we looked at difference in response times on the confidence scale (hereafter: confidence RT) across stress groups. A one-way ANOVA with group as factor and participant as random effect did not reveal any difference ($p > .41$). Still, when we applied the same diffusion analysis methodology as above to confidence RTs, we found that non-decision times were slower for the low stress group (401 *ms*) than for the medium (290 *ms*) and high (297 *ms*) groups ($F(2,24)=3.7$, $p < .05$, $\eta^2=.24$; all other ps $> .18$). But, again neither confidence RT or non-decision times of confidence RT did predict AUC ($p > .70$), suggesting that the reason why high stress participants have lower metacognitive accuracy is not their hastiness to respond on the confidence scale.

References

1. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J. Am. Statist. Assoc. 1979;74, 829–836.

2. Laming DRJ. Information, theory of choice-reaction times. New York: Academic Press; 1968

3. Voss A, Voss J. Fast-dm: A free program for efficient diffusion model analysis. Behav Res Methods. 2007;39, 767–775.

4. Ratcliff RA. A theory of memory retrieval. Psychol. Rev. 1978;85, 59–108.

2.4.4.3 Pilot Study for "Self-knowledge dim-out".

Below is a pilot study for the previous article "Self-knowledge dim-out". The pilot study was an application of the methods of the experiments in the first paper of this dissertation "Introspection during visual search", with the same stress methodology as in the published "Self-knowledege dim-out". The results demonstrated that the method would be useful in assessing the effect of stress on introspection, but the novelty of the introspective measures (SNSI) made the interpretation difficult. This is why in the published version I resorted to the simpler and better studied measure of confidence judgments.

Material and Methods

Participants

Session 1: *Stress-sensitivity assessment.* 120 students (mean age: 20.6, *SD*: 1.5, 60 women) participated in the study. Exclusion criteria were: a body mass index of < 18 or > 30 kg/m²; receiving medical treatment known to affect the HPA (hypothalamus – pituitary– adrenal) axis; a history of psychiatric or neurological disorders; abnormal vision; smokers; pregnant women and women taking oral contraceptives. Participants were asked not to eat or brush their teeth 1 hour before the TSST, and to not drink alcohol or play sports the day before. Participants gave their informed consent and were compensated U$10 for the 2-hour session. Experimental sessions were scheduled from 2:30 to 6:30 PM.

Session 2: *metacognitive abilities measure*. Based on the results of session 1, we constituted three groups of 9 participants with high, medium and low sensitivity to stress. These 27 participants took part approximately 8 months after session 1 (mean age: 20.3, *SD*: 1.7; 11 women). The experimental session lasted 1-hour.

Apparatus & Stimuli & Procedure

Session 1. The TSST consisted of asking participants to prepare an oral presentation and perform mental arithmetic in front of an expert panel. The protocol started with a rest phase of 10 minutes. Then, it was extracted the first sample of saliva and participants were asked to respond to the State-Trait Anxiety Inventory (STAI). Afterward, participants were instructed to stand in a fictional scenario in which they had to present 5 minutes

oral presentation (Fig. 1A). Immediately after this, the expert panel asked participants to do mental arithmetic for 5 minutes. Finally, a second saliva sample was taken and participants were asked to do a second STAI. Five more saliva samples were taken during a post-exposure rest phase to control normal decrease in the stress reactivity curves. Cortisol concentration in saliva was measured with a standard procedure (see main study)

Session 2. Stimuli consisted of a set of black letters (T, L or X, size: 0.8° X 0.6°, luminance: 0.5 cd/m²) on a uniform grey background (luminance: 44.1 cd/m²), presented on an imaginary circle (radius: 6.2°) around a central fixation spot at the center of the screen. Individual orientation for each letter was randomized (0, 90, 180, 270°). Stimuli were equally spaced on the imaginary circle, while its overall orientation was randomized for each trial. Stimuli were presented on a CRT screen (size 17", resolution of 1024 X 768, refresh rate of 100 Hz, viewing distance ~ 55 cm). The experiment took place in a dark room. Stimuli (Fig. 1B) were presented for 200 *ms*, preceded by a fixation spot presented for 500 *ms*. Participants were instructed to decide on the presence or absence of a target (L or X) within the set of distractors (Ts), by pressing the "Q" or "W" key on a standard Spanish "QWERTY" keyboard. Half of the trials were target absent trials. Target present trials contained one "L" or one "X". Set-size (2, 4, 8 or 12 items) and presence-absence of a target were fully crossed. X targets were meant to create easy, "pop-out" searches ("Feature searches"), while L targets were introduced to create difficult, attentional searches ("Conjunction searches"). Immediately after the response, three continuous introspective scales were simultaneously presented: i) Confidence: *Are you certain of your decision?* (anchored at "guess" and "absolutely certain"); ii) Subjective Number of Scanned Items (SNSI): *How many items do you think you examined before reaching your decision?* (Four qualitative categories were presented below the scale: "no item", "some items", "many items" and "all items"); iii) Introspective response time (iRT): *How long do you think that it took you to determine whether the target was present or absent?* (Graduated from 200 *ms* to 1200 *ms* with intervals of 100 *ms*). Experimental session comprised 256 trials divided in 8 blocks, with a 60 second pause between each block. Before beginning the experiment, participants had 32 training trials.

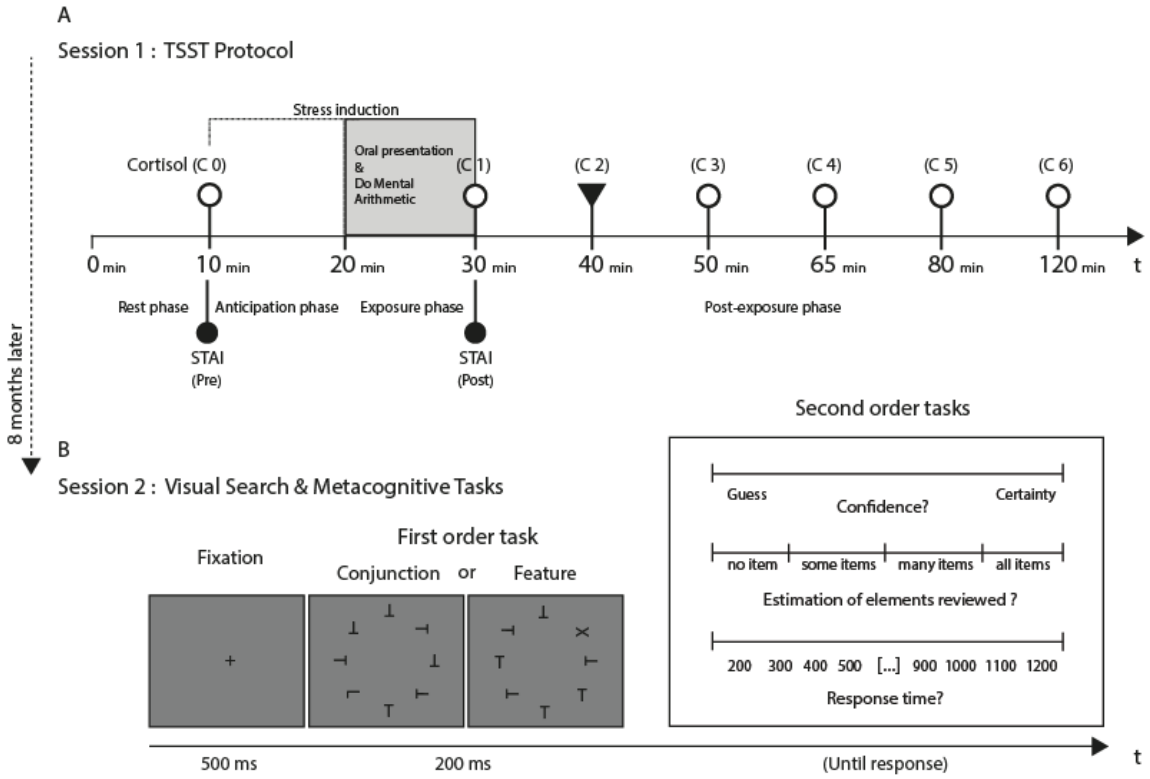


Fig. 1. (A) TSST protocol (*session 1*) and (B) Visual search task with 3 metacognitive scales (*session 2*).

Results

Session 1

As is customary in comparable studies (see 1), cortisol level at C2 was taken as indicative of participants' sensitivity to stress (Fig. 2A). Participants above the 75th percentile were categorized as having high susceptible (M = 10.90 nmol/l; SE = .55; range: [9.7, 14.1]; Cortisol increase [C2-C0 or Baseline] = 2.53), participants below the 25th percentile as low sensitivity (M = 4.57 nmol/l; SE = .22; range: [3.6, 5.4]; Cortisol increase = .18) and participants around the 50th percentile as medium sensitivity (M = 7.49 nmol/l; SE= .35; range: [6.0, 8.8]; Cortisol increase = .32). Classifications reach the criteria for distinguishing cortisol responders from non responders (threshold: 1.1 nmol/l; see 2). Cortisol concentration (C2) evidenced no differences due to gender (p > .88, one-way ANOVA) or correlation with participants' age (p > .11). Moreover, the stress induction procedure increased the score on the STAI (Fig. 2B). A repeated measure ANOVA on STAI scores, with a factor of application (before/after) and stress groups (high, medium and low stress reactivity) evidenced a main effect of application (F(1,24) = 13.2, p < .01, η^2p = .36), without differences between stress groups (p > .47) or interaction (p > .13).

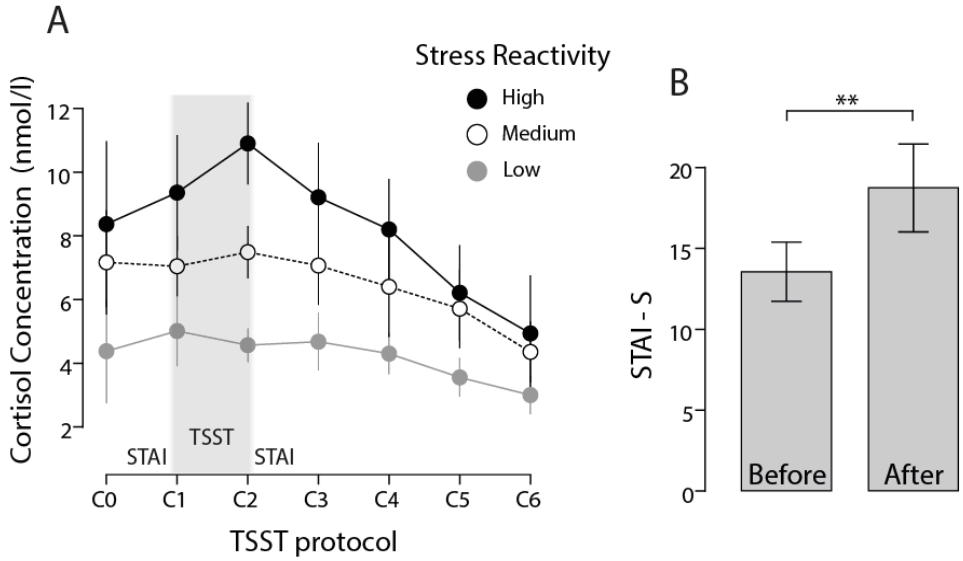


Fig. 2. A) Cortisol concentration as a function of TSST phase for each experimental group. Error bars denote Cousineau-within-subjects 95% confidence intervals. B) State-Trait Anxiety Inventory (STAI), before and after stress induction. Error bars are 2+/- *SE*. Higher scores denote higher levels of stress-anxiety.

Session 2

First order task

Trials with response times below 200 *ms* and response times 2 *SD* above the mean (4.8%) were excluded. Response times and error rates were positively correlated ($r^2(324) = .24$, $\beta = .48$, $p < .001$), thus, it was computed Inverse Efficiency Scores (IES: ratio of median RT over proportion of correct responses). Before computing IES (see 3), RTs were log-transformed to normalize the distribution. The traditional pattern of interaction between search type (FS: feature search (X) *vs*. CS: conjunction search, (L)) and set-size (2, 4, 8, 12) was observed. Crucially, there were no differences between stress groups. A repeated measure ANOVA on IES in target present trials, with a factor of search type, set-size, stress group and their interactions was run. The analysis showed a significant main effect for set-size ($F(2.2,53.1) = 13.9$, $p < .001$, $\eta^2p = .37$; Greenhouse-Geisser corrected), search type ($F(1,24) = 38.8$, $p < .001$, $\eta^2p = .62$) and an interaction between these ($F(2.1,52.0) = 13.9$, $p < .001$, $\eta^2p = .37$; G-G). No other main effects or interactions were found (all *p*s > .12). The same analysis on

absent target trials showed only a main effect for set-size ($F(1.1,27.9) = 8.65$, $p < .01$, $\eta^2p = .26$; G-G, all others ps > .82).

Second order tasks

It was analyzed how well our three metacognitive scales correspond to the objective performance, so as to quantify metacognitive abilities. All p values were Bonferroni corrected (p(cor)), to account for the 3 dependent variables. First, a similar repeated measure ANOVA on SNSI was run in correct-present trials. The results indicated that the three groups estimated reviewing more items depending on the number of items presented ($F(1.1,26.4) = 135.7$, $p < .001$, $\eta^2p = .85$, G-G). This main effect was qualified by an interaction with stress group ($F(6,72) = 4.29$, $p < .01$, $\eta^2p = .26$). Pairwise comparisons revealed a difference only between medium and high stress group ($M = 1.63$, $SE = .59$, $p < .05$). No other main effects or interactions were found (all ps > .09, Fig. 3A). When the analysis was repeated on absent target trials a similar pattern was observed (set-size main effect: $F(1.1,27.7) = 91.4$, $p < .001$, $\eta^2p = .79$, G-G; stress group X set-size: $F(6,72) = 3.77$, $p < .01$, $\eta^2p = .24$; stress group main effect: $F(2,24) = 6.55$, $p < .05$, $\eta^2p = .35$; medium *vs*. high: $M = 1.68$, $SE = .47$, $p < .05$).

Next, an assessment was made as to what extent participants accessed their own response times (RTs). It was found that participants with higher stress reactivity were less sensitive to their own RT than individuals with low reactivity (Fig. 3B). A Linear Mixed Model (LMM) was run across single present target trials. The global RT fixed effect and the interaction effect between RT X stress groups on iRT were investigated. In addition, a random participants' intercept was included. It was observed that the RT main effect ($F(1,2870.2) = 596.7$, $p < .001$) and the interaction ($F(3,2457.3) = 200.3$, $p < .001$) were significant, indicating that the regression coefficient between RT and iRT was different depending on the stress group. Closer inspection evidenced that the low stress reactivity group presented the higher increase, even when all stress groups presented a significant and positive relationship (low: $\beta_{(stand)} = .42$, $t(962) = 14.5$, $p < .001$; medium: $\beta = .22$, $t(982) = 7.18$, $p < .001$; high: $\beta = .23$, $t(938) = 7.10$, $p < .001$). Post-hoc comparison showed a significant difference between them (low *vs* medium: $\beta = .30$, $t(1944) = 4.55$, $p < .001$; low *vs* high: $\beta = .37$, $t(1900) = 6.01$, $p < .001$; medium *vs* high: $p > .32$), meaning that compared to low susceptible participants, high and medium stress participants were less able to track the variability of their own RTs. By repeating the analysis on FS and CS, separately (CS: low: $\beta = .19$,

$t(396) = 3.92$, $p < .001$; medium: $\beta = .14$, $t(413) = 2.92$, $p < .01$; high: $p > .06$; FS: low: $\beta = .48$, $t(564) = 13.1$, $p < .001$; medium: $\beta = .17$, $t(567) = 4.26$, $p < .001$; high: $p > .52$) a similar pattern was observed.

Furthermore, when the same LMM analysis was run on absent target trials, it was observed that there was a significant main effect of RT($F(1,2471.6) = 404.1$, $p < .001$) and an interaction between RT and stress group ($F(2,1705.6) = 30.7$, $p < .001$), showing a significant slope for each group (low: $\beta = .36$, $t(847) = 11.4$, $p < .001$; medium: $\beta = .29$, $t(828) = 8.90$, $p < .001$; high: $\beta = .24$, $t(797) = 7.11$, $p < .001$). Closer inspection revealed no differences between low *vs* medium ($p > .34$) and medium *vs* high ($p > .21$), but a significant difference between low *vs* high: $\beta = .20$, $t(1644) = 2.90$, $p < .01$), confirming again our prediction.

Finally, performance monitoring by means of the correlation of confidence ratings and performance was investigated. To this end, using a similar LMM we regressed confidence on accuracy in the first order task with stress group as covariate and participants as random effect. We found that participants' confidence was higher in correct trials ($F(1,39.3) = 11.70$, $p < .01$), and that participants in different stress groups were equally confident ($p = .09$). Critically there was an interaction between stress group and accuracy, to the effect that the higher the stress reactivity, the lower the increase in confidence from incorrect to correct trials ($F(3,28.3) = 7.11$, $p < .05$). A more detailed assessment of this interaction, indicated a significant difference between low and medium reactivity group ($\beta = -.46$, $SE = .19$, $p < .05$) and also between the low and high reactivity ($\beta = -.83$, $SE = .20$, $p < .001$). Consistent with these results, as shown in Fig. 3C, meta-cognitive sensitivity was better in the lower than in higher stress group, but exclusively for absent target trials. Metacognitive sensitivity was quantified by the relationship between confidence rating and accuracy using type-2 Receiver Operating Characteristic curve (ROC, see 4), calculated separately for each stress group in present (mean accuracy in high: .86; medium: .90; low: .85, one-way ANOVA: $p > .90$) and absent target trials (mean accuracy in high: .79; medium: .81; low: .80, one-way ANOVA: $p > .92$). ROC curves were computed by plotting the cumulative probabilities for each confidence level (i), both for incorrect trials on the x-axis, p (confidence = i | incorrect), as well as for correct trials on the y-axis, p (confidence = i | correct). Higher Area under the type-2 ROC curve (AUC) denotes a stronger link between confidence and accuracy.

The analysis run on present target trials showed no differences between stress groups ($p > .28$, one-way ANOVA), even after having repeated it separately for each search condition (all $ps > .52$). However, when the

analysis was repeated on absent target trials, a significant main effect of stress group was evidenced ($F(2,26) = 4.81$, $p < .05$; high: .04; medium: .18; low: .23), which was characterized by a difference between low and high stress group ($M = .18$, $SE = .06$, $p < .05$). No differences were observed between medium and high ($p > .09$) or low and medium stress groups ($p > .99$).

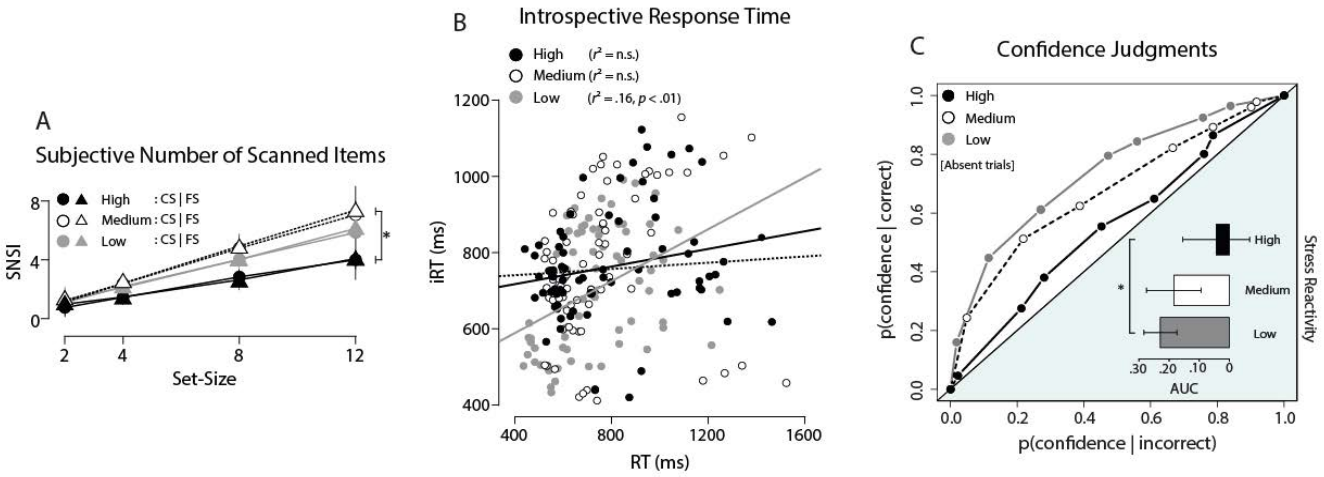


Fig. 3. A) Subjective Number of Scanned Items (SNSI) as a function of set-size and search conditions, for each stress group in present target trials. B) Linear regression of mean RT on mean iRT for each stress group. For presentation purposes only, each dot represents the RT/iRT mean of each participant for each set-size and search condition in present target trials. C) Metacognitive sensitivity quantified by the relationship between confidence rating and accuracy using type-2 ROC curves in absent target trials.

References

5. Allen AP, Kennedy PJ, Cryan JF, Dinan TG, Clarke G. Biological and psychological markers of stress in humans: focus on the Trier Social Stress Test. Neurosci Biobehav Rev. 2014;38, 94–124.

6. Miller R, Plessow F, Kirschbaum C, Stalder T. Classification criteria for distinguishing cortisol responders from nonresponders to psychosocial stress: evaluation of salivary cortisol pulse detection in panel designs. Psychosom. Med. 2013;75, 832–840.

7. Townsend JT, Ashby FG. The Stochastic Modeling of Elementary Psychological Processes. Cambridge: Cambridge University Press; 1993

8. Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychon Bull Rev. 2003;10, 843–867.

3 Conclusion

In this work I presented four experimental studies on introspection. The overall goal of my dissertation was to investigate the extension of the cognitive domain accessible to introspection, and the constraints that apply to it. It seems that the natural direction for experimental psychology was, until recently, to look for *limits* of introspection. It is as though scientific psychologists felt that their duty was to counteract the spontaneous tendency of thinking that we know what is happening in our minds. Serious psychology would start by dispelling this illusion. Perhaps also, scientific psychologists felt that they had to make amends for the unchecked belief in the powers of introspection that was common prior to the behaviorism turn, at the inception of the discipline. Whatever the reasons, seemingly counterintuitive claims about failures of introspection are much more common in the literature than the opposite claims about the efficacy of introspection (Lyons, 1987). However, if there is a boundary between what is accessible and what is not accessible to introspection, it is equally important to describe the functioning of introspection within its proper domain. My work thus starts with the intuition that it is now time to explore the positive side of introspection: what *does* it have access to? What are its mechanisms? What constraints apply to it? For instance, it does not seem a satisfactory methodology to define as accessible to introspection "whatever has not been shown to be inaccessible"!

I took my starting point in Nisbett and Wilson's most famous attack against introspection in their 1977 paper. I wanted to discuss one very specific claim that they make, namely that cognitive processes are inaccessible, contrary to cognitive states. To me, this claim, taken generally, is mistaken. I wanted to show that with the proper definition of the state / process distinction, and with an appropriate experimental methodology, processes could be shown to be accessible to introspection. According to Nisbett and Wilson's theory (Nisbett & Wilson, 1977), only an introspective measure focused on cognitive states (i.e., perceptual states accompanying or preceding the decision) can access subjective information that accompanies cognitive activity. In opposition, access to processes (for instance the sensory transformations making up the decision) would yield an interpretation of the behavior by the subject, and not an introspection. Despite substantial objections (Ericsson & Simon, 1980; Smith & Miller, 1978; White, 1980, 1987, 1988), and recent reformulations (Wilson, 2002, 2003; Wilson & Dunn, 2004), this idea was established as a canon in the literature of experimental metacognition (Johansson et al., 2005; Overgaard, 2006; Overgaard & Sandberg, 2012).

To my knowledge, there has been no attempt in contemporary experimental psychology to directly test whether introspection of processes was possible, with an explicit definition of a process. Experimental work has been more interested in investigating the functional conditionals of access mechanisms (Fleming & Frith, 2014), through the introspection of perceptual cognitive states with low complexity. In contrast, in philosophy of psychology the introspection of processes issue has been explored. Strikingly the contemporary consensus seems to coincide with the traditional negative answer: there would be an apparent impossibility of introspective access to judgments and decisions (Carruthers, 2010). Given this, how to experimentally face the question of the extension of introspective reports to processes? From the reviewed conceptual and experimental literature, I reasoned that a reliable assessment of the extension of introspection would require: (i) to specify the distinction between cognitive processes and states (White, 1980, 1987, 1988); (ii) to determine the relationship between introspective access and report accuracy (Feest, 2012; Goldman, 2004; Jack & Roepstorff, 2002) and thus, the reliability of different introspective procedures (Froese, Gould, & Seth, 2011; Overgaard & Sandberg, 2012; Sandberg et al., 2010) as well as being able; (iii) to identify and control the sources of contamination involved in the introspective process of information retrieval (Piccinini, 2003; Prinz, 2004). In the experimental works that I presented, I tried to implement each of these ideas. I will now review them in turn.

First, the distinction between a cognitive state and process is traditionally defined in terms of its causal relation, (Miller, 1962; Neisser, 1967), an aspect that is not experimentally approachable (Engelbert & Carruthers, 2010). Paradoxically, the Nisbett and Wilson cannon has opted for a definition in terms of accessibility: a cognitive process is a cognitive element that cannot be accessed by participants' consciousness (Nisbett & Wilson, 1977). Such definition comes from evidence that subjective descriptions of sequences of mental actions do not faithfully correspond to the functioning pattern supposed for such mental content. However, given our previous remarks, it is more plausible to maintain that the introspective limit does not derive from the nature of the mental content. It is more parsimonious to maintain that certain mental contents, generated at particular moments of a task, are easier to access than others. The present thesis defended the idea that introspective blindness depends on functional properties of the access mechanism. In other words, I proposed that the reliability of introspection is a matter of the degree of accessibility. It is not an opposition between classes of mental contents. In the first paper presented (*Introspection during visual search*) I proposed the idea that the focus of introspection could vary according to two dimensions: along the self-observation / mental monitoring axis and along the early / late stage of task execution. The distinction between a cognitive state and process would be determined then by the focus

of introspection within this parameter space. In the three first papers of this dissertation, I provided evidence that mental monitoring of intermediate stage of task execution was possible, which is the operational definition of introspection of processes I adopt.

The second point is methodological: the contemporary literature converges on the idea that the validity of an introspective report must be established as a match between the subjective state reported and certain objective consequence attributed to the mental content of interest. Quite simply, in this sense, it is acceptable to conclude that an individual presents an adequate subjective access to its response time when such index reliably correlates with the objective response time over the same experimental task (Corallo et al., 2008; Marti et al., 2010). Critically, all my experimental studies abide by this requirement. Perhaps the most interesting in this respect is the second study on visual search (*Introspective access to an implicit shift of attention*) because I infer a shift of attention from the properties of the eye movements of participants, something that they are not directly aware of.

More generally, as I do not *use* introspection to study mental mechanisms, but *study* introspection of known mental mechanisms, I always need an objective counterpart to introspection, as a hypothetical model of mental activity. Thus, in case there is a discrepancy between introspection and the model, there is the possibility that introspective reports may not be inadequate, but rather the theoretical model supposed for such mental processing. In this line, Jack and Roepstorff (2002) suggested that, like many others studies interested in the access of simple perceptual states, research interested in the access to complex processes must be, in the first place, able to prove that it is capable of objectively model a cognitive process and later to evaluate whether such changes are reliably described by the introspective report. This was the strategy adopted in all the presented research projects. This point is particularly important with respect to the study on memory (*Introspection during Working Memory Scanning*), as it presupposes subtle experimental arguments (McElree and Dosher, 1989) to the effect that item recognition tasks are not (contrary to what had been argued for by Sternberg in his inaugural papers of 1966) serial, but in fact parallel. Notice that the introspective evidence in the same direction can be seen as corroborating the serial / parallel distinction for judgment or recency and item recognition tasks. In this sense, this may be the only instance where my studies contribute to the knowledge of first order mental processes.

The third critical point in the experimental study of introspection is the isolation of the relevant source of information for introspection (and correlatively the identification of the potential contaminants). My experiments show that introspective mechanisms filter both internal and external information. The validity of introspection is determined by the precision with which it focuses on a particular mental or behavioral content and discards others that are irrelevant. In the first study (*Introspection during visual search*), I tried to systematically weigh the role of behavioral information (both explicit: response times, and implicit: eye movements) in the buildup of introspective reports. From my results, it seems that individuals do not control very precisely the diversity of sources of information that can contribute to the generation of an introspective judgment. In consequence, an adequate framework for the investigation of introspection should consider control mechanisms to corroborate that the introspective reports corresponds to the experimental interest and not to other sources of information. Several authors concur with this recommendation: rely only on an introspective data when it co-varies with experience, behavior or physiological responses (Jack & Roepstorff, 2002; Schooler, 2002a; Schooler & Schreiber, 2004). In retrospect, it is probable that the general introspective blindness proposed by Nisbett and Wilson (1977) stems from poor experimental control over the focus of introspection in (almost) all the experiments analyzed by the authors (White, 1980, 1988). The experimental work developed here has always considered this possibility.

An important qualification to all my studies is in order here: in experimental work on introspection, one cannot directly control the source of information (contrary to what can be done in sensory psychophysics, for instance). Thus, one must give precise instructions to participants as to what introspective task is expected, and then trust them as to their performance. But of course compliant participants will try to make sense of instructions in the most relevant way. So it may be that a non introspective interpretation of the instructions is favored by participants. This is in a sense what happened in the first experiment of *Introspection during visual search* where participants seem to report the perceptual load of the stimulus (an objective property) rather than the complexity of the decision process, because of the short presentation time.

In general terms, the first three papers converge on the same idea: the reliability with which individuals access their own mental contents depends on the experimental context. Traditionally, literature has maintained that the veracity of introspection would be determined by the kind of mental content over which a judgment of this type is aimed (Carruthers, 2010). Our experimental work offers empirical evidence in favor of another interpretation:

accessibility does not depend on the nature of a mental content, but more on the accuracy with which introspective mechanisms internally discriminate certain contents with respect to others (Prinz, 2004). In this respect, from our first research project (*Introspection during visual search*) it is possible to extract two main ideas: 1) introspective capacity is less limited than what was previously thought (Nisbett & Wilson, 1977): individuals, under certain particular experimental conditions, are able to report a mental content of high cognitive complexity; 2) the reliability of introspective data is highly dependent on the information surrounding the introspection act (e.g., response time or eye-movement), as well as the experimental context where this is executed (introspective task). Regarding this last idea, even though it is inspired by early theoretical developments (Ericsson & Simon, 1980), it is important to point out that our experimental work constitutes the first attempt at measuring the degree of contamination of an introspective report. Considering both points, we argue that introspective accuracy depends on the degree with which the inner scanning mechanism (Armstrong, 1968) focuses the content of interest and not other(s). My second research project (*Introspective access to an implicit shift of attention*) confirms the previous ideas, but also offers another demonstration that the focus of introspection can be targeted over cognitive processing (operationally: attentional shift), which take place between a sensory processing and motor response, independently of the consciousness of the cause that activates such processing. In other terms, individuals, without knowing the cause of the attentional shift are able to access it and introspectively report it, suggesting that they can access intermediate information between the sensory cause and response. In retrospect, the first and second research projects converge on the idea that it is possible to experimentally control the irrelevant information for introspection (and with this, diminish the probability that participants will form non-introspective judgments), without altering the accuracy of their reports. The third research project extend these ideas to another cognitive modality: introspection is successful at describing the cognitive processing during a memorization task. At the same time, such access to the nature of the decision process, as in previous projects, was influenced by irrelevant contextual information (response times).

The final research project (*Self-knowledge dim-out*) had a different motivation from the first three. Here, I focused on the structural constraints that apply to introspection. I evidenced that biological reactivity to stressful contexts predicts the accuracy of introspection. In line with recent studies which underlie precise cortical areas responsible for intra-individual differences of this capacity, our study suggests that psycho-social stress is a relevant factor.

From the above considerations commented above and, it is possible to propose a new model for future investigations in metacognition. This framework makes a distinction between *self-observation* and introspection. Knowledge about oneself, even about one's own mental processes, can derive both from direct mental contact with the targeted parameter or process, or through inferences based on self-observation of behavior. Both qualify as introspection in a broad sense, but only the first is pure introspection, which may be more adequately termed *mental monitoring*. Schematically, mental monitoring is possible for two types of mental content: (i) low cognitive complexity mental contents (cognitive states); (ii) high cognitive complexity mental contents, defined here as cognitive processes. While this distinction was clearly stated in Nisbett and Wilson's (1977) seminal paper, it may have been under-appreciated in more recent experimental studies of introspection. Moreover, besides the complexity of the mental content it is also important to take into account the moment, over the course of the task, from which the information is read out. In effect, the introspective mechanism can read out information generated early during the experimental task (for instance perceptual load), but also mental state appearing late, or even after completion of the task (subjective perception of error/success in a decision). This second dimension specifies the stage over which the introspective focus is aimed, regardless of the complexity of the mental content. Most importantly, the distinction between early *vs*. late introspective recovery processes has received certain experimental interest in recent years through the question of the limit of the process of accumulating information: some have contended that introspection is limited to the information accumulated until the moment of the first-order decision (Kiani & Shadlen, 2009; Zylberberg et al., 2012), but others have claimed that it essentially extends afterwards (Pleskac & Busemeyer, 2010). Overall, taking into account the distinction by Nelson and Narens' (1990) between *monitoring* and *control*, we propose a descriptive model of introspection that represents two dimensions that define the space within which introspection can flexibly be focused (Figure 3). This model can serve as a framework for further research on the mechanisms of introspection. It should drive attention over the contaminants of introspection: Researches interested in introspection should consider analytical tools that confirm that the reached report correlates with the genuine mental phenomenon of interest (Piccinini, 2003), specifying whether this is a product (self-observation) or a cognitive process (mental-monitoring).

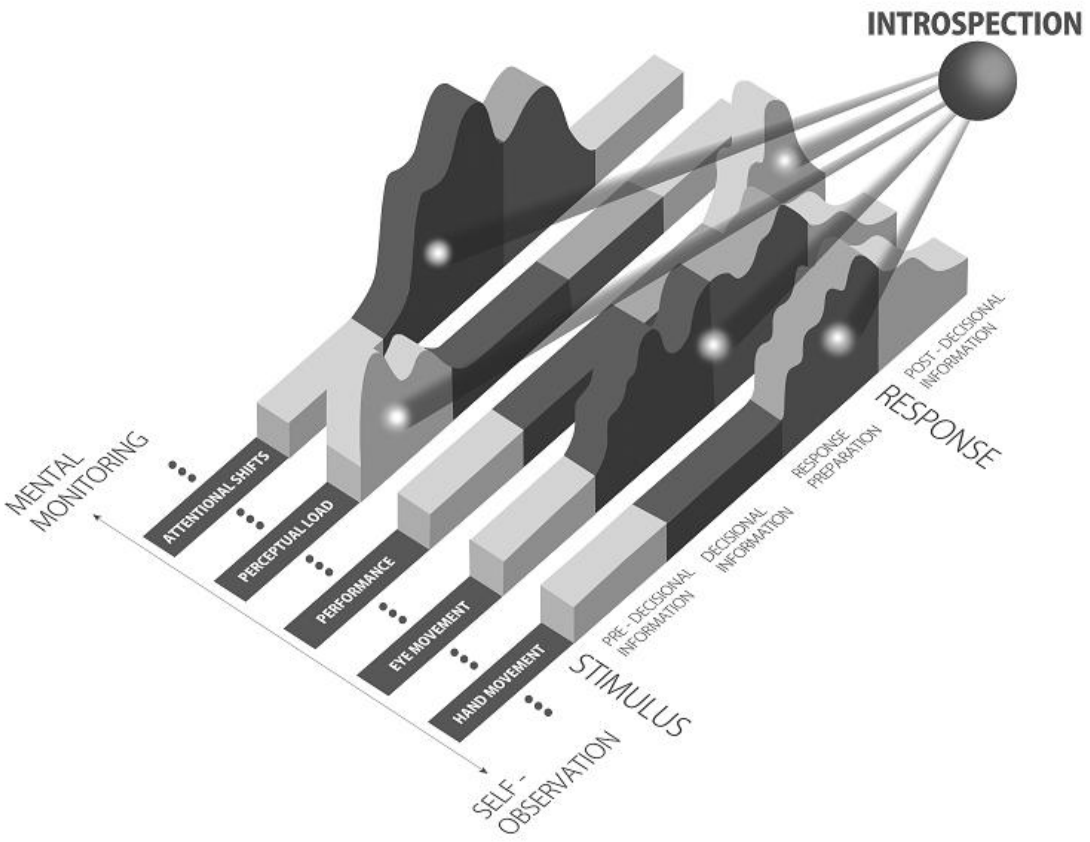


Figure 3. Schematic representation of the flexibility of introspection. The first dimension (mental monitoring *vs*. self observation) organizes and presents sources of information on which an introspective measure is focused. When an introspective index primarily uses behavioral information sources, introspection is conceptualized as a process of *self-observation*. By contrast, when the introspective source of information comes from the cognitive process underlying behavioral products, the index is properly conceptualized as *mental monitoring*. There is a gradation with respect to introspection, with pure mental monitoring at one extreme and pure self-observation at the other. The second dimension (early *vs*. late processes) specifies different stages, during the development of the first order task, at which mental states are available to introspection. Participants can be guided to focus the introspective mechanism on different stages during a cognitive task.

As final words, our results allow suggesting that the introspection of cognitive processes is possible under precise experimental controls. The previous point supposes an advance in the conceptualization and experimentation of the limits of subjective report. Despite this, there are certain aspects that were not part of my studies that require further experimental work.

One general question that emerges from my work is the question of inter-individual variability in introspective capacity. Some experimental evidence (Fleming et al., 2010) show that differences in introspective capacity are related to brain structures. In this context, my thesis proposes a new factor: the biological reactivity to stress. The comparison of this fourth paper and of the structural brain determinants of the accuracy of introspection (Fleming et al. 2010) raises the question of the mechanism through which stress impacts metacognition. Multiple questions come up regarding this: is the effect of stress over metacognition characteristic of a particular personality trait? In which case, is it rooted in the same brain structures as the ones highlighted by (Fleming et al., 2010)? And if there are personality profiles, is there a sensitive period in the development of metacognitive capacity? My studies do not give conclusive evidence regarding the type of stress impacting metacognition. In effect, the impact of stress on metacognition could be mainly due to psychosocial stress or contextual stress. Furthermore, it could also be the case that the evidenced metacognitive variability should not directly relate to reactivity to stress (cortisol), but more to the associated cortical response. Whatever the answers to these questions, it seems possible to dampen introspective mechanism through stress induction. This opens a wide research avenue for unveiling the functional architecture of introspection.

From a general perspective, my thesis is relevant both for the science of consciousness and science of metacognition, as well as for the methods in introspective research. As for the first point, our results evidence a certain access to information related to the sensory treatment that precedes the decision. Speculatively, such high-order representation could have a particular neural signature. In other words, in the same way as recent neuroimaging studies, (Baird et al., 2013; Baird et al., 2014; Fleming et al., 2014) evidenced different neural correlates for introspection over perceptual tasks *v*s. memorization tasks, it is possible to imagine different neural correlate for introspection over cognitive states *vs*. cognitive processes. Regarding the science of metacognition, our experimental work opens a new line of research: while my studies make an argument in favor of the multiplicity of information sources in the constitution of an introspective judgment, I did not investigate how individuals synthesize such information, neither the specific intra-individual conditionals of introspective processing. This is an open question I shall hopefully study in the coming years.

4. General references

Armstrong, D. (1968). *A materialist theory of the mind.* London: Routledge.

Armstrong, D. (1997). *A World of States of Affairs.* Cambridge: Cambridge University Press.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychol. Learn. Motiv., 2*, 89-195.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *J. Neurosci., 33,* 16657–16665.

Baird, B., Mrazek, M. D., Phillips, D. T., & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation raining. *J. Exp. Psychol. Gen., 143,* 1972–1979

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., Roepstorff, A., Rees, G., Frith, C. D., & Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Conscious. Cogn., 26,* 13-23.

Barthelme, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proc. Natl. Acad. Sci. U.S.A., 107,* 20834-20839.

Bayne, T. (2015). Introspective insecurity. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 3(T).* Frankfurt am Main: MIND Group.

Bem, D. J. (1972). Self-perception theory. *Adv. Exp. Soc. Psychol., 6,* 1-62

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends Cogn. Sci., 15,* 567-575.

Bona, S., Cattaneo, Z., Vecchi, T., Soto, D., & Silvanto, J. (2013). Metacognition of Visual Short-Term Memory: Dissociation between Objective and Subjective Components of VSTM. *Front. Psychol., 4,* 62.

Bona, S., & Silvanto, J. (2014). Accuracy and Confidence of Visual Short-Term Memory Do Not Go Hand-In-Hand: Behavioral and Neural Dissociations. *PLoS ONE, 9,* e90808.

Boring, E. G. (1953). A history of introspection. *Psychol. Bull., 3,* 169-189.

Broadbent, D. E. (1958). *Perception and communication.* Elmsford, NY: Pergamon Press

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychol. Bull., 109,* 204–223

Bryce, D., & Bratzke, D. (2014). Introspective reports of reaction times in dual-tasks reflect experienced difficulty rather than timing of cognitive processes. *Conscious. Cogn., 27,* 254-267.

Butler, J. (2013). *Rethinking Introspection: A Pluralist Approach to the First-Person Perspective.* London: Palgrave Macmillan.

Carruthers, P. (2010). Introspection: Divided and Partly Eliminated. *Philos. Phenomenol. Res., 80,* 76-111.

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

Churchland, P. (1984). *Matter and Consciousness.* Cambridge, MA: MIT Press.

Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychol. Sci., 19,* 1110-1117.

Costall, A. (2006). 'Introspectionism' and the mythical origins of scientific psychology. *Conscious. Cogn., 15,* 634–654.

Danziger, K. (1980). The history of introspection reconsidered. *J. Hist. Behav. Sci., 16,* 241-262.

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philos. Trans. R. Soc. Lond. B, 367,* 1379-1390.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan R. J. (2013). Confidence in value-based choice. *Nat. Neurosci., 16,* 105–110.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci., 10,* 204-211.

Dunlosky, J., & Bjork, R. A. (Eds.). (2008). *A handbook of metamemory and memory.* Hillsdale, NJ: Psychology Press.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Los Angeles, CA: SAGE Publications.

Dunlosky, J., Serra, M., & Baker, J. M. C. (2007). Metamemory. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), *Handbook of Applied Cognition* (2nd ed., pp. 137–159). New York, NY: Wiley

Engelbert, M., & Carruthers, P. (2010). Introspection. *Wiley Interdiscip. Rev. Cogn. Sci., 1,* 245-253.

Engelbert, M., & Carruthers, P. (2011). Descriptive Experience Sampling: What is it good for? *J. Conscious. Stud., 18,* 130-149.

Ericsson, K. A. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks: Toward a solution to the central problems of introspection as a source of scientific data. *J. Consciousness Stud., 10,* 1-18.

Ericsson, K. A., & Fox, M. C. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011). *Psychol. Bull., 137,* 351–354.

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychol. Rev., 87,* 215-251.

Feest, U. (2012). Introspection as a method and introspection as a feature of consciousness. *Inquiry, 55,* 1-16.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp.231-236). Hillsdale, NJ: Erlbaum.

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

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philos. Trans. R. Soc. Lond. B, 367,* 1338-1349.

Fleming, S. M., & Frith, C. D. (Eds.). (2014). *The Cognitive Neuroscience of Metacognition*. Berlin: Springer.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci., 32,* 6117–6125.

Fleming, S. M., & Lau, H. (2014). How to measure metacognition. *Front. Hum. Neurosci., 8,* 1–9.

Fleming, S.M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychol. Sci., 26,* 89-98.

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

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541-1543.

Fodor, J. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychol. Bull., 137,* 316–344.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Phil. Trans. R. Soc. Lond. B, 367,* 2213–2223.

Froese, T. (2013). Interactively guided introspection is getting science closer to an effective consciousness meter. *Conscious. Cogn., 22,* 672-676

Froese, T., Gould, C., & Seth, A. K. (2011). Validating and calibrating first and second person methods in the science of consciousness. *J. Conscious. Stud., 18,* 38-64.

Goldman, A. (2004). Epistemology and the evidential status of introspective reports. *J. Conscious. Stud., 11,* 1-16.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* Oxford: Oxford University Press.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *J. Exp. Psychol.- Learn. Mem. Cogn., 6,* 651-675.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. *PLoS ONE, 8,* e60554.

Hart, J. (1965). Memory and the feeling-of-knowing experience. *J. Educ. Psychol., 56,* 208–216

Hart, J. (1966). Methodological note on feeling-of-knowing experiments. *J. Educ. Psychol., 57,* 347-349.

Hart, J. (1967). Memory and the memory-monitoring process. *J. Verbal Learning Verbal Behav., 6,* 685–691.

Hermans, E. J., Henckens, M. J., Joëls, M., & Fernández, G. (2014). Dynamic adaptation of large-scale brain networks in response to acute stressors. *Trends Neurosci., 37,* 304–314.

Hill, C. S. (2009). *Consciousness.* Cambridge: Cambridge University Press.

Hunt, R., & Ellis, H. (2004). *Fundamentals of cognitive psychology.* London: McGraw-Hill.

Hurlburt, R. T., & Heavey, C. L. (2004). To beep or not to beep: Obtaining accurate reports about awareness. *J. Conscious. Stud., 11,* 113–128.

Jack, A. I. (2013). A scientific case for conceptual dualism: The problem of consciousness and the opposing domains hypothesis. In J. Knobe, T. Lombrozo & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy: Vol. 1* (pp. 173-207). Oxford: Oxford University Press.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus-response to script-report. *Trends Cogn. Sci., 6,* 333-338.

Jack, A. I., & Roepstorff, A. (Eds.). (2003). *Trusting the subject? Volume 1.* Exeter: Imprint Academic.

Jack, A. I., & Roepstorff, A. (Eds.). (2004). *Trusting the subject? Volume 2.* Exeter: Imprint Academic.

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

Jacobs, C., & Silvanto, J. (2015). How is working memory content consciously experienced? The 'conscious copy' model of WM introspection. *Neurosci. Biobehav. Rev., 55,* 510-519.

James, W. (1890). *Principles of Psychology. Vols. 1 & 2.* New York, NY: Dover Publications.

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

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

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2013). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It! *J. Behav. Decis. Mak., 27,* 281-289.

Kiani, R., & Shadlen, M. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324,* 759–764

Kosslyn, S. M., & Thompson, W. L. (2000). Shared mechanisms in visual imagery and visual perception: Insights from cognitive neuroscience. In M. S. Gazzaniga (Ed.), *The New Cognitive Neurosciences, 2nd edition* (pp. 975-985). Cambridge, MA: MIT Press.

Kraut, R. E., & Lewis, S. H. (1982). Person perception and introspection: Awareness of the influences on behavior. *J. Pers. Soc. Psychol., 42,* 448-460.

Kroll, E., Kellicutt, M. H., & Parks, T. (1975). Rehearsal of visual and auditory stimuli while shadowing. *J. Exp. Psychol. Hum. Learn., 1,* 215-222.

Lempert, K. M., Chen, Y. L., & Fleming, S. M. (2015). Relating Pupil Dilation and Metacognitive Confidence during Auditory Decision-Making. *PLoS ONE, 10,* e0126588.

Lieberman, D. A. (1979). Behaviorism and the mind: A (limited) call for a return to introspection. *Am. Psychol., 34,* 319-333.

Lycan, W. (1987). *Consciousness.* Cambridge, MA: MIT Press/Bradford Books.

Lyons, K. E., & Zelazo, P. D. (2011). Monitoring, metacognition, and executive function. Elucidating the role of self-reflection in the development of self-regulation. *Adv. Child Dev. Behav., 40,* 379-412.

Lyons, W. (1986). *The Disappearance of Introspection.* Cambridge, MA: MIT Press.

Mandler, G. (1975). *Mind and Emotion.* New York, NY: Wiley.

Mandler, G. (2007). *A history of modern experimental psychology: From James and Wundt to cognitive science.* Cambridge, MA: MIT Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn., 21,* 422–430.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition, 115,* 303-313.

McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgm. Decis. Mak., 8,* 577-588.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *J. Exp. Psychol. Gen., 135,* 391-408.

Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: comparing metacognition for object detection and object categorization. *Atten. Percept. Psychophys., 76,* 1057–1068.

Miller, G. (1962). *Psychology, the science of mental life.* New York, NY: Harper & Row.

Miller, J., Vieweg, P., Kruize, N., & McLea, B. (2010). Subjective reports of stimulus, response, and decision times in speeded tasks: how accurate are decision time reports? *Conscious. Cogn., 19,* 1013-1036.

Muter, P. (1979). Response latencies in discriminations of recency. *J. Exp. Psychol.- Learn. Mem. Cogn., 5,* 160-169.

Neisser, U. (1967). *Cognitive Psychology.* New York, NY: Appleton-Century-Crofts.

Nelson, T.O. (1996). Consciousness and metacognition. *Am. Psychol., 51,* 102-116.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychol. Learn. Motiv., 26,* 125-173.

Nelson, T. O., & Narens, L. (1994). Why Investigate Metacognition? In J. Metcalfe and A. P. Shimamura (Eds.), *Metacognition* (pp. 1-25). Cambridge, MA: MIT Press.

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *J. Pers. Soc. Psychol., 35,* 613-624.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

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

Overgaard, M. (2006). Introspection in science. *Conscious. Cogn.,* 15, 629-633.

Overgaard, M. (2008). Introspection. *Scholarpedia, 3,* 4953.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Phil. Trans. R. Soc. B, 367,* 1287-1296.

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Conscious. Cogn., 28,* 151–160.

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Conscious. Cogn., 22,* 654-669.

Piccinini, G. (2003). Data from introspective reports: upgrading from common sense to science. *J. Conscious. Stud., 10,* 141-156.

Pierce, C.S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Science, 3*, 75-83.

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol. Rev., 117,* 864-901.

Prinz, J. (2004). The fractionation of introspection. *J. Conscious. Stud., 11,* 40-57.

Prinz, J. (2007). Mental pointing: Phenomenal knowledge without concepts. *J. Conscious. Stud., 14,* 184–211.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenol. Cogn. Sci., 3,* 1-23.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461,* 263-266.

Rich, M. C. (1979). Verbal Reports on Mental Processes: Issues of Accuracy and Awareness. *J. Theory Soc. Behav., 9,* 29-37.

Rosenthal, D. M. (2005). *Consciousness and Mind.* Oxford: Clarendon Press.

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. *Cogn. Neurosci., 1,* 165–175.

Sackur, J. (2009). L'Introspection en psychologie expérimentale. *Rev. Hist. Sci. Paris, 62,* 5-28.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Conscious Cogn., 19,* 1069-1078.

Schooler, J. (2002a). Re-representing consciousness: dissociations between consciousness and meta-consciousness. *Trends Cogn. Sci., 6,* 339-344.

Schooler, J. (2002b). Verbalization produces a transfer inappropriate processing shift. *Appl. Cogn. Psychol., 16,* 989-997.

Schooler, J. (2011). Introspecting in the Spirit of William James: Comment on Fox, Ericsson, and Best (2011). *Psychol. Bull., 137,* 345-350

Schooler, J., & Schreiber, C. (2004). Consciousness, meta-consciousness, and the paradox of introspection. *J. Conscious. Stud., 11,* 17-39.

Schwartz, B. L., & Díaz, F. (2014). Quantifying human metacognition for the neurosciences. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 9 -23). New York, NY: Springer.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Psychol. Rev., 117,* 245-273.

Schwitzgebel, E. (2011). *Perplexities of consciousness*. Cambridge, MA: MIT Press.

Schwitzgebel, E. (2012). Introspection, what? In D. Smithies & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 29-47). Oxford: Oxford University Press.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychol. Rev., 85,* 355-362.

Smith, J. D. (2009). The study of animal metacognition. Trends Cogn. Sci., 13, 389-396.

Smith, J. D., Couchman, J. J., & Beran, M. J. (2012). The highs and lows of theoretical interpretation in animal-metacognition research. *Phil. Trans. R. Soc. B, 367,* 1297–1309.

Smithies, D., & Stoljar, D (Eds.). (2012). *Introspection and Consciousness.* Oxford: Oxford University Press.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Conscious. Cogn., 4,* 1787-1792.

Sperling, G. (1960). The information available in brief visual presentations. *Psychol. Monogr., 11,* 1-29.

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science, 153,* 652-654.

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S. J. (2013). The development of metacognitive ability in adolescence. *Conscious. Cogn., 22,* 264-271.

White, P. A. (1980). Limitations on verbal reports of internal events: A Refutation of Nisbett and Wilson and of Bem. *Psychol. Rev., 87,* 105-112.

White, P. A. (1987). Causal report accuracy: retrospect and prospect. *J. Exp. Soc. Psychol., 23,* 311-315.

White, P. A. (1988). Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later. *Br. J. Psychol., 79,* 13-45.

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

Wilson, T. D. (2003). Knowing when to ask. Introspection and the adaptive unconscious. *J. Conscious. Stud., 10,* 131-140.

Wilson, T. D., & Dunn, E. (2004). Self-Knowledge: its limits, value, and potential for improvement. *Annu. Rev. Psychol., 55,* 493-518.

Wright, P., & Rip, P. D. (1981). Retrospective reports on the causes of decisions. *J. Pers. Soc. Psychol., 40,* 601-614.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B, 367,* 1310-1321.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., & Sugiura, M. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neurosci. Res., 68,* 199–206.

Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: Current and future directions. *Stud. Sci. Educ., 49,* 121-169.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Front. Integr. Neurosci., 6,* 79.