Following new task instructions:

Evidence for a dissociation between knowing and doing

Marcel Brass*, Baptist Liefooghe+, Senne Braem* & Jan De Houwer+

*Department of Experimental Psychology, Ghent University, Belgium
+ Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

In Press. Neuroscience & Biobehavioral Reviews.

Corresponding author:

Dr. Marcel Brass
Department of Experimental Psychology, Ghent University
Henri Dunantlaan 2, 9000, Ghent Belgium
Email: marcel.brass@ugent.be

Abstract

The ability to follow new instructions is crucial for acquiring behaviors and the cultural transmission of performance-related knowledge. In this article, we discuss the observation that successful instruction following seems to require both the capacity to understand verbal information, but also the ability to transform this information into a procedural format. Here we review the behavioural and neuroimaging literature on following new instructions and discuss how it contributes to our understanding of the functional mechanisms underlying instruction following. Based on this review, we distinguish three phases of instruction following. In the instruction phase, the declarative information of the task instruction is transformed into a task model consisting of a structured representation of the relevant condition-action rules. In the implementation phase, elements of this task model are transformed into a highly accessible state guiding behaviour. In the application phase, the relevant condition-action rules are applied. We discuss the boundary conditions and capacity limits of these phases, determine their neural correlates, and relate them to recent models of working memory. (170 words)

Key words:

Instruction following

Prepared reflex

Frontoparietal network

Dissociation of knowing and doing

Goal neglect

1. Introduction

In their seminal paper, Nakahara et al. (2002) scanned both macaque monkeys and humans while carrying out a simple version of the Wisconsin Card sorting task, a common neuropsychological test of executive control. They found that similar brain regions were active in monkey and man, leading to the conclusion that performance in this task must be based on similar neurocognitive mechanisms. However, in a commentary on this study, Roepstorff and Frith (2004) raised the question whether one can actually compare the task-relevant processes between both species: while monkeys were trained for months to carry out this task, humans learned the task within a few minutes. This commentary pointed to a fundamental difference between humans and other species: while humans can execute a given instruction almost instantaneously, often without practice, any other species needs effortful trial and error learning to learn new tasks. Admittedly, studying learning via instructions in animals remains difficult as there will always be a language barrier between humans and non-human animals. Furthermore, there is some evidence for simple forms of learning new skills via instructions in some animals (Whiten et al., 1999). However, it is unquestionable that humans have a uniquely developed ability of instruction following that allows for easy cultural transmission of rules, forms the basis for most technological developments of modern societies, and separates them from other non-human animals. This raises the question why humans can follow instructions so easily while this ability is very restricted in other animals?

For one, it seems rather straightforward that our language capacity to represent and understand abstract content in a verbal format (Deacon, 1997) is vital to instruction following. However, as we will argue below, the ability to *understand* instructions is a

necessary but insufficient condition to successfully follow instructions. Following new instructions not only requires understanding these instructions but also the translation of these instructions into actual behavior. For example, it is one thing to read and understand the instruction manual of your new smartphone while it is another to actually operate it. Such dissociation between understanding instructions ('knowing) and following instructions ('doing') has been first proposed more than half a century ago by demonstrating that frontal patients sometimes fail to follow instructions even though they are perfectly able to recapitulate what they were supposed to do (Milner, 1963).

Whereas the dissociation between knowing and doing seems straightforward at first, understanding the neurocognitive dynamics at the origin of this dissociation has become a major challenge in recent years (Demanet et al., 2016; Duncan et al., 1996; Duncan et al., 2008; Liefooghe et al., 2012; Muhle-Karbe et al., 2016). Accordingly, the aim of the current review is to provide an overview of the current state of knowledge on the dissociation between knowing and doing. To this end, three research domains are considered. First, research on 'goal neglect', which argues that participants sometimes fail to implement specific instructions even though they are perfectly able to remember them (Bhandari and Duncan, 2014; Duncan et al., 1996; Duncan et al., 2008). Second, behavioral research on the 'prepared reflex' (Hommel, 2000) or 'intention-based reflexivity' (Meiran et al., 2012) which examines the automatic effect of instructions to respond to stimuli. One important question within this line of research is whether instruction-based automatic response activation depends on the intention to implement a specific instruction or whether it is enough to simply remember the instruction (Liefooghe et al., 2012). Finally, we review brain imaging research, which tried to reveal the functional neuroanatomy of instruction following (Brass et al., 2009; Hartstra et al., 2011; Ruge and Wolfensteller, 2010) and attempted to dissociate

between the implementation and memorization of instructions (Demanet et al., 2016; Muhle-Karbe et al., 2016).

Based on this literature review we will argue that instruction following can be decomposed in three different phases: the instruction phase, the implementation phase and the application phase. The instruction phase refers to the translation of the instruction into a task model. Research on goal neglect and neuroimaging research on complex rule following has extensively investigated this phase (Cole et al., 2010; Duncan et al., 2008). The implementation phase refers to active maintenance of specific aspects of the task model that need to be implemented. The literature on instruction-based congruency and some imaging studies have investigated this phase (Liefooghe et al., 2012; Muhle-Karbe et al., 2016). Finally, the application phase refers to the execution of the instruction. While the application phase is not at the core of the current review, we discuss some interesting findings that are relevant for our broader understanding of instruction following (Bhandari and Duncan, 2014; Ruge and Wolfensteller, 2010).

2. The study of goal neglect

First evidence for the idea that instruction following goes beyond instruction understanding stems from neuropsychological research in prefrontal patients and refers to the dissociation of 'knowing and doing' (Luria, 1980; Milner, 1963). Milner (1963) reported that her frontal leucotomy patients accompany their incorrect actions with correct verbal comments. Teuber (1964) referred to this as the 'curious dissociation of knowing and doing' (page, 333). According to Luria (1980), this dissociation between knowing and doing is neither caused by a lack of instruction understanding nor by motor deficits. While these

findings have been discussed in the literature for decades, little systematic research was conducted to further understand the neurocognitive mechanisms that underlie this dissociation.

2.1. Goal neglect in the cognitive literature

In order to fill this empirical gap, Duncan and colleagues (Duncan et al., 1995; Duncan et al., 1996) introduced the concept of *goal neglect* which tried to capture the dissociation between knowing and doing on an experimental level. Goal neglect is defined by three properties (Bhandari and Duncan, 2014): (a) it reflects a gross failure to follow task rules; (b) performance is limited by the complexity of task instructions rather than by the complexity of task execution; and (c) performance is not explained by a failure of explicit rule recall.

Duncan et al. (1996) were the first to investigate goal neglect by using a letter-monitoring task (Figure 1). In this task, a pair of letters or a pair of numbers is presented in each trial. One character is presented on the left side of the screen, the other on the right side of the screen. At the onset of the task, participants are cued which screen side is relevant and they have to read out loud the letters that are presented on that side. Digits on the same side and letters on the other side have to be ignored (Figure 1). After a few trials, participants receive a symbol that either indicates that they have to switch to the other side or continue the task on the same side. Duncan et al. (1996) observed that some participants did not switch to the other side when they were required to do so, even though they were able to repeat the instructions verbally at the end of the task. Duncan et al. (1996) interpreted this failure to follow the instruction as 'goal neglect' and related it to general intelligence (g) and frontal brain damage. In order to investigate whether goal neglect also depended on task difficulty, a secondary task was introduced during the letter-monitoring

task. During the stream of character pairs a dot could briefly flash either above or below the pairs and participants had to respond to dot position by using a particular response key. The extent by which task difficulty increased goal neglect, depended on the point in time at which the secondary task demand was introduced during the instruction phase. If the additional demand was introduced in the beginning of the instruction phase, it was hardly neglected. If it was introduced at the end of the instruction phase, it was ignored more often.

In a follow-up study, (Duncan et al., 2008) further investigated the relation between task difficulty, instruction structure and goal neglect. In a first experiment, the number of stimuli that had to be monitored was increased. Interestingly, goal neglect was not larger in this more difficult variant of the letter-monitoring task. In another experiment, a secondary task was introduced. Goal neglect was only affected by the secondary task when both the primary and secondary tasks were introduced in the beginning of the experiment. When the secondary task was introduced separately during the experiment, goal neglect was largely unaffected. This indicates that it is the complexity of the task-model created in the instruction phase that matters for goal neglect and not the task difficulty per se. Based on their results Duncan et al. (2008) concluded that there is ‘a limit in constructing and maintaining what we call a task model—a working-memory description of relevant facts, rules, and requirements used to control current behavior’ (Duncan et al., 2008). Bhandari and Duncan (2014) added a number of important aspects to the study of goal neglect. First, goal neglect strongly depends on the chunking of complex task instructions. The task instruction has to be parsed into a set of chunks or subtasks that can be easily executed. The more efficient such chunking is, the less goal neglect will emerge. Second, for goal neglect only the complexity of the relevant task is crucial. Increasing the complexity of another task

does not affect goal neglect in the current task. Finally, in the beginning of rule application participants respond relatively unstably. Only after a few trials, they settle on a specific strategy indicating that the task model stabilizes during its application (Bhandari and Duncan, 2014).

Whereas the previous research primarily relates to the instruction phase and the construction of the task model, De Jong et al. (1999) argued that goal neglect can also arise from a failure to implement the task model. In contrast to the interpretation discussed above, goal neglect is thus not always a consequence of the failure to construct a task model but rather an occasional failure to implement this task model. Here, implementation refers to getting prepared to apply elements of the task model. De Jong et al. (1999) emphasized intra-individual variations in implementing a specific element of the task model in interference tasks. Even though participants might in principle be perfectly able to implement the task model, they sometimes fail because this implementation process is effortful. In situations where most of the time it is not necessary to implement the specific element of the task model, the implementation might be compromised (McVay and Kane, 2009).

2.2. *Goal neglect in the developmental literature*

A similar dissociation between knowing and doing has been discussed in the developmental literature (Diamond, 1991; Zelazo et al., 1996; Zelazo and Reznick, 1991). For example, Zelazo (2004) describes what he calls a ‘knowing-action’ dissociation in children performing a card sorting task. In the dimensional change card sorting task (Zelazo et al., 1996), children have to match cards according to one dimension of the card (e.g., color).

After a number of trials, the matching rule changes so that they have to match the cards according to another dimension (e.g., shape). Up to the age of four, children tend to continue sorting cards according to the pre-switch rule while at the same time being aware that they are supposed to change the sorting rule (Zelazo et al., 1996; Zelazo and Reznick, 1991). Interestingly, Zelazo et al. (1996) showed that such a dissociation between knowing and doing occurs already after a single pre-switch trial, making it unlikely that the failure to implement the new rule is caused by the habitual character of the old rule. He argues for a level of consciousness interpretation of the dissociation. While both rules (the pre-switch rule and the post-switch rule) are represented on a specific level of consciousness, they cannot be integrated and therefore children continue to follow the pre-switch rule. From this perspective the problem is due to a failure to generate a hierarchical rule structure when building the task model (Zelazo et al., 1996).

These developmental findings differ in one crucial aspect from the goal neglect findings reported by (Duncan et al., 1996), namely that they involve a switch of the sorting dimension (Marcovitch et al., 2010). This introduces a number of potential causes for the failure to implement the new rule, such as a failure to inhibit the old rule or priming of the old rule by the stimulus (Marcovitch et al., 2010). Later studies (Roberts and Anderson, 2014; Towse et al., 2007) tested goal neglect in children using very similar paradigms as the one introduced by (Duncan et al., 1996). Towse et al. (2007) found strong indications of goal neglect in preschool children with such a paradigm. Roberts and Anderson (2014) further showed that goal neglect in children is sensitive to the same complexity manipulation as in adults: adding an additional task component leads to more goal neglect in children even when the requirements during task execution are identical.

2.3. Summary

The finding of goal neglect suggests that even though participants have a declarative representation of the task instruction they fail to follow parts of the instruction under specific circumstances. Goal neglect increases when the instructions to carry out the actual task (the task model) become more complex. Goal neglect is neither affected by the difficulty of the task at task execution, nor is it affected by the whole set of instructions as long as they are related to different task models. Furthermore, goal neglect seems to heavily depend on the way task instructions are transformed into smaller chunks of information that guide performance. These observations strongly support the idea that a declarative representation of a task instruction is not sufficient for instruction following. A certain form of transformation has to take place in order to make the declarative representation effective for execution. There seems to be a capacity limit of how many elements of a task model can be transformed into such a task effective representation. This might explain why goal neglect is correlated with working memory capacity (McVay and Kane, 2009), is stronger in young children (Zelazo et al., 1996) and is correlated with general intelligence (Bhandari and Duncan, 2014; Duncan et al., 1996; Duncan et al., 2008). Furthermore, hierarchical structuring and chunking of information seems to be crucial for the construction of the task model. The instruction has to be parsed into a set of rule-like representations that are presumably represented in the form of condition-action rules specifying what happens when a specific stimulus or cue occurs. Finally, the task model stabilizes during the first application trials (Bhandari and Duncan, 2014; Zelazo et al., 1996).

3. Automatic influences of instructions on behavior: Instruction-based congruency effects

In the previous section we reviewed evidence for the idea that a simple declarative representation of a task instruction is not sufficient for instruction following. This evidence primarily deals with the way in which participants transform an instruction into a task model that guides behavior. Furthermore, it investigates the specific format in which an instruction has to be given in order to be implemented. The studies that we review in this section involve more simple instructions. Therefore, the instruction phase and the formation of a task model plays a less important role here. Instead these studies focus on the cognitive mechanism of *implementing* task instructions by investigating the automatic influence of instructed tasks on behavior. Here, the notion of 'automatic' refers to the idea that an instructed stimulus-response (S-R) mapping induces a response tendency without the participant having the intention to execute the instructed S-R mapping at the moment of task execution. Showing such automatic influences of instructions on behavior provides further evidence for the idea that instruction following involves a transformation from a mere declarative representation (i.e. understanding) of the instruction to a representation that is geared towards the execution of this instruction.

The idea that instructions can exert such an automatic influence on behavior goes back to authors like Exner (1879) and Woodworth (1938). In his seminal chapter, Hommel (2000) reviewed modern and historical literature supporting the concept of the so-called 'prepared reflex' (Woodworth, 1938). The basic idea is that participants prepare for a task by setting oneself in a state that ensures that responses are carried out efficiently (Hommel, 2000). While this state of preparation is voluntary and effortful, once it is accomplished, the

S-R translation itself is more or less automatic and effortless. Meiran et al. (2012) propose that when instructed S-R mappings are intended to be executed, S-R associations can be formed without overt practice.

There are a number of conditions that have to be met in order to draw the conclusion that an instructed task exerts an automatic influence on behavior (Meiran et al., 2015a). First, one has to ensure that the instructed S-R mappings do not have a learning history. In other words, one has to investigate paradigms where the instructed S-R mappings are either new on every trial or are never executed. Otherwise one can argue that it is the prior execution of the S-R mapping rather than the instruction that exerts the influence. Furthermore, the task instruction needs to involve at least two S-R mappings to avoid that participants simply prepare the execution of a specific response rather than an S-R mapping.

The most common way to investigate the notion of the prepared reflex is by inducing interference effects via task instructions (instruction-based congruency, IBC). Following classical methodology from the cognitive control literature there are different ways to induce such interference. One way is through a flanker paradigm (Eriksen and Eriksen, 1974). Participants have to respond to a central stimulus while ignoring laterally presented stimuli. The response induced by the lateral stimuli can either be congruent or incongruent to the instructed response. In contrast to classical flanker tasks, the lateral stimuli in IBC studies are newly instructed and never applied (Cohen-Kdoshay and Meiran, 2007, 2009). Another way to induce interference via instructions is by using a dual-task like situation (Liefooghe et al., 2013; Liefooghe et al., 2012; Waszak et al., 2008). Here, participants are instructed to respond to a specific property of a stimulus in task A. After the instruction, participants carry out another task B in which they have to respond to a different property of the same stimulus. The response induced by the first instruction can then either be congruent or

incongruent with the response required for the now relevant second task (i.e., a task-rule congruency effect). Again, the task rules of the irrelevant task have never been applied. Either participants alternate between the two tasks (Waszak et al., 2008) or the secondary task is embedded in the primary task (Liefooghe et al., 2013; Liefooghe et al., 2012; Meiran et al., 2015a). In the following paragraphs we will summarize evidence for IBC from these different approaches.

3.1. Evidence from flanker tasks

Cohen-Kdoshay and Meiran (2007) for the first time showed an IBC effect in a flanker task. At the beginning of each experimental block, participants were presented with a new stimulus-set and with a new pair of category-to-response mappings (e.g., if a number is even, then press left; if a number is odd, then press right). However, only a subset of the instructed stimuli actually appeared as targets, and hence required the application of the instructed mappings. The remaining stimuli exclusively served as distracting flankers. Cohen-Cohen-Kdoshay and Meiran (2007) observed a flanker interference effect for flankers that were newly instructed and never executed overtly before. Such effect even occurred when participants were strongly discouraged to attend the flanker stimuli (e.g. by increasing the distance between flankers and the target). These IBC effects were furthermore sensitive to working-memory load. In each experimental block, a secondary Go/No-Go task was added to the flanker task. During the flanker task one or more digits were presented and participants had to press the spacebar, whenever the digit matched a particular criterion (e.g., divisible by 4, larger than 5, odd,…). This criterion changed across blocks. When such secondary task was added to the flanker task, the IBC effect disappeared. In line with the findings on goal neglect, one can argue that the secondary task increased the complexity of the task model

and prevented the S-R mappings from exerting their automatic influence. In a follow-up study, Meiran and Cohen-Kdoshay (2012) controlled whether their initial results were due to an increased load on the buffer capacity of working memory or to an increased multitasking demand. They demonstrated that IBC effects were only absent when the criterion of the Go/No-Go task changed across blocks, while they were still present when the criterion remained constant, which requires a similar degree of multitasking but induces a smaller load on the buffer capacity of working memory. This latter finding indicates that the secondary task only influenced the IBC effect when it had to be updated together with each new instruction.

One potential alternative interpretation for the IBC effects reported by Cohen-Kdoshay and Meiran (2007) relates to the use of category-response mappings. In a follow-up study, Cohen-Kdoshay and Meiran (2009) argued that performance in the first trials of each experimental block may have led to the formation of category-to-response associations in long-term memory and the automatic retrieval of these associations may induce a flanker effect, even for flanker stimuli that were never responded to. In order to circumvent this problem, Cohen-Kdoshay and Meiran (2009) adapted their initial procedure such that they could measure a flanker effect on the very first trial following the presentation of the instructions. Their results indicated the presence of a significant first-trial flanker effect, thus offering a stronger case for the hypothesis that instructions can be implemented instantaneously. Finally, Wenke et al. (2015) replicated the instruction-based flanker effect and compared it to a classical flanker effect. Overall, the instruction-based flanker effect was smaller than the classical flanker effect and decreased over the course of an experimental block, whereas the classical flanker effect remained constant. To summarize, there is strong evidence from IBC effects in the flanker task for the idea that instructions can lead to

automatic activation of responses. These effects are sensitive to a working memory manipulation that increases the complexity of the instructed task.

3.2. Evidence from dual task approaches

Another way to investigate IBC effects is by creating task-rule congruency effects on the basis of instructions (De Houwer et al., 2005; Liefooghe et al., 2012; Waszak et al., 2008). The basic idea is to instruct specific task rules for one task and then test whether these instructed task rules exert an influence on another task when they apply on an irrelevant stimulus dimension. In Waszak et al. (2008) participants had to switch between a shape and a color task on bivalent stimuli (stimuli that contain both stimulus dimensions). In the shape task, participants had to respond to the shape of a stimulus by pressing a left or a right key. In the color task, participants had to respond to the color of the stimulus. In the instruction phase, four colors and four shapes were mapped onto left and right responses. Four additional values (two colors and two shapes) were not mapped to any response. In the test phase, participants had to switch between the color and the shape task. Importantly, only two of the four values of each task were presented on the relevant stimulus dimension and thus were overtly responded to (applied stimuli). The other two values were only presented on the irrelevant stimulus dimension and were therefore never responded (instructed stimuli). The four values that were not related to a response (uninstructed stimuli) were also selectively presented on the irrelevant stimulus dimension. In this case, univalent stimuli were thus created on which only one task applies. With this design it was possible to test whether stimuli that were related to a response via instruction could induce interference, even when they were never responded to overtly before. The results showed that compared to uninstructed stimuli, both applied and instructed mappings induced interference when

they were presented on the irrelevant stimulus dimension. However, only applied mappings induced a task-rule congruency effect. This result indicates that instructed S-R mappings exert an unspecific influence in this paradigm.

Liefooghe et al. (2012) introduced a paradigm in which two tasks were embedded and therefore could not be processed independently (Figure 2). Participants were first instructed with two S-R mappings of an inducer task (e.g., respond left to the letter k and right to the letter l). These S-R mappings were newly instructed on every run of trials. Immediately following upon the presentation of these mappings, participants had to carry out a diagnostic task. The diagnostic task referred to a different stimulus dimension (e.g., respond left to italic letters and right to upright letters) but used the same stimuli and responses as the inducer task. Only after participants had executed a few trials of the diagnostic task, they could execute the inducer task. Liefooghe et al. (2012) found that instructed S-R mappings of the inducer task caused an interference effect on the diagnostic task. Participants were faster in the diagnostic task when the response of the inducer task (e.g. respond left to k) matched the response in the diagnostic task (e.g. respond left to italic letters) compared to the condition were the response in the inducer task (e.g. respond right to l) was incongruent to the response in the diagnostic task (e.g. respond left to italic letters). It is important to note that whereas the studies of Waszak et al. (2008) and Liefooghe et al. (2012) focussed on the same type of effect, namely instruction-based task-rule congruency effects in dual-task situations, they obtained different results. Possibly, the reason for this is that both dual-task studies used different experimental parameters. First, in the procedure of Liefooghe et al. (2012), the diagnostic task is embedded in the inducer task and participants are encouraged to actively maintain the mappings of the inducer task, while completing the diagnostic task. In the paradigm of Waszak et al. (2008), both tasks are

presented sequentially and in a discrete manner. On each trial attention is drawn towards one task and the S-R mappings of the alternative task do not have to be actively maintained. This reminds on the observation of Bhandari and Duncan (2014) that only the now relevant task model exerts an influence on behavior. Second, in the paradigm of Waszak et al. (2008), participants may have learned that the merely instructed S-R mappings never had to be applied anyway, and therefore lost the intention to maintain these S-R mappings. Third, in the study of Waszak et al. (2008) eight S-R mappings were instructed, whereas Liefooghe et al. (2012) did only instruct 2 S-R mappings for the inducer task and 2 category-response mappings for the diagnostic task. Possibly, the high number of S-R mappings to be maintained in the study of Waszak et al. (2008), made it difficult to keep these S-R mappings in a highly activated state, resulting in the absence of an instruction-based congruency effect.

In a follow-up study, Liefooghe et al. (2013) further manipulated the degree to which participants prepared the inducer task. In two experiments, they could show that an IBC effect occurred in the diagnostic task only when participants were motivated to actively prepare the inducer task. Similar findings were reported by Meiran et al. (2015a) in studies using the NEXT paradigm. In this paradigm participants are also instructed to carry out an inducer task (e.g. X press left, Y press right). However, they were asked to respond only to the stimuli when they were printed in green. When the stimulus of the inducer task was red, participants had to advance to the next trial by pressing a predefined response key (either the left or the right key). This NEXT response was either congruent or incongruent to the responses defined by the instructed S-R mappings. The NEXT procedure also elicits robust IBC effects and this even on the very first NEXT response that follows the instructed S-R mappings. Importantly, the “next” response is not a choice response but rather a simple

go/no-go response. In line with Liefooghe et al. (2013), Meiran et al. (2015a) also observed that the IBC is a function of the amount of preparation the inducer task receives.

A recent study by Braem et al. (2016) on IBC revealed an interesting parallel to the task complexity effect observed in goal neglect and provides further evidence for the idea that IBC effects can also reveal a type of goal neglect. In this study, task complexity was investigated by testing whether the IBC effect can be observed in a context-specific manner. Within the procedure of Liefooghe et al. (2012), participants were instructed that the inducer task would be relevant in one context only (i.e., location on the screen), after which they were asked to perform the diagnostic task in either the same or in another context. This way, Braem et al. (2016) measured whether the IBC effect is restricted to the instructed context, which would suggest that people can integrate the instructed S-R mappings and the instructed task context. This study was motivated by previous observations that overtly practicing a task in a particular context, results in binding between task and context, which causes interference from these tasks to emerge in a context-specific manner (Abrahamse et al., 2016). Interestingly, Braem et al. (2016) observed that IBC effects are not modulated by the context, suggesting that context-specificity cannot be induced on the basis of instructions. Further analyses indicated that people, however, clearly remembered both the instructed mappings and the instructed context, but failed to fully integrate both components. In contrast, moderately practicing the inducer task was sufficient to induce context-specific congruency effects, which suggests that context and rules are only integrated through overt practice.

Although most research on IBC effects focuses on the “doing” part of the dissociation between knowing and doing, research on IBC has also more directly considered this dissociation on itself. Whereas Liefooghe et al. (2012) observed an IBC effect when

participants were instructed to execute the S-R mappings of the inducer task, they failed do find an IBC effect when participants were instructed to memorize instructed S-R mappings for a future recognition test. This led to the conclusion that IBC effects are restricted to situations where participants form the intention to act, which strongly supports the idea that an additional transformation is necessary to implement instructions. Liefooghe et al. (2012) more explicitly proposed that the intention to execute the instructed S-R mappings, results into the formation of a procedural representation, which induces IBC effects. In contrast, if such demand is absent (e.g., for memorization) only declarative representations are created, which do not induce IBC. However, although the distinction between procedural and declarative representations proposed by Liefooghe et al. (2012), fits nicely with the idea of a dissociation of knowing and doing, a recent study by Liefooghe and De Houwer (subm.) questions whether their initial proposal was valid. In a series of experiments they could show that if the memory condition was made more challenging by introducing a response deadline for the inducer task or by forcing participants to memorize all mappings, IBC effects could be induced without the intention to execute the instructed S-R mappings. In addition, Liefooghe and De Houwer (subm.) added one caveat to research on IBC, namely that in all experiments that were conducted on IBC, S-R instructions always involve left and right responses. Because it has been demonstrated that the word left activates a left response and the word right activates the right response (Bundt et al., 2015), relating a stimulus to the concept left or right might be sufficient to induce a response tendency without necessarily requiring the intention to respond to the stimulus. Taken together, the findings of Liefooghe and De Houwer (subm.) raise the possibility that IBC does not necessarily reflect the presence of procedural representation. Instead, the IBC can also be based on the declarative representation, when it is kept highly accessible in working memory and comprises salient

semantic concepts, such as left and right. It becomes clear that further research has to demonstrate whether IBC can be also observed when the semantic concepts that are used in the instruction are not directly linked to response concepts.

3.3. Summary

Taken together, the IBC effects partly support the idea that following new task instructions requires a transformation into a procedural format. When participants form the intention to execute a task instruction, automatic effects of these instructions are stronger than in a situation where they merely intent to remember the instruction (Liefooghe et al., 2013). Furthermore, there is evidence that the complexity of the instruction has an influence on this transformation step and can either lead to a form of goal neglect in IBC (Braem et al., 2016) or even to the elimination of the IBC effect (Cohen-Kdoshay and Meiran, 2007). The degree to which task instructions exert an influence on behavior seems to depend on the degree of preparation (Liefooghe et al., 2013). However, recent research leaves the possibility open that IBC effects can occur even when participants do not have the intention to execute the instructions (Liefooghe and De Houwer, subm.), which challenges the validity of the IBC effect as a proxy of a procedural representation.

4. Neural evidence for the dissociation between knowing and doing

In the previous section, we have primarily reported behavioral evidence for the idea that instruction following of new instructions is characterized by a dissociation of knowing and doing, and requires the transformation of the instruction from a declarative into a procedural format. However, the neuroimaging literature on instruction following is similarly

suggestive of a dissociation between knowing and doing, and provides a hint that such implementation steps can be traced back to specific brain regions. The patient studies discussed above already argued that the dissociation between knowing and doing is primarily related to frontal brain damage (Duncan et al., 1996; Milner, 1963; Teuber, 1964). However, over the last few years, an increasing number of studies also started to investigate instruction following using different imaging techniques (Cole et al., 2010; Dumontheil et al., 2011; Everaert et al., 2014; Hartstra et al., 2011; Ruge and Wolfensteller, 2010).

First, we discuss studies that used neural indices of motor activation to show that, in line with research on IBC, newly instructed S-R mappings can directly activate the motor system (Everaert et al., 2014; Meiran et al., 2015b). Second, we consider studies that systematically compared novel instruction presentation with the presentation of instructions that have been applied before (Brass et al., 2009; Cole et al., 2010; Hartstra et al., 2011; Ruge and Wolfensteller, 2010). There, the crucial question is whether the functional neuroanatomy of implementing new instructions differs from the brain regions involved in following already applied instructions. Whereas these studies do not yet address the question whether instruction following requires a transformation of new instructions into a procedural format they help identifying the brain networks that are involved in following new instructions. Finally, we discuss studies that contrasted the memorization and implementation of instructions (Demanet et al., 2016; Hartstra et al., 2011; Muhle-Karbe et al., 2016). These studies directly addressed the question whether there is neural evidence for the dissociation between knowing and doing.

4.1. Neural evidence for motor activation through instructions

Research on IBC effects strongly suggests that when participants form the intention to respond to a stimulus, that stimulus automatically triggers the instructed response. One way to further test the hypothesis that the IBC effect reflects automatic response activation is by measuring motor evoked potentials with EEG. In particular, the lateralized readiness potential (LRP) allows to investigate whether IBC is caused by a response tendency to automatically apply the instruction. The LRP is a response-locked event-related component that occurs contralateral to the responding hand. However, the LRP can also be induced by a covered response tendency. In classical interference tasks with lateralized responses such as the flanker or the Simon task it has been demonstrated that LRPs are larger in congruent than in incongruent trials (Eimer, 1995). Furthermore, on incongruent trials a small LRP ipsilateral to the response hand can be observed, induced by the interfering response tendency.

A first study that investigated IBC using the LRP was carried out by Everaert et al. (2014). These authors used the procedure of Liefooghe et al. (2012) to investigate whether the responses instructed in the inducer task evoke LRPs in the diagnostic task, which would offer a stronger case for the hypothesis that the IBC effect is based on automatic response activation. It was predicted that if the instructed responses of the inducer task are automatically activated, LRPs in the diagnostic task should be larger when they are congruent with the responses of the inducer task than when they are incongruent. The results indicated that congruent responses in the diagnostic task led to larger LRPs than incongruent responses. Furthermore, there was an initial ipsilateral LRP on incongruent trials, suggesting that the motor cortex of the instructed mapping was activated first (ipsilateral to

the responding hand) and subsequently overwritten by motor cortex activation of the response hand (contralateral to the responding hand).

Independently, Meiran et al. (2015b) carried out a similar EEG study. In their study, participants had to carry out a go/no-go task. First, participants received instructions to respond with a left response to one letter and with a right response to another. However, participants should only respond when the letter was green. In the first few trials after the instruction, the letters were red, so participants were instructed not to respond. During these no-go trials LRPs were measured. Meiran et al. (2015b) found LRPs in accordance with the instructed response. Interestingly, these LRPs only occurred in the first trial after the instruction and quickly disappeared in subsequent trials. Furthermore, the LRP in the first no-go trial was correlated with reaction time in the first go trial, indicating that the LRP effect indexed motor preparedness.

Together, these studies clearly support the idea that instructed S-R mappings can lead to an automatic activation of the instructed motor response and further strengthen the hypothesis that following instructions requires the implementation of a procedural representation, which enables prepared reflexes. Next, we turn to fMRI studies that explored the neural correlates of this implementation process. To get at these neural signatures, researchers used designs were (1) new instructions were contrasted with trained instructions, and (2) the implementation of response-related representations is compared with those that do not require a response. We discuss both approaches separately.

4.2. Comparing neural activity for newly instructed and practiced S-R mappings

The first fMRI study comparing new versus applied instructions was carried out by Brass et al. (2009), who used the aforementioned dual-task design by Waszak et al. (2008).

As outlined above, participants had to alternate between a shape and a color task. On the irrelevant stimulus dimension merely instructed, applied or uninstructed stimulus values were presented. The basic logic of this study was to investigate interference-related brain activation induced by stimulus values on the irrelevant stimulus dimension. Thus, the study indexes the influence of instructions indirectly by the interference they induce on the neural level. When comparing the applied versus the uninstructed condition, a typical conflict-related brain network was observed consisting of the ACC, the preSMA, the frontolateral cortex and the parietal cortex. When contrasting the instructed with the uninstructed condition, only parts of this conflict-related network was observed including the preSMA and frontolateral cortex. Interestingly, no activation was found in the ACC. A direct contrast of the applied and instructed condition yielded activation in the ACC, the inferior parietal cortex, and the dorsal premotor cortex, indicating that overcoming conflict from applied S-R mappings leads to additional activation in conflict related areas (Ridderinkhof et al., 2004). These data support the idea that instructed mappings share some properties with already applied S-R mappings and that they exert an automatic influence on behavior. However, it also demonstrates that applied mappings exert a stronger automatic influence than merely instructed mappings. In line with task-switching studies including univalent stimuli on which only one task can be applied (Rogers and Monsell, 1995; Steinhauser and Hubner, 2007), one possibility is that the instructed S-R mappings only lead to the activation of a more general representation of the task they are associated with (i.e. a general task conflict). In contrast, the stronger interference elicited by applied S-R mappings may suggest that applied S-R mappings additionally lead to the activation of specific responses within the task representation they belong to. The additional activation in the ACC for overcoming interference from applied compared to merely instructed mappings on the irrelevant

stimulus dimension may be suggestive of such additional response conflict and is most likely related to what was observed on the behavioural level in this paradigm, namely IBC effects for applied but not for instructed mappings

Partly motivated by the dissociation of applied and instructed mappings, Brass et al. (2009) introduced a model of instruction following that distinguishes three stages of instruction implementation and following. The first stage is a linguistic stage were the instructions are represented on a semantic level. This stage is followed by a sublinguistic stage where the instructions can already exert an automatic influence on behavior but do not lead to full-blown response activation. The third stage is the sensorimotor stage, during which the instructed S-R mappings have travelled deep into the sensorimotor system. Interestingly, as outlined above, no IBC effect for merely instructed mappings was found in the original behavioral study by Waszak et al. (2008) and the imaging study suggesting that in this experimental setup the instructions do not directly activate the instructed motor response (Meiran et al., 2012).

While the previous fMRI study investigated the implementation of instructed S-R mappings indirectly through interference caused by instructed and practiced S-R mappings, Ruge and Wolfensteller (2010) directly investigated the implementation of newly instructed S-R mappings. In their study, participants got four new stimuli in the instruction phase of each trial. Two stimuli were related to a left response and two stimuli were related to a right response. After the instruction phase, stimuli were repeatedly presented allowing them to investigate the dynamics of instruction following over practice. A frontoparietal network was active in the instruction phase, before participants applied the S-R mappings for the first time, thought to be responsible for the implementation of new task instructions. Parts of this frontoparietal network showed a strong decline in activity following practice, sometimes

even after the first application of a new mapping. Furthermore, activation in the frontolateral and parietal cortex during the instruction phase was reliably correlated with task performance during practice. These results show that the degree of preparation in the instruction phase determines how easily instructed S-R mappings can be implemented during practice. Furthermore, the strong activation decline in some of the instruction-related brain regions further suggests that these regions have only a very transient role in guiding the implementation of instructions.

Another study directly comparing instructed and implemented S-R mappings was carried out by Hartstra et al. (2011). In their study, two S-R mappings were presented in the instruction phase of each trial. The instruction presentation phase was separated by a variable jitter interval allowing to model the instruction phase and application phase independently. Importantly, half of the trials were already applied in a training session that preceded the scanning session. This allowed to directly compare the instruction phase of newly and already trained S-R mappings. Contrasting brain activation in the instruction phase of newly instructed and trained mappings yielded significant activations in the frontolateral cortex, more precisely in the inferior frontal junction area (IFJ). However, as will be outlined below, this activation was not specific to new S-R instructions. In a follow up study, (Hartstra et al., 2012) tried to dissociate brain areas that are involved in creating the link between the stimulus and response from brain areas that are related to activating the motor system when new instructions are given. They found the frontolateral cortex along the inferior frontal sulcus to be related to S-R formation.

Finally, a series of studies by Cole and colleagues (Cole et al., 2010; Cole et al., 2011) employed a paradigm in which participants had to combine three types of rules (four sensori-semantic rules, four logic rules and four response rules) to determine the correct

response. The combination of these three rules allowed them to create 64 unique task instructions. Given the complexity of the instructions, these studies are more similar to the later studies on goal neglect (Bhandari and Duncan, 2014) than the studies on IBC. In their first combined fMRI and MEG study (Cole et al., 2010), they presented 60 of these instructions for the first time in the scanning session whereas four had been practiced during a training session. When comparing new instructions and trained instructions during the instruction phase they found the typical frontoparietal network to be active, including the frontolateral cortex (DLPFC and IFJ) and parietal regions. Interestingly, like in Ruge and Wolfensteller (2010), brain activation in these areas strongly declined for new rules when the task instructions were applied. More anterior parts of the prefrontal cortex showed an opposite pattern, leading Cole et al. (2010) to conclude that whereas the DLPFC might be responsible for the formation of (simple) task-rule representations, the anterior prefrontal cortex is involved in creating more higher-level integrated task representations. Finally, there is a study by Dumontheil et al. (2011) investigating instruction following in a design similar to the goal neglect study by Bhandari and Duncan (2014). In this study participants were scanned while receiving a series of complex rules. Furthermore, the effect of rule complexity on brain activation was investigated. Dumontheil et al. (2011) showed that a frontoparietal network was involved in the building of the task model and was active whenever a new rule was presented. This network increased activity with the number of rules added.

Taken together these studies suggest that in the instruction phase a frontoparietal network is more active for newly instructed compared to practiced instructions. This activation also correlated with performance when implementing new instructions. Moreover, the involvement of this network strongly decreases as soon as the instructions have been

applied a few times, suggesting that the frontoparietal network is primarily responsible for building a new task model and for maintaining the S-R relations until they become more habitual.

4.3. *Comparing neural activity for implementing or memorizing instructions*

While the studies discussed so far provide information about the neural implementation of new task instructions, they do not directly address the question whether there is a difference in brain activity between instructions to execute S-R mappings and instructions that have to be simply memorized without the intention to execute them. To our knowledge only three studies have tried to address this question using brain imaging. The first study was already briefly discussed above and was carried out by Hartstra et al. (2011). As outlined before, on each trial two S-R instructions were given. In addition to the newly instructed and trained S-R mappings, participants were also presented with so called object-color mappings (O-C mappings). O-C mappings relate an object (e.g. a jacket) to a color (blue). The information content is very similar to S-R mappings. However, in the O-C task participants simply had to evaluate whether a colored object (e.g. a blue jacket) that was presented in the application phase, matched the mapping presented in the instruction phase. Hence, the mappings were not directly related to a response but only provided relational information about two properties of an object (form and color). When comparing the instruction phase of S-R and O-C mappings independently of whether they were new or trained, activation in premotor and parietal cortex was found. This indicates that in the instruction phase of S-R mappings the motor system is already activated. Importantly, however, no brain region was specifically active for newly instructed S-R mappings that was not active for newly instructed OC mappings. This seems to suggest that frontoparietal

activation that has been reported when comparing new versus applied instructions is not specifically related to the implementation of instructed behavior, but reflects more general processes related to maintain the information conveyed by the instruction screen.

However, one problem of the previous study was that the information provided in the instruction phase for S-R mappings differs from the information provided for O-C mappings. In order to address this issue, Demanet et al. (2016) carried out a study in which they used the logic introduced by Liefooghe et al. (2012) where participants were either instructed to implement or to memorize the mappings that were given in the instruction phase (Figure 3a). In a between-subject design they show that a region in the right frontolateral cortex was more active in the group that had to implement the given mappings compared to the group that had to remember the mappings (Figure 3b). The frontolateral activation in the implementation group was correlated with performance when implementing the mappings.

Finally, there is a recent study (Muhle-Karbe et al., 2016) that compared implementation and memory instructions using multivoxel pattern analysis (MVPA). MVPA allows to identify fine grained pattern of voxel activity in the brain (Haxby et al., 2001; Haynes and Rees, 2006) . In their study, participants were instructed to either memorize or implement simple S-R mappings that either combined two houses with two responses or two faces with two responses. In both implementation or memorization blocks, a two-second instruction screen showed two pictures of houses or faces that were related to two responses, and was followed by a delay phase in which a fixation cross was presented for a variable delay. Then, depending on the block type, participants either had to respond with the instructed response to one of the two stimuli (implementation block), or indicate whether displays matched or mismatched that of the instruction screen (memorization block). In a first analysis, a pattern classifier was trained to discriminate house and face

instructions in the instruction and delay phase. The crucial question was whether the pattern of brain activity differed between implementation and memory blocks. Interestingly, the pattern of brain activity was identical for the implementation and memory blocks in the instruction phase. This is not surprising given that the same stimuli were presented. In the delay phase, however, decoding was possible from more widespread areas in the implementation compared to the memory block. While in the implementation block the mappings could be decoded from frontoparietal and visual brain areas, decoding in the memory block was restricted to visual brain areas. The latter finding is consistent with recent observations from the working memory literature demonstrating that maintenance of information is restricted to posterior brain regions (Riggall and Postle, 2012). When the same information has to be implemented, however, frontoparietal brain regions come into play. In a second analysis, representational similarity analysis (RSA) was used to correlate patterns of brain activity across different conditions to study the similarity of these patterns (Kriegeskorte et al., 2008). The analysis revealed that the instruction phase and the delay phase were more strongly correlated for memory instructions compared to implementation instructions, suggesting that in the implementation condition an additional transformation took place between the instruction and delay phase. Overall, this study demonstrated that memory and implementation instructions can be dissociated on the neural level. Furthermore, it demonstrates that the frontoparietal network is more involved when participants have to implement new instructions.

4.4. Summary

While there are still some inconsistencies in the literature, the few existing fMRI studies provide relevant information about the implementation of new instructions. First,

building a task model for new instructions involves frontoparietal brain regions (Dumontheil et al., 2011; Ruge and Wolfensteller, 2010). Second, these frontoparietal brain regions are also involved in maintaining the instructions when the goal is to implement them (Muhle-Karbe et al., 2016). After a few applications of the new mappings, activation strongly decreases in these areas (Cole et al., 2010; Ruge and Wolfensteller, 2010). Importantly, implementation and memory instructions can be dissociated using univariate and multivariate analysis (Demanet et al., 2016; Muhle-Karbe et al., 2016). Finally, activation in the frontopariatal network during the instruction phase correlates with performance (Demanet et al., 2016; Ruge and Wolfensteller, 2010).

5. A cognitive model of implementing new instructions

After having reviewed behavioral and imaging findings on following new instructions, we will summarize these findings in a heuristic model. As mentioned in the introductory part of the current review, our model distinguishes different phases of instruction following (see Figure 4). The first phase relates to the construction of the task model where participants have to integrate and structure the declarative information conveyed by the task instructions. We propose that the instruction phase results into the formation of a procedural representation containing structures of different condition-action rules. The second phase refers to the implementation of the task model. Implementation is the process through which the whole task model or its relevant parts become highly accessible, such that the corresponding condition-action rules are ready to be applied. The implementation phase thus results into a high state of preparedness and offers the basis of prepared reflexes.

The last phase is the application phase where one specific condition-action rule is selected out of the set of highly accessible condition-action rules. Such selection is highly automatized and reflexive.

The three phases we proposed can be partly mapped on the different components of the working memory model of Oberauer (Oberauer, 2009; Oberauer, 2010). This model assumes that working memory consists of three layers of representations and dissociates between declarative and procedural working memory. The first layer is the activated long-term memory (ALTM), which supposedly contains procedural and declarative representations at a subthreshold level of activations. At the second layer, information is highly accessible. For declarative information, this layer is labelled the direct access region (DA). For procedural information, this layer is labelled the bridge. According to Oberauer (2009) "the bridge holds the currently operative task set, that is, the task set that is currently in control of thought and action." (p. 58). The third layer only contains the most accessible element. This constitutes the focus of attention (FOA) in declarative working memory and the response focus in procedural working memory. The hypothesis is that as information is passed on through these different layers, the higher they are activated and the more they are prone to capacity limitations. We propose that the three phases underlying instruction following are each bound to one of these three layers.

5.1. The instruction phase

In the instruction phase, linguistic information is translated into a task model. For simple tasks, the construction of the task model only involves compiling a number of verbally instructed S-R mappings into condition-action rules (Hartstra et al., 2012; Liefooghe et al., 2012; Ruge and Wolfensteller, 2010). For more complex tasks (Bhandari and Duncan,

2014; Cole et al., 2010; Dumontheil et al., 2011), the construction of the task model is more complicated. The different sets of relevant rules need to be structured into a format, which is suitable for the implementation phase. This involves creating a hierarchical structure and chunking of information. Furthermore, it involves the formation of condition-action rules. When the task model becomes too complex, specific elements will not be included and will be neglected in the following phases. The way the instructions are given determines whether an element is integrated into the task model or not (Duncan et al., 1996). Furthermore, even if participants are told that a specific aspect of the task model is not relevant in a given context, the complexity of the original task model still impacts on its implementation (Duncan et al., 2008). The construction of the task model activates a broad frontoparietal network, the so called multiple demand network (Dumontheil et al., 2011; Duncan, 2013) and is highly related to fluid intelligence.

This instruction phase results in a procedural task-model in which all necessary condition-action rules are represented. How exactly the transformation from the declarative instruction to a procedural task-model is achieved is an open question. Certainly, the structuring of information and the break down into smaller informational units plays a crucial role. Furthermore, mental simulation and motor imagery might be important. For instance, Ruge and Wolfensteller (2010) observed that stronger activation in the lateral premotor cortex and prefrontal cortex during the encoding of the novel S-R mappings, predicted enhanced performance improvement during the application of these mappings. This finding led to the suggestion that the implementation of novel S-R mappings may be completed by mentally simulating the overt application of these mappings. Finally, learned associations between specific words and motor programs might also play a role (Bundt et al., 2015). Depending on the complexity of the instructions, such a model can be relatively

simple (e.g., two condition-action rules) or very complex, consisting of different sets of condition-action rules and higher-order conditions, which indicate when a particular set is relevant. In principle, these procedural representations can be represented in ALTM. However, in order to result in overt behavior and instruction following to occur, the task model or its relevant parts need to become highly accessible. To this end, it needs to be represented in the next layer of procedural working memory: the bridge. This uploading occurs during the implementation phase.

5.2. The implementation phase

While the information represented in the task model is in a procedural format, another transformation step is needed in order to execute an instruction. De Jong et al. (1999) first put forward the idea that goal neglect can result from a failure to implement the task model. From this perspective, a problem can not only arise during construction of the task model but also during its implementation. Both behavioral research on IBC and brain imaging research support the importance of such an additional implementation step. Behavioral research suggests that only when the intention to implement the task model is maintained, instructions exert an automatic influence on behavior (Liefooghe et al., 2013). Furthermore, not all elements of the task model necessarily lead to IBC effects (Braem et al., 2016). Brain imaging research suggests that stimulus-response mappings that participants intend to implement are kept in a highly accessible state in parts of the MD network, presumably in the frontolateral cortex (Demanet et al., 2016). These findings indicate that the implementation and maintenance of the task model is an effortful process that is susceptible to strategies. For complex task models, where sequences of rules have to be implemented, not all elements that are included in the task model are transformed into such

an active state but only the elements that are needed for the next operation. For more simple tasks, all elements might be transformed into the implementation stage.

Within the framework of Oberauer (Oberauer, 2009; Oberauer, 2010), we propose that implementation consists of loading the task-model into the bridge. It is crucial to mention that for instruction following only the now relevant task model is crucial. Other task models that are not required in the current trial do not influence the implementation process directly. While these other task models seem to be also accessible to some degree, they are kept in a less activated state, in ALTM.

5.3. The application phase

The last phase of instruction following is the application phase. In this phase the relevant condition-action rule is selected out of the different task rules, which are represented in the bridge. In the WM terminology of Oberauer (2009), the condition-action rule that is the most relevant for the next cognitive operation is loaded into the response focus, which can only contain one response representation at any time, which coincides with the structural bottleneck to response selection (Pashler, 1994).

During this application phase, we are able to almost automatically respond to the relevant task stimulus, due to the "prepared reflex" mechanisms described above (Hommel, 2000; Meiran et al., 2012). Importantly, however, another transformation step occurs during this application phase. During, or swiftly following, the application of a response, a memory trace will be formed (e.g. in LTM) that is qualitatively different from how (or "where") it was encoded thus far. The idea that the initial application of newly instructed S-R mappings also leads to the formation of additional S-R associations, has been echoed in the computational model of Ramamoorthy and Verguts (2012). Their model supposes the presence of two

processing routes. The first route quickly learns novel S-R associations on the basis of instructions, but leads to slow responding. The second route slowly learns novel S-R associations, but elicits fast responses. Ramamoorthy and Verguts (2012) propose that the second route learns S-R associations on the basis of Hebbian learning, following the application of these S-R associations through the first route.

In addition, it is crucial to emphasize two aspects that are relevant for our understanding of instruction following. First, the imaging literature on instruction following shows that the brain networks that are involved in constructing the task model and implementing it become rapidly disengaged as soon as the instruction has been applied (Ruge and Wolfensteller, 2010). This indicates that specific neurocognitive mechanisms are involved in instruction following of new instructions. Second, for complex tasks the application phase can be used to consolidate the task model. In such cases the application phase acts as a kind of extension of the instruction phase (Bhandari and Duncan, 2014).

6. Open questions and future research

The aim of the current review was to investigate the dissociation of knowing and doing when following new instructions. We tried to integrate and structure the existing literature and formulate a heuristic model of instruction following. This model might help to integrate existing research domains in the area of instruction following. Furthermore, it might help to generate hypotheses about the factors that influence the different phases of instruction following. However, a number of questions are still open. One question relates to the specific nature of the task model. In the model we outlined above we assume that the task model has already a procedural format. However, only when elements of the task

model are implemented they exert an automatic influence on behavior. In other words, there seem to be different transformation steps from a pure declarative to a completely procedural representation. Furthermore, the specific cognitive mechanisms that mediate the transformation from a declarative to a procedural representation are still poorly understood. A second issue relates to potential capacity limits of the different phases of instruction following. The literature on goal neglect suggests that the complexity of the task model is strongly limited. However, the factors determining the capacity limit of the task model are not well understood. Furthermore, in the implementation phase an even stronger capacity limit seems to apply. The IBC literature suggests that at least two S-R mappings can be represented in the implementation phase. This might, however, depend on the characteristics of these S-R mappings. Another open question relates to the functional neuroanatomy of instruction following. There is strong evidence that the MD network is involved in instruction following (Cole et al., 2010; Dumontheil et al., 2011; Duncan, 2013; Ruge and Wolfensteller, 2010). However, it is still an open question whether different regions of this network become differentially involved in the different phases of instruction following. We suggest that the construction of the task model involves the MD network (Dumontheil et al., 2011). The implementation phase, however, is based only on parts of this network, presumably the frontolateral cortex (Demanet et al., 2016). In this context it is also crucial to briefly discuss other pathologies that might be related to instruction following. We have referred primarily to frontal brain damage and the dissociation of knowing and doing in frontal patients (Milner, 1963). However, there is an interesting literature on apraxia, in particular ideomotor apraxia, that might be relevant for our understanding of the functional neuroanatomy of instruction following as well (Wheaton and Hallett, 2007). Patients with ideomotor apraxia sometimes fail to implement instructions even though they have

comprehended the instructions and are also able to carry out the instructed behavior spontaneously (Heilman and Rothi, 2003). This seems to suggest that they have a problem in the transformation of instructions into motor programs. The literature on ideomotor apraxia is very complex and it would go beyond the scope of the present review to provide a detailed characterization. However, models of ideomotor apraxia assume that the failure to implement instructions can have different causes, including a failure to access motor representations, damage of such representations or a failure to translate motor representations into motor programs (Rothi et al., 1997). There are two differences between the dissociation of knowing and doing in frontal patients and ideomotor apraxia. First, ideomotor apraxia is usually tested by using direct verbal commands of trained actions such as 'wave goodbye', whereas the dissociation between knowing and doing usually refers to the application of new condition-action rules. Secondly, ideomotor apraxia has been mainly attributed to the left posterior parietal and premotor cortex, whereas the present dissociation between knowing and doing seems to be most related to frontal patients. While the dissociation of knowing and doing presumably relates to the first two stages of our heuristic model of instruction following, ideomotor apraxia presumably relates to malfunctions in the application stage.

Furthermore, we have outlined how different phases of instruction following map onto different working memory components. This mapping, however, should be only understood as an attempt to integrate the two literatures. At the present time, we mainly used the framework proposed by Oberauer. However, other candidate models could be considered. For instance, Vandierendonck (2016) recently conceptualized procedural working memory by extending the multi-component working memory model of Baddeley and Hitch (1974) with an additional executive memory module. Within this model, one could

assume that implementation and maintenance of instructions is a function of this additional module. Within the multi-component model of (Baddeley and Hitch, 1974) the question arises how the episodic buffer (Baddeley, 2000) plays a role in the creation of task model on the basis of verbal information. This is especially pertinent, because the episodic buffer is supposed to be involved in the integration of declarative information and language comprehension more generally (Baddeley et al., 2009).

Finally, in the computational neuroscience domain there are models that try to explain a capacity limit in cognitive control (Badre, 2008; Chatham et al., 2014). Chatham et al. (2014), for example, distinguish an input and an output gate in working memory. Input gating determines which information enters working memory. The output gating mechanism (Kriete and Noelle, 2011) determines which information influences behavior and thus corresponds to the implementation phase in our model. It becomes clear that understanding instruction following and the dissociation between knowing and doing, will require the further integration of different research domains in order to develop models that go beyond the heuristic framework we propose here.

Acknowledgement

This research was supported by a BELSPO Research Network funded under the ‘‘Interuniversity Poles of Attraction” Program, Grant P7/33, and the FWO project (G.0231.13). Marcel Brass was supported by Ghent University BOF-ZAP Grant and Jan De Houwer by a Ghent University Methusalem Grant (BOF09/01M00209). Senne Braem (12K6316N) was supported by FWO - Research Foundation Flanders.

References

Abrahamse, E., Braem, S., Notebaert, W., Verguts, T., 2016. Grounding cognitive control in associative learning. Psychol Bull 142, 693-728.

Baddeley, A., 2000. The episodic buffer: a new component of working memory? Trends Cogn Sci 4, 417-423.

Baddeley, A., Hitch, G., Allen, R., 2009. Working memory and binding in sentence recall. Journal of Memory and Language 61, 438-456.

Baddeley, A.D., Hitch, G., 1974. Working memory. Psychology of learning and motivation 8, 47-89.

Badre, D., 2008. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. Trends Cogn Sci 12, 193-200.

Bhandari, A., Duncan, J., 2014. Goal neglect and knowledge chunking in the construction of novel behaviour. Cognition 130, 11-30.

Braem, S., Liefooghe, B., De Houwer, J., Brass, M., Abrahamse, E.L., 2016. There Are Limits to the Effects of Task Instructions: Making the Automatic Effects of Task Instructions Context-Specific Takes Practice. J Exp Psychol Learn Mem Cogn.

Brass, M., Wenke, D., Spengler, S., Waszak, F., 2009. Neural correlates of overcoming interference from instructed and implemented stimulus-response associations. The Journal of neuroscience : the official journal of the Society for Neuroscience 29, 1766-1772.

Bundt, C., Bardi, L., Abrahamse, E.L., Brass, M., Notebaert, W., 2015. It wasn't me! Motor activation from irrelevant spatial information in the absence of a response. Frontiers in human neuroscience 9.

Chatham, C.H., Frank, M.J., Badre, D., 2014. Corticostriatal output gating during selection from working memory. Neuron 81, 930-942.

Cohen-Kdoshay, O., Meiran, N., 2007. The representation of instructions in working memory leads to autonomous response activation: evidence from the first trials in the flanker paradigm. Q J Exp Psychol (Hove) 60, 1140-1154.

Cohen-Kdoshay, O., Meiran, N., 2009. The representation of instructions operates like a prepared reflex: flanker compatibility effects found in first trial following S-R instructions. Exp Psychol 56, 128-133.

Cole, M.W., Bagic, A., Kass, R., Schneider, W., 2010. Prefrontal dynamics underlying rapid instructed task learning reverse with practice. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 14245-14254.

Cole, M.W., Etzel, J.A., Zacks, J.M., Schneider, W., Braver, T.S., 2011. Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. Frontiers in human neuroscience 5, 142.

De Houwer, J., Beckers, T., Vandorpe, S., Custers, R., 2005. Further evidence for the role of mode-independent short-term associations in spatial Simon effects. Perception & Psychophysics 67, 659-666.

De Jong, R., Berendsen, E., Cools, R., 1999. Goal neglect and inhibitory limitations: dissociable causes of interference effects in conflict situations. Acta psychologica 101, 379-394.

Deacon, T.W., 1997. The symbolic species: The co-evolution of the brain and language. New York: WW Norton&Co.

Demanet, J., Liefooghe, B., Hartstra, E., Wenke, D., De Houwer, J., Brass, M., 2016. There is more into 'doing' than 'knowing': The function of the right inferior frontal sulcus is specific for implementing versus memorising verbal instructions. Neuroimage 141, 350-356.

Diamond, A., 1991. Frontal lobe involvement in cognitive changes during the first year of life. Brain maturation and cognitive development: Comparative and cross-cultural perspectives, 127-180.

Dumontheil, I., Thompson, R., Duncan, J., 2011. Assembly and use of new task rules in fronto-parietal cortex. Journal of Cognitive Neuroscience 23, 168-182.

Duncan, J., 2013. The structure of cognition: attentional episodes in mind and brain. Neuron 80, 35-50.

Duncan, J., Burgess, P., Emslie, H., 1995. Fluid intelligence after frontal lobe lesions. Neuropsychologia 33, 261-268.

Duncan, J., Emslie, H., Williams, P., Johnson, R., Freer, C., 1996. Intelligence and the frontal lobe: the organization of goal-directed behavior. Cogn Psychol 30, 257-303.

Duncan, J., Parr, A., Woolgar, A., Thompson, R., Bright, P., Cox, S., Bishop, S., Nimmo-Smith, I., 2008. Goal neglect and Spearman's g: competing parts of a complex task. J Exp Psychol Gen 137, 131-148.

Eimer, M., 1995. Stimulus-response compatibility and automatic response activation: evidence from psychophysiological studies. Journal of Experimental Psychology: Human Perception and Performance 21, 837.

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

Everaert, T., Theeuwes, M., Liefooghe, B., De Houwer, J., 2014. Automatic motor activation by mere instruction. Cogn Affect Behav Neurosci 14, 1300-1309.

Exner, S., 1879. Physiologie der Grosshirnrinde. Handbuch der physiologie 2, 189-350.

Hartstra, E., Kuhn, S., Verguts, T., Brass, M., 2011. The implementation of verbal instructions: an fMRI study. Human brain mapping 32, 1811-1824.

Hartstra, E., Waszak, F., Brass, M., 2012. The implementation of verbal instructions: dissociating motor preparation from the formation of stimulus-response associations. Neuroimage 63, 1143-1153.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425-2430.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat Rev Neurosci 7, 523-534.

Heilman, K., Rothi, L., 2003. Apraxia, in: KM, H., E, V. (Eds.), Clincial Neuropsychology. Oxford University Press, New York, pp. 215–235.

Hommel, B., 2000. The prepared reflex: Automaticity and control in stimulus-response translation., in: Monsell, S., Driver, J. (Eds.), Control of cognitive processes: Attention and performance XVIII. MIT Press, Cambridge, MA, pp. 247-273.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2, 4.

Kriete, T., Noelle, D.C., 2011. Generalisation benefits of output gating in a model of prefrontal cortex. Connection Science 23, 119-129.

Liefooghe, B., De Houwer, J., subm. Automatic effects of instructions do not necessarily reflect the implementation of an action plan.

Liefooghe, B., De Houwer, J., Wenke, D., 2013. Instruction-based response activation depends on task preparation. Psychon Bull Rev 20, 481-487.

Liefooghe, B., Wenke, D., De Houwer, J., 2012. Instruction-based task-rule congruency effects. J Exp Psychol Learn Mem Cogn 38, 1325-1335.

Luria, A.R., 1980. Higher order functions in man. Consultants Bureau, New York.

Marcovitch, S., Boseovski, J.J., Knapp, R.J., Kane, M.J., 2010. Goal neglect and working memory capacity in 4- to 6-year-old children. Child Dev 81, 1687-1695.

McVay, J.C., Kane, M.J., 2009. Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. J Exp Psychol Learn Mem Cogn 35, 196-204.

Meiran, N., Cohen-Kdoshay, O., 2012. Working memory load but not multitasking eliminates the prepared reflex: further evidence from the adapted flanker paradigm. Acta psychologica 139, 309-313.

Meiran, N., Cole, M.W., Braver, T.S., 2012. When planning results in loss of control: intention-based reflexivity and working-memory. Frontiers in human neuroscience 6, 104.

Meiran, N., Pereg, M., Kessler, Y., Cole, M.W., Braver, T.S., 2015a. The power of instructions: Proactive configuration of stimulus-response translation. J Exp Psychol Learn Mem Cogn 41, 768-786.

Meiran, N., Pereg, M., Kessler, Y., Cole, M.W., Braver, T.S., 2015b. Reflexive activation of newly instructed stimulus-response rules: evidence from lateralized readiness potentials in no-go trials. Cogn Affect Behav Neurosci 15, 365-373.

Milner, B., 1963. Effects of different brain lesions on card sorting. . Archives of Neurology 9, 90 -100.

Muhle-Karbe, P.S., Duncan, J., De Baene, W., Mitchell, D.J., Brass, M., 2016. Neural Coding for Instruction-Based Task Sets in Human Frontoparietal and Visual Cortex. Cereb Cortex.

Nakahara, K., Hayashi, T., Konishi, S., Miyashita, Y., 2002. Functional MRI of macaque monkeys performing a cognitive set-shifting task. Science 295, 1532-1536.

Oberauer, K., 2009. Design for a working memory. Psychology of learning and motivation 51, 45-100.

Oberauer, K., 2010. Declarative and Procedural Working Memory: Common Principles, Common Capacity Limits? Psychol Belg 50, 277-308.

Pashler, H., 1994. Dual-task interference in simple tasks: data and theory. Psychol Bull 116, 220-244.

Ramamoorthy, A., Verguts, T., 2012. Word and deed: a computational model of instruction following. Brain Res 1439, 54-65.

Ridderinkhof, K.R., Ullsperger, M., Crone, E.A., Nieuwenhuis, S., 2004. The role of the medial frontal cortex in cognitive control. Science 306, 443-447.

Riggall, A.C., Postle, B.R., 2012. The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. The Journal of neuroscience 32, 12990-12998.

Roberts, G., Anderson, M., 2014. Task structure complexity and goal neglect in typically developing children. J Exp Child Psychol 120, 59-72.

Roepstorff, A., Frith, C., 2004. What's at the top in the top-down control of action? Script-sharing and 'top-top' control of action in cognitive experiments. Psychol Res 68, 189-198.

Rogers, R.D., Monsell, S., 1995. Costs of a Predictable Switch between Simple Cognitive Tasks. Journal of Experimental Psychology-General 124, 207-231.

Rothi, L.j.G., Ochipa, C., Heilman, K.M., 1997. A cognitive neuropsychological model of limp praxis and apraxia, in: Rothi, L.J.G., Heilman, K.M. (Eds.), Apraxia: The Neuropsychology of Action, Psychology Press, East Sussex.

Ruge, H., Wolfensteller, U., 2010. Rapid formation of pragmatic rule representations in the human brain during instruction-based learning. Cereb Cortex 20, 1656-1667.

Steinhauser, M., Hubner, R., 2007. Automatic activation of task-related representations in task shifting. Mem Cognit 35, 138-155.

Teuber, H.L., 1964. Discussion, in: Warren, J.M., Akert, K. (Eds.), The frontal granular cortex and behavior. McGraw-Hill, New York, p. 333.

Towse, J.N., Lewis, C., Knowles, M., 2007. When knowledge is not enough: the phenomenon of goal neglect in preschool children. J Exp Child Psychol 96, 320-332.

Vandierendonck, A., 2016. A Working Memory System With Distributed Executive Control. Perspect Psychol Sci 11, 74-100.

Waszak, F., Wenke, D., Brass, M., 2008. Cross-talk of instructed and applied arbitrary visuomotor mappings. Acta psychologica 127, 30-35.

Wenke, D., De Houwer, J., De Winne, J., Liefooghe, B., 2015. Learning through instructions vs. learning through practice: flanker congruency effects from instructed and applied S-R mappings. Psychol Res 79, 899-912.

Wheaton, L.A., Hallett, M., 2007. Ideomotor apraxia: a review. Journal of the neurological sciences 260, 1-10.

Whiten, A., Goodall, J., McGrew, W.C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C.E., Wrangham, R.W., Boesch, C., 1999. Cultures in chimpanzees. Nature 399, 682-685.

Woodworth, R., 1938. Experimental psychology. Holt, Rinehart and Winston, New York.

Zelazo, P.D., 2004. The development of conscious control in childhood. Trends Cogn Sci 8, 12-17.

Zelazo, P.D., Frye, D., Rapus, T., 1996. An age-related dissociation between knowing rules and using them. Cognitive Dev 11, 37-63.

Zelazo, P.D., Reznick, J.S., 1991. Age-Related Asynchrony of Knowledge and Action. Child Dev 62, 719-735.

Figure captions

Figure 1. Goal neglect paradigm introduced by Duncan et al. (1996). Participants were instructed to attend to the stimuli on the side that was indicated by the first side cue (e.g. Watch Right). They should read the letters and ignore numbers on this side. Before the last three trials, another side cue was presented. A '+' indicated to attend to the right side and a '-' indicated to attend to the left side.

Figure 2. Schematic outline of the procedure used by Liefooghe et al. (2012). On each run of trials, new S-R mappings were first instructed. These instructions were followed by a variable number of diagnostic trials. Each run ended, with the presentation of a probe stimulus of the inducer task. Both tasks shared the same responses and stimuli.

Figure 3. Experimental procedure and brain imaging results from the study by Demanet et al. (2016). a) In the instruction phase, the *implementation group* should prepare to execute the instructed S-R mappings when a go signal occurred (66 % of the trials) and could stop preparing when a no-go signal occurred (33 % of the trials). In the probe phase, one of the stimuli of the instruction phase was presented and participants had to respond to the stimulus with the instructed response. The *memorization group* had to memorize the instructions when a go signal occurred and could stop memorizing the instructions when a no-go signal occurred. In the probe phase they had to indicate whether the presented stimuli matched the stimuli of the instruction phase. b) Brain activation for the interaction of group (Implementation vs. memorization) and signal (go versus no-go) in the instruction phase.

Figure 4. Heuristic model of instruction following. The model consists of three phases. Whereas the degree of proceduralization increases in each phase, the working memory capacity decreases.

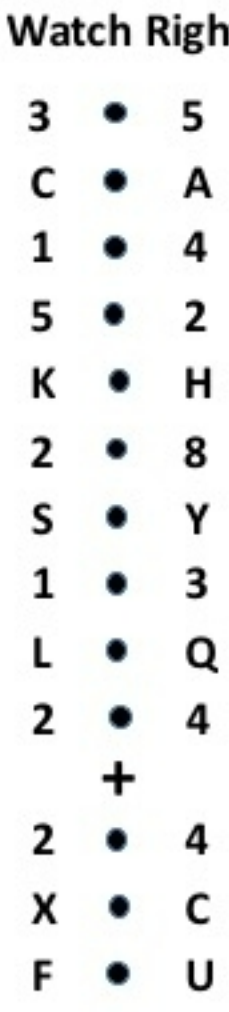

Figure 1

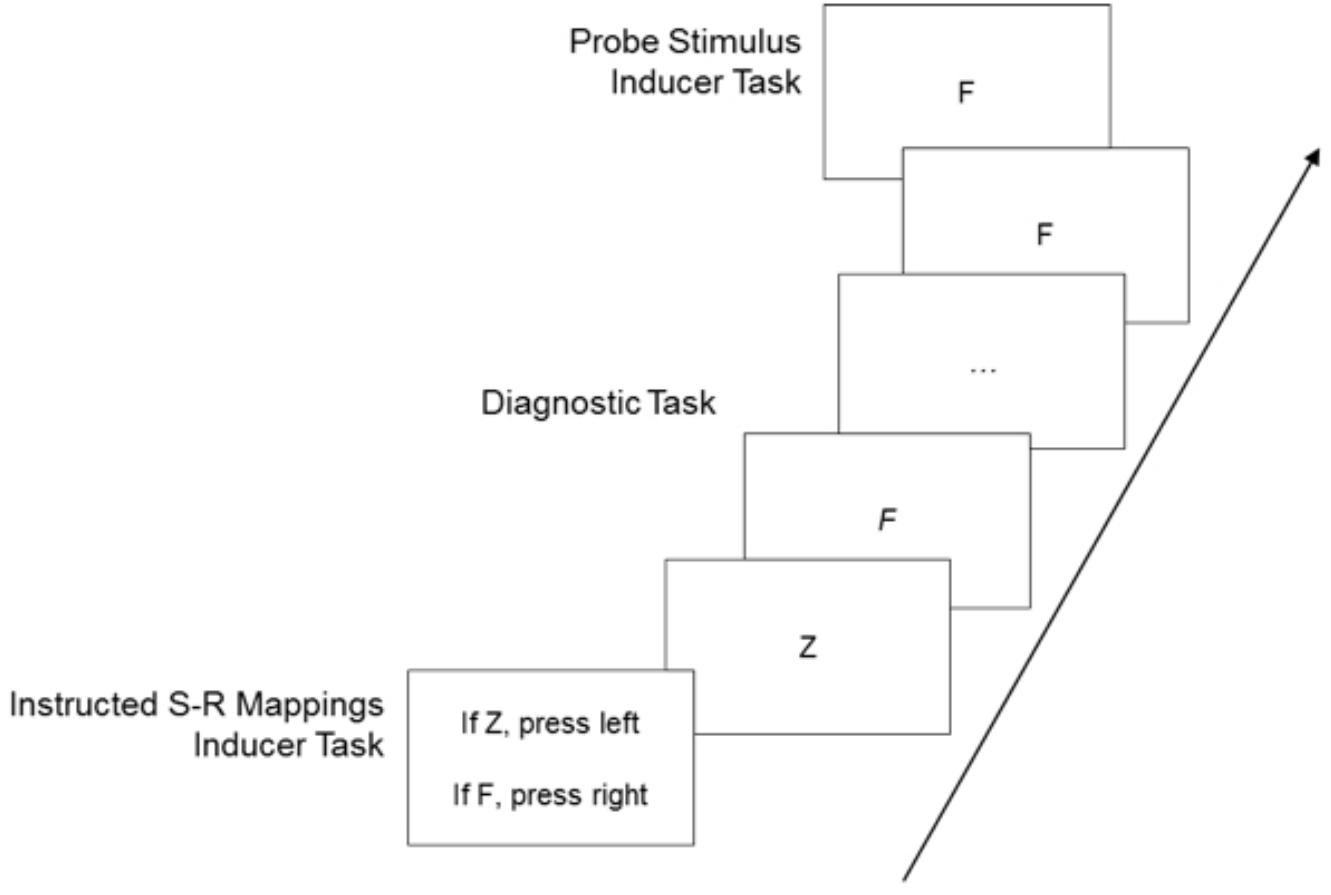
Probe Stimulus
Inducer Task
F
F
...
Diagnostic Task
F
Z
Instructed S-R Mappings
Inducer Task
If Z, press left
If F, press right

Figure 2

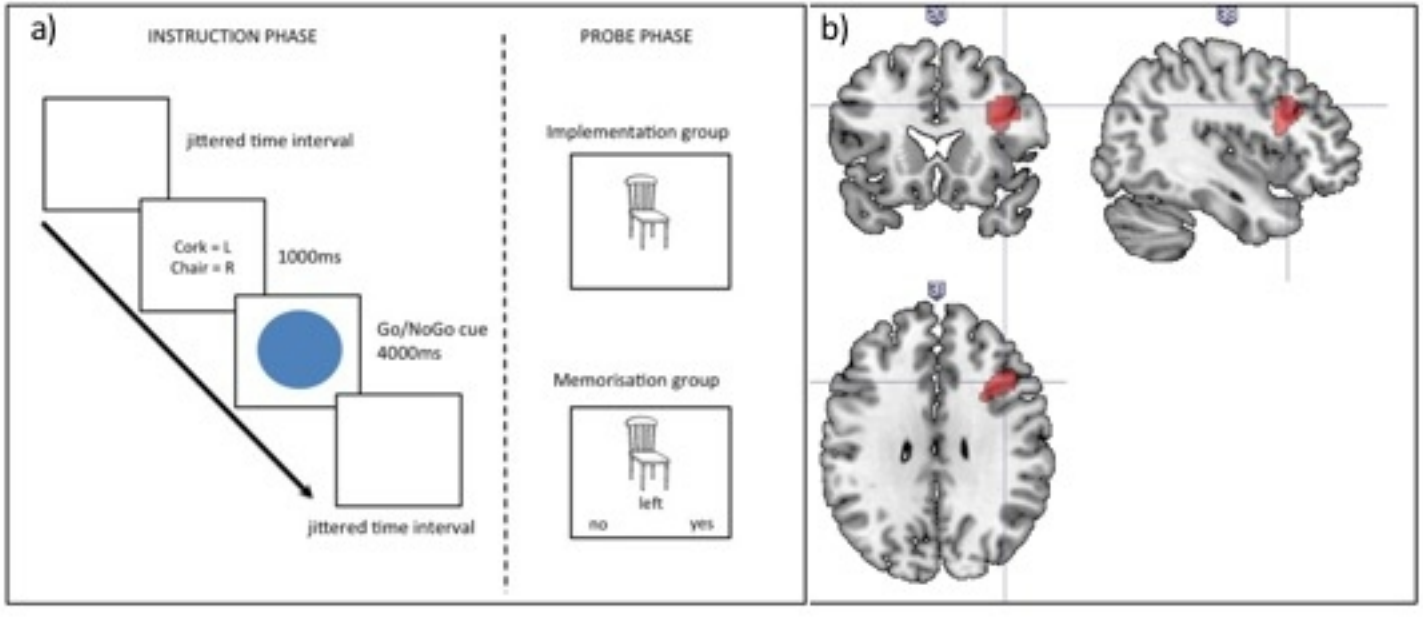
a)
INSTRUCTION PHASE
PROBE PHASE
jittered time interval
Cork = L
Chair = R
1000ms
Go/NoGo cue
4000ms
jittered time interval
Implementation group
Memorisation group
left
no
yes
b)

Figure 3

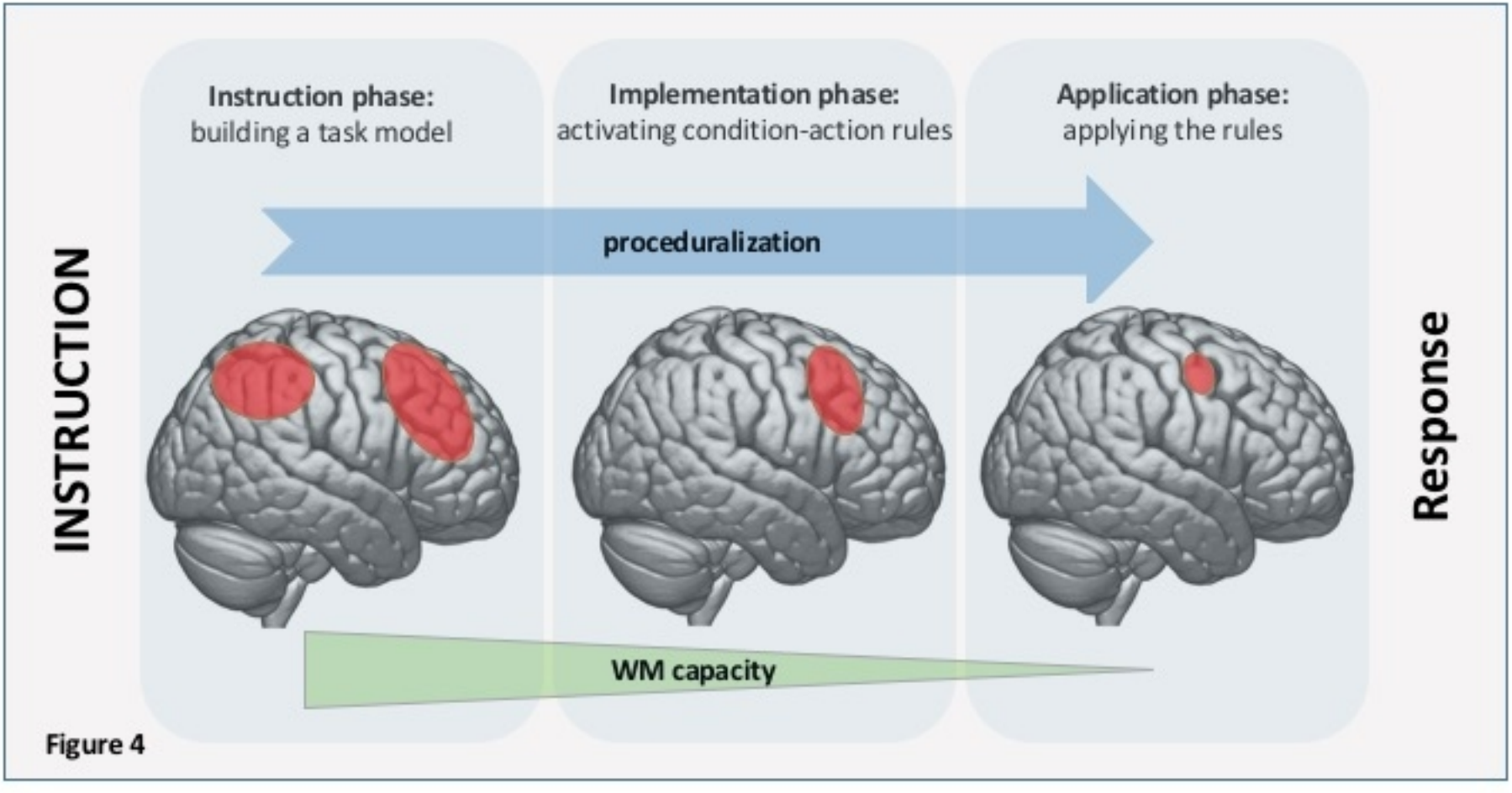
INSTRUCTION
Instruction phase:
building a task model
Implementation phase:
activating condition-action rules
Application phase:
applying the rules
proceduralization
WM capacity
Response

Figure 4

Highlights

- Reviews cognitive neuroscience research on instruction following
- Evaluates the evidence for a dissociation of knowing and doing
- Proposes a heuristic model of following new instructions
- Determines the role of the frontoparietal network in instruction following
- Integrates cognitive control and working memory research on instruction following