In Press. Journal of Experimental Psychology: General

Tenacious instructions: How to dismantle newly instructed task rules?

Elger Abrahamse[1,2], Senne Braem[3], Jan De Houwer[4] & Baptist Liefooghe[5]

[1] Department of Communication and Cognition, Tilburg University, The Netherlands

[2] Department of Educational Sciences, Atlántico Medio University, Spain

[3] Department of Experimental Psychology, Ghent University, Belgium

[4] Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

[5] Department of Psychology, Utrecht University, Utrecht, The Netherlands

Author note

Elger Abrahamse and Baptist Liefooghe contributed equally to this manuscript. This research was supported by grant BOF16/MET_V/002 of Ghent University to J.D.H., and by grant G00951N of the Flemish Government to B.L. and J.D.H. S.B. was supported by an ERC Starting grant (European Union's Horizon 2020 research and innovation program, Grant agreement 852570). Raw data and corresponding processing scripts are available at https://osf.io/au7pv/. Correspondence concerning this article should be addressed to Baptist Liefooghe, b.liefooghe@uu.nl. The ideas and data appearing in the current manuscript have not been disseminated elsewhere before.

Abstract

Humans excel in instruction following to boost performance in unfamiliar situations. We can do so through so-called prepared reflexes: Abstract instructions are instantly translated into appropriate task rules in procedural working memory, after which imperative stimuli directly trigger their corresponding responses in a ballistic, reflex-like manner. But how much control do we have over these instructed task rules when their reflexes suddenly lose their relevance? Inspired by the phenomenon of directed forgetting in declarative working memory, we here tested across four experiments if the presentation of (implicit or explicit) task cancellation cues results in the directed dismantling of recently instructed task rules. Our findings suggest that – even when cancelation cues are actively processed – such dismantling does not occur (Experiment 1-3) unless the no-longer relevant task rules are replaced by a new set of rules (Experiment 4). These findings and their implications are discussed in the broader context of action control and working memory.

Keywords: Instructions; Action Control; Cognitive Control; Working Memory; Prepared Reflex; Flexibility; Instruction Following; Directed Forgetting

Tenacious instructions: How to dismantle newly instructed task rules?

In every-day life, goal-directed behavior requires constant adaptation to new challenges and demands. For most species, discovering how to act best in an unfamiliar context is a laborious process that involves costly trial-and-error learning of the appropriate set of stimulus-response rules. Fortunately, due to extraordinary language abilities, humans can boost learning and performance through explicit instruction (e.g., Rastle et al., 2021) and circumvent the trial-and-error process. For example, a driving instructor can instruct the student driver how to maneuver the car, preparing the student's mind for the most appropriate procedures to be executed in traffic. Hence, instruction following is a core mechanism in goal-directed action control.

Behavior is governed by a multitude of instructions presented at different moments in time, and no-longer-relevant instructions need to be quickly updated by relevant ones. The student driver may need to quickly forget old instructions, and adapt to new ones, when an oncoming car unexpectedly changes the traffic situation. Previous research has shown that, once prepared for, instructed stimulus-response rules are applied and executed in a rather effortless and reflex-like manner (for reviews, see Brass et al., 2017; Meiran et al., 2012). Yet, little is known about the level of adaptive control humans hold over already prepared but not yet executed instructions. When the context requires us to do so, can we 'dismantle' new stimulus-response instructions as easily as we can apply them?

Instructions at the origin of human action control

Action control is an umbrella term for adaptive processes that serve goal-directed behavior, monitoring and adjusting the information-processing stream between environmental input and behavioral output in order to reach a goal. Such control is highly versatile and encompasses various more dedicated functions (cf. Miyake et al., 2000), such as task switching (e.g., Koch et al., 2018), response inhibition (e.g., Verbruggen & Logan, 2008), and conflict adaptation

(e.g., Botvinick et al., 2001). A general challenge is to understand action control as a self-regulating system that does not postulate ill-defined sets of homunculi (Monsell & Driver, 2000). Several proposals to face this challenge have been made, conceptualizing action control as the interaction between computationally well-defined basic mechanisms (e.g., Verbruggen, McLaren, & Chambers, 2014) and/or as the product of basic learning mechanisms (e.g., Abrahamse et al., 2016; Botvinick et al., 2001; Egner, 2014; Frings et al., 2020; Schmidt et al., 2016). These accounts generally highlight the importance of instructions in configuring the control system when a new task needs to be performed; yet, the mechanisms through which this is achieved, remain underspecified.

Interest into the mechanisms underlying instruction following increased substantially over recent years (for reviews, see Brass et al., 2017; Cole et al., 2017; Meiran et al., 2012; Meiran et al., 2017). Most of this research has been inspired by the observation that instructions – once prepared for – automatically impact behavior (e.g., Cohen-Kdoshay & Meiran, 2007; De Houwer et al., 2005; Liefooghe et al., 2012; Meiran et al., 2015a; Wenke et al., 2007; Whitehead & Egner, 2018a,b). One task that demonstrates such automaticity in instruction following is the inducer-diagnostic paradigm (e.g., Braem et al., 2017; Everaert et al., 2014; Liefooghe et al., 2012, 2013, 2016; Liefooghe & De Houwer, 2018; Theeuwes et al., 2014, 2015). The paradigm is outlined in Figure 1 (please note that Figure 1 displays an example of the *standard* paradigm, while for the current study we adjusted it on several aspects that are elaborated on in the below Method sections). It demonstrates an automatic impact of yet unexecuted stimulus-response instructions (prepared for an upcoming 'inducer task') on a currently ongoing 'diagnostic task' via overlapping response options between the two tasks. Specifically, performance on congruent diagnostic trials (in which the stimulus-response rules of both tasks indicate the same response) is typically better than performance on incongruent trials (for reviews, see Brass et al., 2017; Cole et al., 2017; Meiran, et al. 2017). This robust effect is referred to as the Instruction-Based Congruency (IBC) effect.

INSERT FIGURE 1 AROUND HERE

The IBC effect has been shown to vary as a function of instruction preparation (e.g., Braem et al., 2019; Liefooghe et al., 2012; Liefooghe et al., 2013; Meiran et al., 2015b; Wenke et al., 2009; Whitehead & Egner, 2018b). Such preparation likely involves the formation and active maintenance in working memory of action-oriented, procedural representations that enable stimulus-response reflexes (e.g., Brass et al., 2017; Liefooghe et al., 2012; Meiran & Cohen-Kdoshay, 2012): Once an imperative stimulus (feature) is encountered, the instructed corresponding response is automatically triggered. This aligns with Exner's (1879) classic notion of the "prepared reflex" – a core feature of (voluntary) action control more generally (e.g., Hommel, 2009). The notion that instruction following is (in part) completed through prepared reflexes, provides for a mechanistic, homunculus-free incorporation of instruction following into the control chain that is targeted in aforementioned general action control accounts (Abrahamse et al., 2016; Egner, 2014; Frings et al., 2020; Schmidt et al., 2016; Verbruggen et al., 2014). Namely, determining the appropriate instructed response at each moment evolves directly from specific probes in the stimulus context, and does not require an unspecified intelligent controller (cf. general-purpose systems; e.g. Baddeley, 1998; Norman & Shallice, 1986).

Placing instruction following in the broader context of action control, raises the question about the level of control that the cognitive system maintains over newly encoded instructions. Prepared reflexes may be a core mechanism through which people are enabled to instantly use new instructions to steer ongoing behavior – which is of great benefit in terms of flexible adaptation to new and/or changing environments. However, in the absence of any additional control mechanism, prepared reflexes may lead to rigid and erroneous instruction

following once these instructions become outdated and thus no longer relevant. As will be discussed below, the instructed removal of no longer relevant information has been explored extensively for *declarative* representations (cf. 'directed forgetting'). However, much less is known about the 'directed dismantling' of outdated *procedural* representations in working memory – even though this ability seems just as critical for goal-directed behavior.

Directed dismantling in working memory

Working memory ultimately serves goal-directed *action* (e.g., Manohar et al., 2019; Nobre & Stokes, 2019; Wolff et al., 2017). Hence, besides declarative content (underlying for example the typical recall and recognition tasks), working memory involves a procedural system for representations that control processing (including stimulus-response rules, and executive control settings; Oberauer, 2009; Oberauer, 2010; Oberauer et al., 2013). Instruction following requires a transfer from declarative to procedural working memory (Brass et al., 2017): When instructed about how to perform an upcoming task, a successful participant is required to translate the declarative input into appropriate procedural representations. Without such translation, instructions may still be retained (i.e., in a declarative format) but cannot be appropriately executed – as observed for example in frontal lesion patients (Milner, 1963). The IBC effect provides a window into this still poorly understood interplay between declarative and procedural working memory, which is increasingly recognized as a critical aspect of goal-directed action (e.g., Panichello & Buschman, 2021).

Whether declarative and procedural working memory representations actually share dedicated resources and processes (Barouillet et al., 2015), or are independent systems that work according to common principles (Oberauer, 2009; 2010) – it is likely that similar control mechanisms are at play between them. Specifically, in order for declarative and procedural working memory to function properly in line with their limited capacities for novel information (e.g., Oberauer, 2010), both require its content to be restricted to the most

relevant information for the task at hand. As such, with an eye on the constantly changing tasks and their demands that working memory is faced with, proper working memory function not only requires selection of the most relevant information, but also the ability to rapidly update its content when demands change.

Updating in declarative working memory has been proposed to involve an active removal mechanism (Lewis-Peacock et al., 2018). This mechanism alters the representational status of declarative working memory content when it is no longer relevant in order to reduce its access (e.g., via synaptic weight changes; Lewis-Peacock et al., 2018). Such active removal is increasingly supported and understood through studies demonstrating the impact of *cues* that indicate the participant to either keep remembering or forget information that is in working memory (e.g., Dames & Oberauer, 2021; Ecker et al., 2014a; Ecker et al., 2014b; Festini & Reuter-Lorenz, 2013; Festini & Reuter-Lorenz, 2014; Wolff et al., 2017). For example, Festini and Reuter-Lorenz (2013) demonstrated that directed forgetting in declarative working memory (via 'forget cues') reduces semantic interference, a well-known effect that occurs when lists of semantically related words are studied. Importantly, directed forgetting in declarative memory can be selective. Various studies have shown that specific items from a studied list (e.g., sentences in the list about one character called 'Tom') – but not other items (e.g., sentences in the list about another character called 'Alex') – can be selectively affected by forget cues (e.g., Aguirre et al., 2017; Delaney et al., 2009). This fits the notion of active and directed removal from declarative working memory.

Whereas directed forgetting has been extensively investigated for declarative information, little is known about the ability to remove information from procedural working memory when no longer relevant. In view of its importance for goal-directed behavior, we here target 'directed dismantling' of newly instructed stimulus-response rules.

The present study

Previous work in the domain of instruction following indicates that stimulus-response representations can be quickly configured in procedural working memory without much effort, as they instantly generate so-called prepared reflexes after being instructed (cf. the IBC effect; Brass et al., 2017). But can such newly installed representations be actively and efficiently removed from capacity-limited procedural working memory upon demand as well? In a series of four experiments, we here tested the impact of (implicit or explicit) 'cancellation cues' that followed the presentation of newly instructed stimulus-response rules. We adapted the inducer-diagnostic paradigm (see Figure 1) by adding a second series of trials of the diagnostic task after the completion (i.e., implicit cancellation; Experiment 1), explicit cancellation (Experiments 2 and 3), or replacement (Experiment 4) of the initial inducer task rules. The main question throughout this series of experiments, then, was whether the stimuli from the – now no longer relevant – inducer task instructions would still trigger an IBC effect in this second series of trials of the diagnostic task.[1] Our rationale was that if the cognitive system can dismantle newly formed procedural task representations as efficiently as it can create them, then the IBC effect should be eliminated or attenuated after instructions are canceled and no longer relevant.[2]

Experiment 1

As mentioned above, participants were presented with an adapted variant of the inducer-diagnostic paradigm. Participants were instructed that on the diagnostic task trials, they had to decide whether a stimulus was presented in italic or upright (e.g., upright, press left; italic, press right). These diagnostic task rules remained the same across all runs throughout the

[1] Hence, the congruency manipulation underlying the IBC effect in the second series of diagnostic trials was determined no longer in reference to an upcoming inducer task (cf. Figure 1), but in reference to a – now no longer relevant – inducer task from earlier in the 'run'.

[2] In the General Discussion we discuss how such dismantling is different from more general task switching, and how these two domains may nevertheless inform each other.

experiment (and used across all experiments reported in the current paper). An outline of the different run types for Experiment 1 is presented in Figure 2. Each run started with the instruction of two new stimulus-response mappings of the inducer task (e.g., if 'doll', press right; if 'car', press left). Following the encoding of these instructions, the first diagnostic task was presented. After the last trial of this first diagnostic task, the inducer task was performed. Finally, the second diagnostic task was presented.

On Filler Runs (Figure 2) both inducer task stimulus-response mappings were executed in between the first and second diagnostic task – as both the inducer task probes (i.e., stimuli) were presented once. Critically, in the runs of interest (Target Runs in Figure 2) only one of the two instructed stimuli of the inducer task was probed (either once or twice) and had to be responded to (e.g., the word 'doll' printed in green) – leaving the other instructed stimulus (e.g., the word 'car') not to be responded to. In the second diagnostic task, which immediately followed the probe of the inducer task, a distinction for Target Runs could thus be made between stimuli in the diagnostic task that were executed in the context of the inducer task (applied inducer stimulus; e.g., "doll"), and stimuli that were never probed in the inducer task context and thus never responded to (unapplied inducer stimulus; e.g., "car"). The unapplied inducer stimuli were thus part of an instructed stimulus-response mapping of the inducer task (if 'car', press left), that was prepared for, that was never executed, and that was no longer relevant. We tested whether or not an IBC effect still emerged for especially the unapplied inducer stimuli in the second diagnostic task (i.e., thus when the inducer task was completed and no longer relevant).

INSERT FIGURE 2 AROUND HERE

Method

Participants. Thirty-eight students from Ghent University participated in return for 10 euro. Participants were naive to the purpose of the experiment. Our experiments built upon the basic procedure for measuring IBC effects developed Liefooghe et al. (2012) in which the responses and stimuli of the inducer and diagnostic task overlap (see Figure 1). Two earlier experiments in which this procedure was used, showed medium-sized IBC effects with Cohen's d's of .55 (Experiment 1; Liefooghe et al., 2012) and .58 (Experiment 2; Liefooghe et al., 2012). As such, we determined a desired sample size of n = 32, aiming for a power of .80 to detect a medium-sized effect (i.e., d=.5, Cohen, 1962, 1988). Across the current study, only Experiment 2 below fell short of this desired sample size. Experiments in this study were not pre-registered. Ethical approval was obtained from the ethical committee of Ghent University for the overall research project that all Experiments 1-4 reported in the current study are part of.

Tasks & Materials. For each participant, 48 pairs of inducer task stimulus-response mappings were randomly constructed on the basis of 96 Dutch nouns selected from the SUBTLEX-NL database (Keuleers et al., 2010). Each pair of mappings was used for the inducer task only once (i.e., each for a single run), and consisted of one stimulus assigned to a left response, and one stimulus to a right response. During the inducer task trials either both stimulus-response mappings (i.e., both left and right response assignments) had to be actually executed (16 Filler Runs), or only one of the two inducer task stimulus-response mappings was actually executed (32 Target Runs: 16 Target Runs with 1 probe of the inducer task, and 16 Target Runs with twice the same probe of the inducer task). Target Runs were the main focus of interest (i.e., containing both an applied and unapplied inducer stimulus), while filler runs were included so that participants would infer that both inducer responses could actually be probed within a run.

In the diagnostic task trials, participants always judged whether a stimulus was printed upright or in italic. The left-right response assignment of the diagnostic task was counterbalanced across participants. The stimulus-response mapping instructions of the inducer task were followed by 0, 4, 8, or 12 trials of the first diagnostic task, and this number of trials was balanced across the Filler Runs and Target Runs. In line with previous studies, the number of trials of the first diagnostic task on each run was unpredictable for the participant. Accordingly, the onset of the probe of the inducer task was unpredictable, intended to encourage participants to keep prepared for the stimulus-response mappings of the inducer task (e.g., Liefooghe et al., 2012, 2013; Meiran et al., 2015a, 2015b). Following the probes of the inducer task, participants were always presented with the second diagnostic task, which included 20 trials. Similar to the first diagnostic task, congruency for the second diagnostic task was determined on the basis of having either a match (i.e., congruent) or mismatch (i.e., incongruent) between the specific responses signaled by the stimulus-response instructions of the inducer task versus those of the diagnostic task.

Instructed stimulus-response mappings and stimuli were presented in ARIAL font, size 16. Stimuli in the diagnostic task were presented in black on a white background, which was also true for the instructed mappings of the inducer task. Stimuli in the inducer task were presented in green, with the color thus signaling that the inducer task had to be executed. The A-key (left) and the P-key (right) of an AZERTY keyboard were used.

Procedure. Participants were tested in groups of two or three. Each participant was tested in a separate cubicle in which (s)he was placed in front of a 17-inch laptop with an AZERTY keyboard attached to it. The experiment was programmed by using the Tscope library for C/C++ (Stevens et al., 2006). At the start of the experiment, the overall instructions (including the diagnostic task stimulus-response mappings) were presented and paraphrased if necessary. The instructions did not explicitly state that the inducer task would not return

anymore after having responded to the inducer task probe(s) in between the first and second diagnostic tasks, but this was systematically the case and could easily be learned during the practice block (but see below). The experiment started with a practice block of 6 practice runs, followed by 4 experimental blocks, with a small break after each block. Each experimental block consisted of 12 runs: 1 Filler Run and 2 Target Runs for each of the four above stated trial-number conditions (0, 4, 8, or 12 trials) of the first diagnostic task. The practice block contained 12 runs in which the length of the first diagnostic task was determined randomly. Stimuli in this practice block were given names (e.g., Tom, James) rather than the nouns used for the experimental blocks, such that practice block stimuli did not overlap with experimental blocks.

Each run started with the presentation of the two stimulus-response mappings of the inducer task (Figure 2). The position on the screen of these stimulus-response mappings was determined randomly, so that the instructions for each response could be presented either above or below the screen center. Instructions remained on screen for a maximum of 20 seconds or until participants pressed the spacebar. For both the diagnostic task and the inducer task trials, the maximum response time was 2000ms, while the inter-trial interval was set to 500ms. Following incorrect or late responses, the screen flashed red for 100ms to denote an error. The different types of runs were presented in a random order within each block. The experiment lasted for approximately one hour.

Results

All data processing and analyses were performed by using R (R Core Team, 2017). ANOVAs on both reaction times (RTs) and proportions of correct trials (PCs)[3] were

[3] Accuracy was high across all Experiments reported in the current paper. In order to rule out that ceiling effects (and the resulting deviations from normality) had an impact on the PCs results, we also analyzed arc-sin

calculated by using 'afex' (Singmann et al., 2015), and follow-up contrasts on the model estimates by using 'phia' (De Rosario-Martinez et al., 2015). Raw data and corresponding processing scripts are available at https://osf.io/au7pv/

The overall accuracy in the inducer task was .93 (*SD*= .07). The overall accuracy was .93 (*SD*= .05) for the first diagnostic task and .94 (*SD*= .04) for the second diagnostic task. One participant had an accuracy of .64 in the inducer task, which was below 2.5 standard deviation of the group mean accuracy. Another participant had an accuracy of .79 in the first diagnostic task and .82 on the second diagnostic task, which was in both cases below 2.5 standard deviation of the group mean accuracy. Both participants were excluded from the analyses. In addition, only Target Runs were taken into account on which the inducer task was performed correctly (cf. Everaert et al., 2015; Liefooghe et al., 2012, 2013, 2016; Theeuwes et al., 2014).

First diagnostic task. To ensure that we replicated previous findings (e.g., Liefooghe, 2012, 2013), we tested for the presence of an IBC effect on the first diagnostic task, which preceded the probe(s) of the inducer task. Only trials with correct responses were considered for analysis of the RTs. This led to the removal of 6.45% of the total number of trials. Next, for each participant, trials with RTs more than 2.5 standard deviation above each cell mean were considered as outliers. This led to the removal of 3.11% of the total number of correct trials. RTs and PCs were each subjected to an ANOVA with the factor Congruency (congruent, incongruent) as a repeated-measures factor. RTs on congruent trials (*M*= 566, *SE*= 13) were smaller than RTs on incongruent trials (*M*= 577, *SE*= 14), $F(1, 35)= 12.69$, *MSE*= 196, $p<.01$, $\eta_p^2= .27$. PCs were higher on congruent trials (*M*= .94, *SE*= .007)

transformed PCs. Across all Experiments reported in the current paper, the main findings remained unchanged after transformation was applied; we do not return to this in Experiments 2-4 below (but see the processing scripts available at https://osf.io/au7pv/).

compared to incongruent trials (M= .93, SE= .008), $F(1, 35)$= 11.28, MSE= .0005, p<.01, η_p^2= .24. In line with previous studies, an IBC effect was thus present both for the RTs and the PCs.

Second diagnostic task. In order to test whether unapplied and no-longer relevant instructions still impact behavior automatically, the IBC effect measured on the unapplied and applied inducer stimuli was considered. For this analysis, Filler Runs were not considered. For the RTs, 5.97% of incorrect trials and 2.97% of outliers were discarded. RTs and PCs were each subjected to a 2 (Congruency) by 2 (Stimulus Type: applied inducer stimulus, unapplied inducer stimulus) repeated-measures ANOVA. Cell means and corresponding standard errors are presented in Table 1 (see also Figure 3 for RTs). RTs were smaller on congruent trials (M= 531, SE= 11) than on incongruent trials (M= 537, SE= 12), $F(1, 35)$= 7.02, MSE= 143, p<.01, η_p^2= .17. The main effect of Stimulus Type was not significant, F< 1, nor was its interaction with Congruency, F< 1. In view of our research question, additional contrasts were conducted to specifically test whether an IBC effect was present for both the unapplied and applied inducer stimuli. Both for the applied inducer stimuli, $F(1,35)$= 4.18, p < .05, η_p^2= .11 and for the unapplied inducer stimuli, $F(1,35)$= 6.09, p < .05, η_p^2= .15, the IBC effect was significant.

PCs were higher on congruent trials (M= .95, SE= .01) compared to incongruent trials (M= .93, SE= .01), $F(1, 35)$= 7.45, MSE= .001, p< .01, η_p^2= .18. The main effect of Stimulus Type, nor the interaction between Stimulus Type and Congruency were significant, F< 1. Additional contrasts again indicated that the IBC effect was significant both for applied inducer stimuli, $F(1, 35)$= 7.21, p< .05, η_p^2= .17 and unapplied inducer stimuli, $F(1, 35)$= 4.71, p< .05, η_p^2= .12.

INSERT TABLE 1 AROUND HERE

INSERT FIGURE 3 AROUND HERE

Additional analysis. An additional analysis was conducted in which we compared performance on the first and second diagnostic task. The first diagnostic task constitutes a dual-task situation as two sets of stimulus response mappings need to be maintained in memory (i.e., diagnostic task + pending inducer task), whereas the second diagnostic task reflects a single-task situation (i.e., diagnostic task alone). If the intention to perform the inducer task is relaxed after its completion, then performance on the second diagnostic task should be speeded. However, such difference could also be attributed to more extensive practice on the second diagnostic task compared to the first diagnostic task. In order to minimize the contribution of such practice effects, we focused on the final three trials of the first diagnostic task (respective means: 560ms, 571ms, and 561ms) and compared these with the second (550ms), third (546ms), and fourth (530ms) trials of the second diagnostic task (the first trial of the second diagnostic task is a switch trial, which is known to increase RTs; e.g., Vandierendonck et al., 2010). A 2 (Congruency) by 2 (Stimulus Type) by 2 (Diagnostic Task: first, second) repeated measures ANOVA was conducted. RTs were higher on the first diagnostic task (M= 563, SE= 14) compared to the second diagnostic task (M= 542ms, SE= 12), $F(1, 35)$= 18.42, MSE= 1741, p< .001, η_p^2= .34. In line with the previous analyses, RTs on congruent trials (M= 548, SE= 12) were significantly faster than RTs on incongruent trials (M= 557, SE= 13), $F(1, 35)$= 11.03, MSE= 543, p< .01, η_p^2= .24. Neither the main effect of Stimulus Type, nor the interactions were significant, all Fs<1.

Discussion

An IBC effect was observed for both applied and unapplied inducer stimuli, and there was no reliable statistical difference between both. The additional analysis indicated that the second diagnostic task was performed significantly faster than the first diagnostic task. A shift from a dual-task situation (diagnostic task + pending inducer task) to a single-task situation (diagnostic task alone) was thus apparent, which indicates that the intention to perform the inducer task was relaxed almost immediately after its completion. However, the unapplied instructions of this task still biased performance automatically, which suggest difficulties in disengaging from prepared instructions.

Whereas the results of Experiment 1 are interesting as they provide no support for directed dismantling, two concerns can be raised. First, participants had to infer that the instructed stimulus-response mappings of the inducer task were no longer relevant when this task was completed; an inference mainly driven by gaining experience over several practice runs (and potentially the earlier runs of the first experimental block) that such was the case. It could be argued that this constitutes a rather indirect and implicit way for participants to disengage themselves from the prepared instructions. Second, although the additional analysis minimized differences in practice between the first and second diagnostic task, Experiment 1 can never completely rule out the contribution of practice effects to the difference in overall reaction time between the first and second diagnostic task. In the Experiments below, we considered these concerns.

INSERT FIGURE 4 AROUND HERE

Experiment 2

In Experiment 2, participants were presented with Filler Runs and Target Runs (see Figure 4). As in Experiment 1, the diagnostic task required participants to decide whether a stimulus was presented in italic or upright. The Filler Runs started with the presentation of two new stimulus-response mappings of the inducer task (e.g., if 'kra', press right; if 'ple', press left). Following the encoding of these instructions, the diagnostic task started, which in turn was followed by a probe stimulus of the inducer task (e.g., 'kra' printed in green). In the Target Runs a similar sequence of events occurred. However, at some moment during the diagnostic task a cue (i.e., a circle surrounding the word) was presented – and remained on screen until the end of the run. The meaning of this cue was different across three between-subject conditions. In the inhibit-inducer-response condition, the cue indicated that the inducer task was no longer relevant: the probe of the inducer task would be presented at the end of the run, but participants did not have to respond to this probe. In the no-inducer-task condition, the cue also signaled that the inducer task was cancelled and that the probe of the inducer task was no longer presented at the end of the run. Both conditions thus included different operationalizations of cancelling the inducer task. In the control condition, the cue had no informational value and participants had to perform the inducer task at the end of the run, as it was the case for the Filler Runs. The precise meaning of the cue (or lack thereof for the control condition) was presented to participants at the start of the experiment, together with the other general information. Using the control condition as a baseline, the inhibit-inducer-response and no-inducer-task conditions allowed for exploring the impact of a canceled set of inducer task instructions on the trials after onset of the cue; we considered both the inhibit-inducer-response and no-inducer-task conditions to control for any potential impact of the mere presentation of the probe inducer at the end of the run. Note that in each condition, 1/3rd of the runs were Target Runs, while 2/3rd of the runs were Filler Runs in which no cue was presented and the inducer task was always presented at the end of a filler run. This way, an

overall context was created in which participants were encouraged to prepare for the instructions of the inducer task in the first place.

Participants were thus explicitly cued in Experiment 2 about the status of the inducer task during the ongoing diagnostic task, resulting in pre-cue (cf. first diagnostic task in Experiment 1) and post-cue (cf. second diagnostic task in Experiment 1) diagnostic trials, for both of which congruency was determined on the basis of having either a match (i.e., congruent) or mismatch (i.e., incongruent) between the specific responses signaled by the stimulus-response instructions of the inducer task versus those of the diagnostic task. We had two main research aims. First, we explored whether the shift from the dual-task situation (i.e., diagnostic task + pending inducer task) to the single-task situation (i.e. diagnostic task alone) would become immediately apparent in the performance on the diagnostic task. If the intention to perform the inducer task is relaxed or fully cancelled, overall RTs in the cued trials (i.e., trials presented with or after the onset of the cue, the latter of which stayed on the screen during the remainder of the run) of the diagnostic task in the Target Runs should be faster compared to their counterpart trials in the Filler Runs. Specifically, we compared the trials of the diagnostic task of the Target Runs on which the cue was presented, with the trials of the diagnostic task of the Filler Runs which had the same serial position within the set of diagnostic task trials. This way, the effect on overall RTs of cueing that the inducer task was cancelled (no-inducer-task condition) or should not be responded to (inhibit-inducer-response condition) was tested while controlling for practice effects, as Filler Run trials allowed equal opportunity for practice of the diagnostic task as the trials in the Target Runs. The second – and more central question – was whether prepared, but unexecuted and no longer relevant instructions still automatically impact behavior. To this end, we tested whether the IBC effect (dis)appeared on the cued trials of the diagnostic task of the Target Runs in the no-inducer-task and inhibit-inducer-response conditions.

Method

Participants. Fifty-eight new participants were recruited (No Inducer Task: n= 20; Inhibit Inducer Response: n= 19; Control: n=19) and were paid 10 Euro for participation. As noted above, ethical approval was obtained from the ethical committee of Ghent University.

Task & Materials. For each participant, 84 pairs of stimulus-response mappings of the inducer task were randomly created on the basis of a list of 168 three-letter non-words, which was generated using WordGen (Duyck et al., 2004). The diagnostic task consisted of two parts (i.e., pre-cue and post-cue trials), each consisting of either 4 or 8 trials. Therefore, the total length of the diagnostic task (across pre-cue and post-cue parts) could be 8 trials (for n=21 runs), 12 trials (for n=21 runs with 4 trials in the pre-cue and 8 trials in the post-cue diagnostic task; and for n=21 runs with 8 trials in the pre-cue and 4 trials in the post-cue diagnostic task), or 16 trials (for n=21 runs).

Per 21 runs of varying length (see above), fourteen runs were Filler Runs in which no cue was presented during the diagnostic task. The remaining seven runs were Target Runs on which a cue (a circle surrounding the target stimulus) appeared when the first trial of the second half of the diagnostic task was presented, thus either after 4 or 8 trials of the diagnostic task. This cue remained on screen until the end of the diagnostic task.

Procedure. Testing conditions were similar as in Experiment 1. Before starting the experiment, participants were instructed on the stimulus-response mappings for the diagnostic task as well as about the condition-specific meaning of the cue (i.e., no-inducer-task, inhibit-inducer-response, or control; between-subject manipulation). The experiment consisted of 1 practice block and 6 experimental blocks of 12 runs each. Each block consisted of 4 Target Runs (one of each of the aforementioned combinations of pre-cue and post-cue trial sets: 8|8 trials, 4|8 trials, 8|4 trials, 8|8 trials) and 8 Filler Runs (2 x 4|4 trials, 2 x 4|8 trials, 2 x 8|4

trials, 2 x 8|8 trials). Each run started with the presentation of two new stimulus-response mappings of the inducer task for a maximum time of 20sec. or until participants pressed the spacebar. The response deadline and inter-trial interval for both tasks were 2000ms and 750ms respectively. The experiment lasted for approximately one hour.

Results

The first block of the experiment was a practice block and not included in the analyses. Across the experimental blocks, overall accuracy was .89[4] (*SD*= .09) for the inducer task and .94 (SD= .07) for the diagnostic task. We used the same outlier criteria as in Experiment 1 and the data of two participants was removed from the analyses (inducer task: .54 accuracy; diagnostic task: .52 accuracy).

Filler Runs. In order to rule out differences in instruction encoding due to the presence of specific target runs in each between-subjects condition, we analyzed the IBC effect in the Filler Runs of the three conditions. As such, we assessed if the presence of Target Runs in which the inducer task is either canceled or should not be responded to, induced any observable change in the degree by which participants prepared for the stimulus-response mappings of the inducer task in the first place.

For the RT analysis, 4.99% of incorrect trials and 2.72% of outliers were removed. RTs and PCs were each subjected to a 3 (Condition) by 2 (Congruency) mixed ANOVA with repeated measures on the last factor. Cell means and corresponding standard errors are presented in Table 2. RTs were significantly shorter on congruent trials (*M*= 538, *SE*= 11) than on incongruent trials (*M*= 550, *SE*= 11), $F(1, 53)= 34.58$, *MSE*= 111, $p< .001$, $\eta_p^2= .39$.

[4] Accuracies were .90 in the Control condition, .88 in the No inducer task condition and .90 in the inhibit inducer response condition.

Neither the main effect of Condition, $F<1$, nor the two-way interaction was significant, $F(2, 53)= 1.80$, $MSE= 111$, $p= .17$, $\eta_p^2= .06$. PCs were higher on congruent trials ($M= .96$, $SE= .01$) than on incongruent trials ($M= .94$, $SE= .01$), $F(1, 53)= 23.67$, $MSE= .001$, $p< .001$, $\eta_p^2= .31$. The main effect of Condition, $F<1$ and the two-way interaction were not significant, $F < 1$. No significant differences were thus detected in the way instructions were encoded in the three conditions.

INSERT TABLE 2 AROUND HERE

INSERT FIGURE 5 AROUND HERE

Comparison Filler versus Target Runs. Only cued trials of the diagnostic task and their serial counterparts in the Filler Runs were considered in these analyses. Our main question was if the IBC effect is modulated by the informational value of the cue ('no inducer task', 'inhibit inducer response', 'control') and whether overall performance on the diagnostic task trials of the Target Runs was faster compared to performance on the diagnostic task trials of the Filler Runs (see Figure 5).

For the RT analyses, 5.22% of the trials were incorrect and 2.58% of the remaining trials were outliers. RTs and PCs were each subjected to a 3 (Condition) by 2 (Run Type: filler run, target run) by 2 (Congruency) mixed ANOVA with repeated measures on the last two factors (see Table 2 for cell means and corresponding standard errors, and see Figure 5 for specifically RTs). RTs were significantly shorter on congruent trials ($M= 527$, $SE= 10$) than on incongruent trials ($M= 537$, $SE= 11$), $F(1, 53)= 22.92$, $MSE= 268$, $p< .001$, $\eta_p^2= .30$. RTs were significantly shorter on target runs ($M= 525$, $SE= 11$) than on filler runs ($M= 538$,

SE= 11), $F(1, 53)= 14.24$, *MSE*= 698, $p< .001$, $\eta_p^2= .21$. The main effect of Condition was not significant, $F< 1$. On the one hand, Condition and Run Type interacted, $F(1, 53)= 8.18$, *MSE*= 698, $p< .01$, $\eta_p^2= .24$: RTs were shorter in the Target Runs compared to the Filler Runs for the no-inducer-task condition, $M_{diff.}= 24$, $F(1, 53)= 15.67$, $p< .001$, $\eta_p^2= .23$, and for the inhibit-inducer-response condition, $M_{diff}= 23$, $F(1, 53)= 13.34$, $p< .001$, $\eta_p^2= .20$, but not for the control condition, $M_{diff}= 7$, $F(1, 53)= 1.24$, $p= .27$, $\eta_p^2= .02$. RTs thus decreased when the inducer task was cancelled or no longer had to be responded to. On the other hand, the interaction between Condition, Run Type, and Congruency was not significant, $F< 1$. Additional contrasts indicated that the IBC effect was still significant in the no-inducer-task condition, $M_{diff}= 14$, $F(1, 53)= 5.38$ $p< .05$, $\eta_p^2= .09$, and the inhibit-inducer-response condition, $M_{diff}= 13$, $F(1, 53)= 4.57$, $p< .05$, $\eta_p^2= .08$.

For the PCs, only the main effect of Congruency was significant, indicating that PCs were higher on congruent trials (*M*= .95, *SE*= .00) than on incongruent trials (*M*= .94, *SE*= .00), $F(1, 53)= 15.56$, $p< .001$, $\eta_p^2= .23$. None of the remaining main effects or interactions were significant, largest F-value: $F(1, 53)= 1.69$, *MSE*= .001, $p= .19$, $\eta_p^2= .06$.

Discussion

The results of Experiment 2 are consistent with those of Experiment 1. Responses on the cued trials of the diagnostic task of the Target Runs were faster compared to their counterparts in the Filler Runs for the no-inducer-task and inhibit-inducer-response conditions. This was not the case for the control condition. This faster performance suggests that the intention to perform the inducer task was almost immediately relaxed, when the context shifted from a dual-task situation in the initial un-cued trials of the diagnostic task (diagnostic task + pending inducer task) to a single-task situation in the later cued trials of the diagnostic task (diagnostic task alone). However, despite this performance speed-up indicating processing of the cancellation cues, a significant IBC effect was still observed for the cued trials of the

diagnostic task in the target runs of the no-inducer-task and inhibit-inducer-response conditions.

Taken together, the results of Experiment 2 again provide no support for directed dismantling: Whereas the intention to perform the inducer task was clearly cancelled, the automatic impact of its instructions remained present. The critical reader may be concerned that the experimental demands of Experiment 2 were simply too hard for participants. Specifically, due to the fast pace of events in the runs of Experiment 2, participants may not have had sufficient time to dismantle the (no-longer relevant) instructions. In Experiment 3, we aimed to remedy this potential problem by presenting the cancellation cue as a separate event (not within a trial of the ongoing diagnostic task) with self-paced processing time.

INSERT FIGURE 6 AROUND HERE

Experiment 3

As in the experiments above, the diagnostic task in Experiment 3 required participants to decide whether a stimulus was presented in italic or upright. As outlined in Figure 6, participants were presented with three types of runs. The Completion Runs started with the presentation of the stimulus-response mappings of the inducer task, followed by the first diagnostic task. Next, the probe of the inducer task was presented, which was followed by the second diagnostic task. In the Cancellation Runs the same sequence of events occurred, with the exception that the probe of the inducer task was replaced by a cancellation cue in the form of the message ‘cancelled’, indicating that the inducer task was cancelled (this was explicitly explained to participants in the general instructions). Participants had to acknowledge this message by pressing the spacebar. In Proceed Runs, the explicit message ‘proceed’ was

presented between the first and the second diagnostic tasks, indicating that the probe of the inducer task would be presented at the end of the entire run (i.e., after the second diagnostic task). Again, participants had to acknowledge this message by pressing the spacebar. At the end of a Proceed Run, the probe of the inducer task was actually presented. Across all run types, and for both the first and second diagnostic task trials, congruency was determined on the basis of having either a match (i.e., congruent) or mismatch (i.e., incongruent) between the specific responses signaled by the stimulus-response instructions of the inducer task versus those of the diagnostic task.

In contrast to Experiment 2, the cues to abandon or proceed with the instructions of the inducer task were presented as messages separated from the diagnostic task trials, and required a response to be acknowledged. This allowed for more time (in comparison to Experiment 2) to further disengage from the prepared instructions and thus potentially attenuate the automatic impact of these instructions. An additional change compared to both the previous experiments is that the first diagnostic task (i.e., prior to the proceed/cancel messages or probe of the inducer task) contained either 0 or 4 trials. The reason for this manipulation was that in Experiments 1 and 2 the IBC effect could have potentially been inflated by unintentional (misapplication) or intentional (mental simulation) partial execution of the inducer task during the diagnostic task (Meiran et al., 2015a). This would leave open the possibility that the IBC effect observed in the previous experiments was not purely based on instructed representations alone (but see Meiran et al., 2017). Therefore, Experiment 3 also allowed us to test in a straightforward manner whether the IBC effect was still present for instructions that were instantly cancelled or completed (namely, in the case of 0 trials on the first diagnostic task). Finally, we increased the power of our experimental contrasts by (a) using a completely within-subjects design and (b) increasing the sample size. For the latter purpose, Experiment 3 was conducted as an online internet study.

Method

Participants. A total of 101 English-speaking volunteers participated online via the Prolific Academic website (https://prolific.ac). As noted above, ethical approval was obtained from the ethical committee of Ghent University.

Task & Materials. For each participant, 48 pairs of stimulus-response mappings of the inducer task were randomly created on the basis of a list of 96 four-letter highly frequent English nouns, which was selected from the SUBTLEX-UK database (see Van Heuven et al., 2014). These pairs were randomly assigned to 16 Cancellation Runs; 16 Proceed Runs; and 16 Completion Runs. An outline of the three different types of runs is presented in Figure 6.

A Cancelation Run started with the presentation of two stimulus-response mappings of the inducer task. Following the encoding of these mappings, either four or no trials of the diagnostic task were presented, which were followed by the message 'CANCELLED' in the middle of the screen. Participants acknowledged this message by pressing the spacebar. Following the cancellation message, 12 trials of the second diagnostic task were presented. A run ended after the completion of this second diagnostic task. Completion Runs were similar to Cancellation Runs except that the cancellation message was replaced by a probe of the inducer task and participants thus had to apply one of the two stimulus-response mappings of the inducer task (instructed at the onset of the run). Following this probe, the second diagnostic task was presented and a run ended after the completion of this second series. Proceed Runs differed in two ways compared to the previous two runs. First, the message 'PROCEED' was now presented during the run and participants had to acknowledge this message by pressing the spacebar. This response was again followed by a second series of 12 trials of the diagnostic task. However, at the end of this second diagnostic task, the probe stimulus of the inducer task was presented, which participants had to respond to by applying the corresponding stimulus-response mapping of the inducer task. Participants were thus now

required to maintain the stimulus-response mappings of the inducer task also during the second diagnostic task.

Procedure. The experiment was programmed in Inquisit 4.0 and hosted via Inquisit Web (Millisecond Software, Seattle, WA). After providing informed consent and completing demographic questions, the general instructions of the experiment were provided. Since there was no opportunity to paraphrase the instructions, we added different figures in our instructions outlining the different types of runs and participants could scroll back and forth between the different instruction pages, until they completely understood the different task demands.

Following the general instructions (i.e., stimulus-response mappings for the diagnostic task; precise meaning of the cue), a practice block was presented, which consisted of 6 runs: 2 Cancellation Runs, 2 Proceed Runs, and 2 Completion Runs. For each run type either four trials of the diagnostic task followed the instruction of the stimulus-response mappings of the inducer task or these mappings were immediately followed by the imperative message (i.e., 'cancel', 'proceed') or the probe of the inducer task in the completion runs. After this practice block, the actual experiment started, which consisted of 7 experimental blocks each consisting of 6 types of runs (i.e., 2 runs for each type). Blocks were separated by a small break. After the experiment, participants received the opportunity to provide some feedback about their experiences during the experiment. More specifically, we asked whether participants had the impression that they provided useful results and were able to commit to the different task demands or whether they experienced (technical) difficulties during the experiment and had serious doubt about their performance.

The start of each run was announced by presenting five exclamation marks "!!!!!" for 1500ms, printed in white on the center of a black screen. After the presentation of two stimulus-response mappings for the inducer task, participants started the diagnostic task,

responding to upright stimuli by pressing the ‘v’ key (left) and to italic stimuli by pressing the ‘n’ key (right). Incorrect and slow (maximum response deadline of 2000ms) responses were followed by error feedback, which consisted of the word “WRONG” presented in the screen center in red for 200ms. The messages ‘CANCELLED’ (Cancellation Runs) or ‘PROCEED’ (Proceed Runs) were presented in green for 2000ms, or until participants pressed spacebar. Similarly, the probes of the inducer task in the Completion Runs were also presented for 2000ms or until participants pressed one of the two response keys (i.e., left ‘v’-key; right ‘n’-key). Messages and probe stimuli were preceded by a 750ms lasting fixation cross also presented in green. Slow and incorrect responses were followed by error feedback, using the same parameters as in the trials of the diagnostic task.

The second diagnostic task started immediately after the messages or probe item and employed the same parameters as the first diagnostic task. The Cancellation and Completion Runs ended at the end of this second diagnostic task. For the Proceed Runs, an additional probe stimulus of the inducer task was presented and the run ended following a response to this stimulus.

Results

Across experimental blocks, overall accuracy was .65 (*SD*= .48) for the inducer task and .85 (*SD*= .36) for the diagnostic task. Performance was thus worse as compared to the above reported lab-based experiments. In addition, we detected missing cells for our target analyses on the diagnostic task, when applying the restrictions of the previous experiments, namely (a) only trials of the diagnostic task are considered for runs on which the inducer task was performed correctly and (b) only RTs of correct trials of the diagnostic task are analyzed. Accordingly, 15 participants with empty cells were removed from the initial sample. Four more participants were considered as group outliers and discarded (inducer task: .25;

diagnostic task: .49, .39, .50). A remaining sample of 82 participants was thus available for further analyses.

Overall performance in this final sample was .69 (*SD*= .46) for the inducer task[5] and .92 (*SD*= .27) for the diagnostic task. Performance on the inducer task thus remained lower compared to Experiments 1 and 2, whereas performance on the diagnostic task was in the same range. However, inter-individual variability was large in both cases. We were concerned with the fact that the number of observations per cell would still be too low for the RT analyses. At the same time, we were reluctant to set an arbitrary minimum for the number of observations per cell. First, this would further reduce the sample of participants and thus the overall power of our analyses. Second, our conclusions may differ depending on the exact value of this arbitrary cut-off. Accordingly, besides reporting RTs and PCs as in Experiments 1 and 2, we also report the Inverse Efficiency Score[6] (IES, Townsend & Ashby, 1978). The IES is an estimation of the RT adapted for the frequency of incorrect responses, which in the current context is relatively high and variable over participants. Accordingly, we tested whether the IES analyses converged with RTs and PCs analyses, without further truncating our sample. We first consider the first diagnostic task, which preceded (on half of the runs) the messages (Cancellation Runs, Proceed Runs) or the probe of the inducer task (Completion Runs). Next, we turn to the second diagnostic task, which follows upon these interventions.

First diagnostic task. For the RT analysis, 8.41% incorrect trials and 2.19% outliers were removed. RTs and PCs were each subjected to a repeated measures ANOVA with only one factor, namely Congruency. RTs on incongruent trials were significantly longer (*M*= 607, *SE*= 8.43) than RTs on congruent trials (*M*= 584, *SE*= 8.09), $F(1, 81)= 25.58$, *MSE*=

[5] Accuracies of .71 and .64 for Completion and Proceed Runs, respectively.

[6] The Inverse Efficiency Score takes the ratio of the average correct RT and PC (i.e., RT/PC) per cell per participant.

915, $p < .001$, $\eta_p^2 = .24$. PCs on congruent trials were significantly larger ($M = .95$, $SE = .01$) than PCs on incongruent trials ($M = .88$, $SE = .01$), $F(1, 81) = 32.64$, $MSE = .01$, $p < .001$, $\eta_p^2 = .29$. The difference between congruent and incongruent trials was also present for the IES, $F(1, 81) = 32.47$, $MSE = 8990$, $p < .001$, $\eta_p^2 = .29$. An IBC effect was thus present at the early onset of a run, indicating that participants immediately prepared for the instructed stimulus-response mappings of the inducer task.

INSERT TABLE 3 AROUND HERE

INSERT FIGURE 7 AROUND HERE

Second diagnostic task. For the RT analysis, 7.36% of incorrect trials and 2.41% of outliers were removed. RTs and PCs were each subjected to a 3 (Run Type: Cancellation Run, Completion Run, Proceed Run) by 2 (First Diagnostic Task: absent, present) by 2 (Congruency) repeated measures ANOVA. Cell means and corresponding standard errors are presented in Table 3.

For the RTs, the main effect of Run Type was significant, $F(2, 162) = 6.98$, $MSE = 2191$, $p < .01$, $\eta_p^2 = .08$. RTs were longer on Proceed Runs ($M = 563$, $SE = 7.48$) compared to Completion Runs ($M = 554$, $SE = 6.91$), $F(1, 81) = 4.97$, $p < .05$, $\eta_p^2 = .06$, and Cancellation Runs ($M = 550$, $SE = 7.66$), $F(1, 81) = 16.39$, $p < .001$, $\eta_p^2 = 17$. Cancelling or completing the inducer task thus resulted in an overall drop in RTs in the second diagnostic task, which again suggests that participants could relatively easily relax the intention to perform the inducer task.

The main effect of Congruency was also significant, $F(1, 81)= 43.29$, $MSE= 762$, $p< .001$, $\eta_p^2= .35$, and longer RTs were observed on incongruent trials ($M= 562$, $SE= 7.17$) compared to congruent trials ($M= 550$, $SE= 7.04$). Finally, the main effect of First Diagnostic Task was significant, $F(1, 81)= 4.54$, $MSE= 1187$, $p< .05$, $\eta_p^2= .05$. RTs were longer when a first diagnostic task had to be performed ($M= 558$, $SE= 7.34$) then when this was not the case ($M= 554$, $SE= 6.92$).

The interaction between Run Type and Congruency was not significant, $F<1$. An IBC effect was present in the Cancellation Runs, $M_{diff}= 10.15$, $F(1, 81)= 15.69$, $p< .001$, $\eta_p^2= .16$, the Completion Runs, $M_{diff}= 11.35$, $F(1, 81)= 19.79$, $p< .001$, $\eta_p^2= .20$, and the Proceed Runs, $M_{diff}= 13.49$, $F(1, 81)= 20.82$, $p< .001$, $\eta_p^2= .14$. Hence, even if the inducer task was cancelled or completed, an IBC effect remained present.

The remaining two-way interactions were not significant, all $Fs< 1$. However, the three-way interaction was marginally significant, $F(2, 162)= 2.64$, $MSE= 489$, $p= .07$, $\eta_p^2= .03$. This interaction is presented in Figure 5. If anything, this trend towards an interaction seems to be driven by slightly smaller IBC effects for the Cancellation Runs and the Completion Runs when a diagnostic task was present prior to the cancellation message or the probe stimulus, respectively. Nevertheless, IBC effects were present in all three run types, also specifically for the case in which no diagnostic task (i.e., 0 trials) was presented after instruction encoding (Fs>7.8, ps<.01) – a finding we return to in the Discussion section below.

For the PCs, the main effect of Congruency was significant, $F(1, 81)= 21.17$, $MSE= .01$, $p< .001$, $\eta_p^2= .21$, indicating that PCs were higher on congruent trials ($M= .93$, $SE= .07$) compared to incongruent trials ($M= .90$, $SE= .01$). None of the remaining main effects and interactions were significant. Of prime interest, was the absence of an interaction between

Congruency and Run Type, $F < 1$, which indicated that the IBC effect was present in all three types of runs: Cancellation runs, $M_{diff} = .03$, $F(1, 81) = 20.37$, $p < .001$, $\eta_p^2 = .20$, Completion runs, $M_{diff} = .04$, $F(1, 81) = 15.53$, $p < .001$, $\eta_p^2 = .16$, Proceed runs, $M_{diff} = .03$, $F(1, 81) = 10.27$, $p < .001$, $\eta_p^2 = .11$.

For the IES, the main effects of Congruency and Run Type were significant, $F(1, 81) = 24.04$, $MSE = 20114$, $p < .001$, $\eta_p^2 = .23$, and, $F(2, 162) = 5.76$, $MSE = 4976$, $p < .01$, $\eta_p^2 = .07$, respectively. IES were higher on Proceed Runs compared to Completion runs, $F(1, 81) = 7.13$, $p < .01$, $\eta_p^2 = .08$, and Cancellation runs, $F(1, 81) = 10.47$, $p < .01$, $\eta_p^2 = .11$. The interaction between Congruency and Run Type was not significant, $F < 1$. An IBC effect was present in all three types of runs: Cancellation Runs, $F(1, 81) = 21.92$, $p < .001$, $\eta_p^2 = .21$, Completion Runs, $F(1, 81) = 18.57$, $p < .001$, $\eta_p^2 = .19$, Proceed Runs, $F(1, 81) = 14.24$, $p < .001$, $\eta_p^2 = .15$.

Discussion

The findings of Experiment 3 further corroborate the findings of the previous two experiments. Overall RTs in the second diagnostic task were slower on the Proceed Runs compared to the Cancellation and Completion Runs. This finding again suggests that the dual-task demand could be relaxed and the intention to perform the inducer task was easily cancelled. However, the IBC effect in the second diagnostic task did not differ significantly between the three types of runs and a significant IBC effect was present even in the (explicit) Cancellation Runs. Automatic effects were thus present for unexecuted instructions that were no longer relevant.

The results of Experiment 3 further indicate that these automatic effects were present even if no diagnostic task was presented before cancelling or executing the instructions. This finding rules out the possibility that the IBC effect observed in the second diagnostic task is based on the misapplication of the inducer task during the first diagnostic task (cf. Meiran et

al., 2015a). Indeed, the fact that we observed clear IBC effects both in conditions with and without a first diagnostic task, indicates that in the task exploited in the current paper, inducer task mappings are instantly implemented at the level of procedural working memory (cf. Wenke et al., 2009). Taken together, the results of Experiments 1-3 indicate that it is very difficult to disengage from prepared instructions, which keep on triggering automatic effects even when the task they belong to was cancelled in different ways. Hence, the current study repeatedly failed to find any support for directed dismantling, and thus speak against an active removal mechanism operating to serve capacity-limited procedural working memory. In Experiment 4, we extended our exploration by not merely cancelling instructions, but by replacing them by a new set of instructions.

INSERT FIGURE 8 AROUND HERE

Experiment 4

Experiment 4 was also conducted online, and as in the experiments above, the diagnostic task required participants to decide whether a stimulus was presented in italic or upright. Experiment 4 consisted of Proceed Runs (as in Experiment 3), and Replace Runs in which the initial set of instructed stimulus-response mappings for the inducer task (presented at the start of the run) was later replaced by a new set of mappings (see Figure 8). The Replace Runs consisted of the following phases: (a) presentation of a first set of stimulus-response mappings of the inducer task, (b) a first diagnostic task with stimuli of the first set of stimulus-response mappings of the inducer task, (c) the presentation of a new set of stimulus-response mappings of the inducer task to replace the first set, (d) a second diagnostic task *again* with stimuli of the *first* set of instructions of the inducer task, and (e) a trial of the inducer task probing the *second* set of instructions of the inducer task. In the Replace Runs, instructions of the inducer

task are thus not only cancelled but actually replaced by a new set of instructions. Nevertheless, for all run types, congruency for both the first and second diagnostic task trials was determined on the basis of having either a match (i.e., congruent) or mismatch (i.e., incongruent) between the specific responses signaled by the stimulus-response instructions of the *initial* inducer task versus those of the diagnostic task (i.e., the second set of inducer task instructions that replaced the initial ones were not considered with respect to congruency).

Because we wanted to create an overall context that encouraged participants to prepare instructions in the first place, Filler Runs were also presented. The Filler Runs only included one series of trials of the diagnostic task (i.e., instructed stimulus-response mappings of the inducer task – diagnostic task – probe of the inducer task). This way participants had to apply the initially prepared stimulus-response mappings of the inducer task in 2/3rd of the runs – thus probing them to do so across all runs (as the run type was not known at the start of the run). Within this design, our central question was whether cancelled and replaced instructions would still impact behavior automatically in the second diagnostic task of the Replace Runs.

Method

A sample of 102 new participants was tested using Prolific. As noted above, ethical approval was obtained from the ethical committee of Ghent University. Sixty-Four pairs of stimulus-response mappings of the inducer task were randomly created on the basis of a list of 128 four-letter highly frequent English nouns, which was selected from the SUBTLEX-UK database (see Van Heuven et al., 2014). These pairs were randomly assigned to 16 Filler Runs; 16 Proceed Runs; and 16 Replace Runs. Note that for the Replace Runs two pairs of stimulus-response mappings were created: one for the first inducer task and one for the second inducer task.

Replace Runs started with the presentation of two stimulus-response mappings of the inducer task. After either 0 or 4 trials of the first diagnostic task, two new stimulus-response mappings were presented. The presentation of these new mappings was followed by the second diagnostic task, which still employed the stimuli related to the first set of stimulus-response mappings of the inducer task that was instructed at the onset of the run (hence, stimuli from the second set of instructions were not used here). At the end of the Replace Runs, a probe stimulus was presented, which required the application of the second set of instructed stimulus-response mappings. Proceed Runs were identical to the Replace Runs except for the fact that no new instructions were provided – but rather the word 'PROCEED' was presented. Filler Runs started with the instruction of two stimulus-response mappings of the inducer task, followed by either 0 or 4 trials of the first diagnostic task, after which the probe stimulus of the inducer task was presented. These runs ended after responding to this probe stimulus. All 3 types of runs are illustrated in Figure 9. Timing parameters were the same as in Experiment 3 and the experiment lasted for approximately 30 minutes.

Results

Overall accuracy was .85 (*SD*= .36) in the diagnostic task and .72 (*SD*= .45) in the inducer task. Performance was thus in line with Experiment 3 and clear inter-individual variability was observed. After controlling for missing cell values, 5 participants were no longer considered for analyses. An additional 6 participants were considered as outliers (diagnostic task accuracies of .50 or less). The remaining sample thus consisted of 91 participants and overall accuracy was .90 (*SD*= .29) in the diagnostic task and .75 (*SD*= .43) in the inducer task[7]. As in Experiment 3, we also considered IES in addition to RTs and PCs.

[7] Accuracies per run type were .78 (Completion Runs), .72 (Proceed Runs), and .72 (Replace Runs).

First diagnostic task. For the RT analyses, 9.02% of incorrect trials and 2.41% of outliers were removed. RTs and PCs were each subjected to a repeated measures ANOVA with Congruency as a factor. RTs on incongruent trials were significantly longer (M= 633, SE= 10.19) than RTs on congruent trials (M= 613, SE= 11.55), $F(1, 90)= 8.44$, $MSE= 2200$, $p< .01$, $\eta_p^2= .09$. PCs on congruent trials were significantly larger (M= .92, SE= .01) than PCs on incongruent trials (M= .87, SE= .01), $F(1, 90)= 20.01$, $MSE= .01$, $p< .001$, $\eta_p^2= .18$. Similarly, the IBC effect was significant for the IES, $F(1, 90)= 20.01$, $MSE= 25980$, $p< .001$, $\eta_p^2= .11$.

INSERT TABLE 4 AROUND HERE

INSERT FIGURE 9 AROUND HERE

Second diagnostic task. For the RT analysis, 7.65% of incorrect trials and 2.86% of outliers were removed. RTs and PCs were each subjected to a 2 (Run Type: Replace Run, Proceed Run) by 2 (First Diagnostic Task: absent, present) by 2 (Congruency) repeated measures ANOVA. Cell means and corresponding standard errors are presented in Table 4 (see also Figure 9).

For the RTs, the main effect of Run Type was not significant, $F<1$. The main effect of Congruency was significant, $F(1, 90)= 5.39$, $MSE= 1898$, $p< .05$, $\eta_p^2= .06$, and longer RTs were observed on incongruent trials (M= 585, SE= 9.67) compared to congruent trials (M= 578, SE= 9.24). Finally, the main effect of First Diagnostic Task was not significant, $F< 1$.

The interaction between Run Type and Congruency was marginally significant, $F(1, 90)= 3.64$, $MSE= 1155$, $p= .06$, $\eta_p^2= .04$. A significant IBC effect was present in the Proceed runs, $M_{diff}= 12.31$, $F(1, 90)= 12.07$, $p< .001$, $\eta_p^2= .12$, but not in the Replace Runs, $M_{diff}= 2.69$, $F< 1$. The remaining interactions were not significant, largest F-value: 1.20.

For the PCs, the main effect of Congruency was significant, $F(1, 90)= 12.30$, $MSE= .01$, $p< .001$, $\eta_p^2= .12$, indicating that PCs were higher on congruent trials ($M= .92$, $SE= .01$) compared to incongruent trials ($M= .90$, $SE= .01$). Neither the main effect of Run Type, nor the main effect of First Diagnostic Task were significant, both $Fs< 1$. The interaction between Congruency and Run Type was significant, $F(1, 90)= 12.30$, $MSE= .0001$, $p< .001$, $\eta_p^2= .12$. A significant IBC effect was observed in the Proceed Runs $M_{diff}= .03$, $F(1, 90)= 19.39$, $p< .001$, $\eta_p^2= .18$, but not in the Replace Runs, $M_{diff}= .00$, $F< 1$. All remaining interactions were not significant, all $Fs< 1$.

The IES mirrored the previous pattern of results. An IBC effect was present, $F(1, 90)= 10.57$, $MSE= 11382$, $p< .001$, $\eta_p^2= .11$, and the interaction between Congruency and Run Type was significant, $F(1, 90)= 7.53$, $MSE= 8855$, $p< .01$, $\eta_p^2= .08$. The IBC effect was significant for the Proceed Runs, $F(1, 90)= 13.69$, $p< .001$, $\eta_p^2= .13$, but not for the Replace Runs, $F<1$.

Discussion

Replacing initial instructions by a new set of instructions in the Replace Runs attenuated the IBC effect. The interaction between the IBC effect and run type (Replace versus Proceed Runs) was significant for the PCs and IES and marginally significant for RTs. In all three measures, there was a significant IBC effect in the Proceed Runs but not in the Replace Runs. Taken together, the automatic impact of prepared but unexecuted instructions persists even

when these instructions are no longer relevant (Experiments 1-3), and to reduce this impact instructions need not only to be cancelled but to be replaced by new instructions.

General Discussion

Instructions are pivotal in enabling and boosting performance across a large range of activities in modern society (including almost every laboratory task used in the experimental study of human cognition). We can instantly (re)configure our control systems through abstract instructions received from experts, processing rich and context-specific declarative input that can be quickly translated into appropriate settings in procedural working memory (e.g., Brass et al., 2017). With their reflexive qualities, instructions allow for fast yet appropriate actions given a task context. Yet, goals, tasks, and contexts – and thus appropriate reflexes – can change at any moment. The present study focused on how flexibly we can dismantle newly instructed stimulus-response rules that suddenly lose their relevance.

To test our ability of 'directed dismantling', we measured IBC effects in the inducer-diagnostic paradigm (Liefooghe et al., 2012) following the cancellation of the inducer task – before the inducer task was ever executed. The latter ensured that we studied procedural working memory, with as little influence from more permanent memory traces as possible (see below). Cancellation of the inducer task was cued in different ways across Experiment 1-3, yet the results of these three experiments converge upon two main findings. First, retaining task instructions comes with a dual-task cost that can be easily undone once people are instructed that the inducer task is no longer relevant. This was evidenced by an instant speed-up in performance when shifting from a dual-task context (diagnostic task + pending inducer task) to a single-task context (diagnostic task alone). This speed-up is generally relevant here because it indicates that the cancellation cues were processed. The second main finding of the present study is that IBC effects are still present when the inducer task is either completed or

cancelled, which indicates that the automatic impact of the instructed stimulus-response mappings of the inducer task remained. Finally, Experiment 4 further suggests that the impact of no longer relevant instructions can be attenuated only when new instructions need to be encoded and prepared for in replacement of the canceled instructions. Taken together, the results of the present study indicate that it is difficult to dismantle newly formed procedural representations and their reflexes once they are prepared for, and they provide no support for an active removal mechanism serving capacity-limited procedural working memory.

Instructions in action

Humans have the ability to flexibly chose whether or not to encode instructions (e.g., Wenke et al., 2009), and if so, to effortlessly implement them (e.g., Liefooghe et al., 2012). From there, the procedural representations elicited by instructions are rather self-supporting in narrowing cognitive flexibility to attain a specific action (cf. Hommel, 2015; see also Cole et al., 2017). Specifically, as indicated by the IBC effect, stimulus-response mappings maintained in procedural working memory will automatically generate response (selection) processes once an imperative stimulus is processed. As noted above, this prepared reflex mechanism provides for a mechanistic, homunculus-free understanding of instruction following, enriching the various more general theories that pursue a mechanistic understanding of action control (e.g., Abrahamse et al., 2016; Botvinick et al., 2005; Egner, 2014).

A central finding of the current study is that the cognitive system has little control over the impact of newly encoded stimulus-response rules. In Experiments 2 and 3 it was observed that explicitly instructing participants that an initial set of instructions was no longer relevant, is not sufficient to cancel out the automatic effects of these initial instructions. The latter automatic impact demonstrates the power of instructions (once prepared for) to steer cognition and behavior. Yet, whereas (implementation) instructions seem fully adequate to set

up procedural representations and their corresponding action tendencies, further (cancellation) instructions seem paradoxically limited in undoing again these tendencies. This exposes a highly selective power of instructions in action control. Indeed, in line with the current findings, Braem and colleagues (2017) failed to modulate automatic effects of instructions on the basis of additional context instructions (see also Liefooghe & Verbruggen, 2019 for a similar manipulation). Although we do not rule out that more powerful cancellation manipulations may better succeed in eliminating the automatic impact of instructions, Experiment 4 suggests that perhaps the only way to dismantle newly instructed stimulus-response mappings is to replace them by new mappings.

Directed forgetting in declarative working memory has been linked to inhibitory processes that weaken (the access to) declarative representations (e.g., Aguirre et al., 2017). The current findings do not indicate a similar inhibitory mechanism at play for no longer relevant procedural representations (i.e., we observed no directed dismantling). This seems at odds with the often emphasized importance of different inhibitory mechanisms that are needed to flexibly coordinate different actions or tasks. For example, Koch et al. (2010) highlighted the role of inhibition in task switching, which is a major paradigm to study such cognitive flexibility. In this paradigm participants are required to switch between different tasks and the performance costs (i.e., the switch cost; see for a review Vandierendonck et al., 2010) that this imposes, has often been interpreted in terms of inhibitory mechanisms needed to expel the currently irrelevant task. Consider, for instance, the n-2 task repetition cost (e.g., Schuch & Koch, 2003): 'AB<u>A</u>' task sequences have consistently been observed to result in worse performance on the third trial as compared to 'CB<u>A</u>' sequences. Classically, this cost has been attributed to selective inhibition of specific stimulus-response rules when switching away from the task on the first trial, such that on the third trial of the sequence this task is harder to access again. If such an inhibitory mechanism exists, why could it not be triggered

into play through the cancellation cues in the current study? Different points need to be considered here. First, the current inducer-diagnostic paradigm significantly differs from the task-switching paradigm. Hence, (a) in task switching, participants are aware that the currently irrelevant set of stimulus-response rules will become relevant again eventually; and (b) each set of stimulus-response mappings is applied more frequently such that episodic or procedural long-term memory may come into play. Despite the fact that in both paradigms different task sets are required to be maintained, then, caution is required in generalizing across these paradigms. Second, recent research demonstrates that the n-2 task repetition cost may largely reflect episodic retrieval effects rather than inhibitory processes (e.g., Grange et al., 2017). In this respect, the current findings may be cautiously taken to align with research on the n-2 task repetition cost in questioning an extensive role for top-down inhibition of task sets in setting the cognitive flexibility-stability balance (cf. Schmidt & Liefooghe, 2016; Schmidt, Liefooghe, & De Houwer, 2020a, 2020b). We hope that future studies will provide more dedicated explorations into the links between instruction following and task switching paradigms.

Overall, the current findings on instruction following align well with (and may inspire new research on) a general insight arising from the action control literature: The distinction between goal-directed intentions and habitual action may not be so clear-cut as traditionally believed (cf. Hommel & Wiers, 2017; Moors et al., 2017). On the one hand, even though instruction following enables flexible, goal-directed action, it is only to a limited extent under intentional control. Namely, whereas the act of encoding newly instructed stimulus-response mappings may be under intentional control (Wenke et al., 2009)[8], once implemented and installed in procedural working memory, the impact that these mappings have on cognition

[8] For a critical discussion of the paradigm used by Wenke et al. (2009), please refer to Liefooghe et al. (2012).

and performance is rather ballistic until they are replaced by new mappings. This impact cannot be easily undone by instructed changes with respect to the goal of the task (i.e., cancellation cues). On the other hand, whereas habitual action is classically assumed to evolve from repetitive behavior (e.g., Wood & Rünger, 2016), the ballistic and enduring impact on behavior of new instructions in working memory indicates at least an alternative route to habit-like behavior; one that – through its initiation at the level of working memory – is much closer linked to goal-directed processes than previously assumed. Indeed, this aligns well with recent work suggesting that (seemingly) habitual behavior in humans may relate mainly to dysfunctions in goal-directed behavior rather than overactive habit learning (e.g., de Wit et al., 2018).

Instructions in memory

Despite having evolved into rather separate research domains, working memory and action control are intricately related (e.g., Manohar et al., 2019; Nobre & Stokes, 2019; Wolff et al., 2017). Accordingly, next to a declarative system containing the representations of entities in the world (e.g., stimuli like objects and symbols) and the relationships between them, the working memory model by Oberauer (2009; 2010) postulates a procedural system that is responsible for the mental and motor operations performed on this information. Both systems are claimed to involve severe capacity limitations, and both may thus benefit from mechanisms that actively remove outdated information.

In the domain of declarative (working) memory, studies on (selective) directed forgetting indeed support an active removal mechanism operating on declarative content that has become irrelevant (Lewis-Peacock et al., 2018). Based on the tight commonalities (or even actual mechanical overlap) between declarative and procedural working memory processes (e.g., Barouillet et al., 2015; Oberauer et al., 2010; 2013), the clear support for active removal in declarative working memory (Lewis-Peacock et al., 2018) motivated our

search for a similar mechanism operating in procedural working memory. Indeed, actively removing irrelevant procedures would serve the cognitive agent as much as does removal of outdated declarative content. However, in contrast to the effectiveness of 'forget cues' in declarative working memory (i.e., directed forgetting; e.g., Aguirre et al., 2017; Delaney et al., 2009), 'cancellation cues' were here not effective in removing stimulus-response rules from procedural working memory (i.e., no directed dismantling). This does not provide support for the existence of an active removal mechanism at the level of procedural working memory, and thus argues against too strong versions of the claim that declarative and procedural working memory either share their resources and processes (Barouillet et al., 2015) or work according to common principles (Oberauer, 2009; 2010).

The finding that general performance improved after the cancellation cues is relevant mainly in demonstrating that participants processed the cancellation cues. But how can we understand such a speed-up in performance? One may argue that the speed-up indicates that the cancellation cues removed the inducer task stimulus-response rules from capacity-limited procedural working memory – thus freeing up resources for managing more effectively the (second set of) diagnostic task trials – while long-term memory traces kept generating the IBC effect. However, this would leave us with the question about how these newly instructed inducer task stimulus-response rules reached long-term memory before the cancellation cues were presented. Indeed, Meiran and colleagues (2012, p. 3) claim that stimulus and response codes are only linked together in more permanent format in cases that the instructed stimulus-response mappings are actually applied or practiced. One may argue that covert rehearsal or simulation of the inducer task mappings leads to representations in long-term memory (cf. Theeuwes et al., 2018) – but this seems especially far-fetched for the conditions in the current study in which cancellation cues followed immediately after the presentation of the inducer task mappings (i.e., 0 trials of the first diagnostic task).

Rather, we believe that the cancellation cues (even though processed) did not succeed in removing inducer task mappings from procedural working memory – unless actively replaced by new inducer task mappings (Experiment 4). We believe that the dual-task cost reflects the fact that initially (before the cancellation cue) the system was required to constantly monitor for the appearance of the imperative stimulus of the inducer task – the onset of which was (relatively) unpredictable within a run because the number of trials of the diagnostic task varied across runs (cf. Smith et al., 2007). After cancellation, the monitoring was no longer required, and this freed up resources to more effectively manage the ongoing diagnostic task. Monitoring processes are well-integrated with major action control theories (e.g., Botvinick et al., 2001), and monitoring has been considered an intrinsic part of the executive control of working memory (e.g., Vandierendonck et al., 1998).

The current findings will inform current and future models of working memory in general. For example, rather than assigning declarative and procedural representations (and the arbitrary bindings between them) to different systems such as in the conceptual model by Oberauer (2009; 2010), a recent computational model by Manohar et al. (2020) treats stimulus and response codes as highly equivalent nodes that can be bound to each other in any combination through changes in synaptic weights – depending on the task context. This model fits well with recent action control models that assume cognitive control to arise from learning of contextually appropriate associations between stimulus, response, and/or control representations (e.g., Abrahamse et al., 2016; Hommel et al., 2001; Egner, 2014). Whereas the model by Manohar (2020) may be well-suited to explain the observation that cancellation cues leave intact the IBC effect (as synaptic weights are not affected) unless the inducer task is actively replaced by another, we believe that this model cannot easily explain the speed-up in response times after the inducer task is canceled.

Finally, the inducer-diagnostic paradigm can be considered the (pure) working (or short-term) memory counterpart of the more general prospective memory tasks. Even though the current study, to the best of our knowledge, is the first to explore directed dismantling at the level of procedural working memory, our results are generally consistent with at least one prior finding from the prospective memory literature. Bugg and Scullin (2013) focused on the impact of no longer relevant task instructions, but the canceled task instructions had to be remembered over much longer periods of time. Specifically, like in the current study, Bugg and Scullin (2013) informed participants that a previously encoded stimulus-response rule was no longer relevant, after which they measured 'commission errors' (i.e., inappropriate execution) of this stimulus-response rule in a later stage of the task. They observed that participants found it hard to prevent such commission errors, especially if they never executed the now irrelevant stimulus-response rule. This finding aligns with the outcome of the current study. However, although an explicit message similarly stated that the prospective-memory instructions were no longer relevant in the study by Bugg and Scullin (2013), their study provided no measure to attest to the notion that participants effectively disengaged from the intention to apply the instructions. The drop in response times after the cancellation cues in the current study suggests such disengagement. More importantly, in their design, Bugg and Scullin (2013) inserted both a long delay (during which an otherwise irrelevant vocabulary task and a demographics form were administered) and a rather long series of alternative task trials between the encoding and forgetting instructions. This makes it hard to distinguish the effects of working memory from long-term memory effects; indeed, directed forgetting has been observed in declarative (Macleod, 1999) and procedural long-term memory (Schmidt et al., 2021) but may involve different mechanisms. The current study specifically targeted procedural working memory, where novel stimulus-response bindings are maintained (cf. the 'bridge'; Oberauer 2009; 2010), from the notion that active removal serves capacity

limitations (cf. Lewis-Peacock et al., 2018). Future studies should further close the gap between these literatures.

Overall, we can conclude that procedural representations formed on the basis of instructions are more persistent than may be assumed from the floating nature of working memory content. We believe that further progress on linking instruction following to general models of working memory may follow from computational simulations of our data (e.g., by models presented in Manohar et al., 2020 and Oberauer et al., 2013), and from neuroimaging studies that offer a window into (latent) instructed task representations (e.g., Bourguignon et al., 2018; González-García et al., 2017; Muhle-Karbe et al., 2017; Palenciano et al., 2019; Ruge at al., 2019). Moreover, explicit links to working memory input gating mechanisms (Chatham & Badre, 2015) may help to further explain instruction following in a fully mechanistic manner, accounting for the intitial, intention-driven set up of stimulus-response rules in working memory.

Limitations and directions for future studies

The current study aimed to explore the directed dismantling of stimulus-response rules in capacity-limited procedural working memory. In order to study procedural working memory proper – and to prevent an impact of more permanent memory traces of the stimulus-response rules – we employed the inducer-diagnostic paradigm. This task allows for exploring the impact on ongoing behavior of prepared but (formally) not yet executed instructions, opening a relatively clean window into the operations of procedural working memory. Specifically, in the current study we opted for instructed stimulus-response rules that were composed of arbitrary bindings between semantic (stimulus) representations, and left or right response representations. We cannot exclude the possibility that reading the words ‘left’ and ‘right’ during encoding of the inducer task instructions resulted in corresponding sub-threshold motor activation via automatic priming (e.g., Bundt et al., 2015), thus implying potential

subthreshold execution-like processing for each inducer task stimulus-response rule in the instruction phase. On the one hand, this does not threaten current conclusions: The finding remains that reading instructions necessary to perform a future task quickly results in the formation of a "prepared reflex" that is, as we now show, hard to cancel once installed at the level of procedural working memory. Without time and opportunity for more extended practice, we fail to see how mere sub-threshold motor activation would promote stimulus-response representations beyond the realm of capacity-limited procedural working memory (e.g., activated long-term memory traces; Meiran et al., 2012). Indeed, if our current design with left- versus right-hand responses would have somehow allowed for more permanent memory traces of the initial instructions in the first encoding phase, then why is a simple (and instant, in the case of 0 trials for the first diagnostic task) replacement by a new set of instructions in Experiment 4 sufficient to attenuate the impact of the initial set of instructions? If these would have already reached long-term memory, they would be expected to still induce an IBC effect after having been replaced by new instructions (e.g., Logan, 1988). On the other hand, the extent to which both the prepared reflex and its insensitivity to cancellation cues are related to the degree with which the stimulus-response rule can be automatically activated (or simulated) is an interesting research question for future studies that will help inform to inform the level of generalizability of the current proof-of-principle study. For example, future studies may replicate current findings using response sets that have been argued to be less likely to induced sub-threshold activation, such as bimanual, finger-specific response options (Gonzalez-Garcia et al., 2019; but see, Formica et al., 2021).

Finally, as noted above, it may be that more powerful cancellation manipulations would succeed in eliminating the automatic impact of task instructions. Indeed, a study by Whitehead and Egner (2018b) manipulated the proportion congruency across diagnostic trials in a different type of inducer-diagnostic task, and observed that under some conditions it is

possible to partially block the automatic impact of procedural working memory representations. Follow-up studies should try to 'up the stakes' for more complete cancellation in the current study design, as participants in the current study may not have been pushed enough to truly engage in full dismantling. This may be done in several ways. Participants may receive overall performance-contingent reward in order to enhance motivation. Alternatively, performance can be associated with higher rewards for runs with cancellation cues than for other types of runs. Previous work showed the reflexive nature of instruction following (Liefooghe et al., 2012), and the current study took the next step in showing that this even occurs when people have demonstrably disengaged from these instructions when they became irrelevant. Reinforcing cancellation policies could provide an even more stringent test of the true automatic nature of applying instructions.

To conclude, the current study provides novel insight into instruction following, demonstrating that it is difficult to dismantle new instructions maintained in procedural working memory once they become outdated. We suggest that the preparation of new instructions can result in a representation that has the potential to automatically impact task performance through reflexes, but that this representation is ballistic in nature as no mechanisms are available to counteract the effect of prepared instructions. Thus far, it seems that only encoding and preparing new instructions has the potential to override older, no-longer-relevant instructions. Hence, the only way for our student driver from the introduction to not suffer from unexpected changes in the traffic situation, would be to immediately receive a set of adjusted instructions from the expert instructor.

Context of the research

"Please write a context paragraph!" Instruction following is critical for optimal learning and performance in our society, rendering it a research topic with high applied relevance. Even though instruction following has grown into a separate research domain over the last decade,

it cannot be studied in isolation. Understanding its underlying mechanisms requires an integration of insights from various more overarching research domains. The current paper is the result of a collaboration of researchers working across the domains of instruction following, action control/cognitive flexibility, and working memory. Indeed, instructions allow for flexible action control without the need for practice, through configuring procedural working memory with stimulus-response rules that are initially encoded in declarative working memory. This requires and motivates crossing these literatures. Here we explored if a forgetting mechanism studied mainly in the domain of declarative working memory, also operates on new instructions in procedural working memory. This type of research forces both empirical and theoretical integration across research domains. In a next step, we hope to engage language researchers to extent our understanding about how linguistic input sets up declarative and procedural working memory content, and about what linguistic features affect instruction following in which way.

References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693-728.

Aguirre, C., Gómez-Ariza, C. J., Andrés, P., Mazzoni, G., & Bajo, M. T. (2017). Exploring mechanisms of selective directed forgetting. *Frontiers in Psychology*, 8:316.

Baddeley, A. (1998). The central executive: A concept and some misconceptions. *Journal of the International Neuropsychological Society*, *4*(5), 523-526.

Barrouillet, P., Corbin, L., Dagry, I., & Camos, V. (2015). An empirical test of the independence between declarative and procedural working memory in Oberauer's (2009) theory. *Psychonomic Bulletin & Review*, *22*(4), 1035-1040.

Bourguignon, N. J., Braem, S., Hartstra, E., De Houwer, J., & Brass, M. (2018). Encoding of novel verbal instructions for prospective action in the lateral prefrontal cortex: evidence from univariate and multivariate functional magnetic resonance imaging analysis. *Journal of cognitive neuroscience*, *30*(8), 1170-1184.

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

Braem, S., Deltomme, B., & Liefooghe, B. (2019). The instruction-based congruency effect predicts task execution efficiency: Evidence from inter-and intra-individual differences. *Memory & Cognition*, *47*(8), 1582-1591.

Braem, S., Liefooghe, B., De Houwer, J., Brass, M., & Abrahamse E.L. (2017). There are limits to the effects of task instructions: Making the automatic effects of task instructions context-specific takes practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 43,* 394-403.

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews, 81,* 16-28.

Bugg, J. M., & Scullin, M. K. (2013). Controlling intentions: The surprising ease of stopping after going relative to stopping after never having gone. *Psychological Science*, *24*(12), 2463-2471.

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: 539.

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, *1*, 23-31.

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. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1140-1154.

Cole, M. W., Braver, T. S., & Meiran, N. (2017). The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. *Neuroscience & Biobehavioral Reviews, 81,* 4-15.

Dames, H., & Oberauer, K. (2021, October 18). Directed-forgetting in working memory. https://doi.org/10.31234/osf.io/93cru.

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*(4), 659-666.

De Rosario-Martinez, H., Fox, J., Team, R. C., & De Rosario-Martinez, M. H. (2015). Package 'phia'. *CRAN repository. Retrieved*, *1*, 2015.

Delaney, P. F., Nghiem, K. N., & Waldum, E. R. (2009). Short article: The selective directed forgetting effect: Can people forget only part of a text? *Quarterly Journal of Experimental Psychology*, *62*(8), 1542-1550.

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488-499.

Ecker, U.K., Lewandowsky, S. & Oberauer, K. (2014a). Removal of information from working memory: A specific updating process. *Journal of Memory & Language*, *74*, 77–90.

Ecker, U.K., Oberauer, K. & Lewandowsky, S. (2014b). Working memory updating involves item-specific removal. *Journal of Memory & Language*, *74*, 1–15.

Egner, T. (2014). Creatures of habit (and control): a multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology*, 5:1247.

Everaert, T., Theeuwes, M., Liefooghe, B., & De Houwer, J. (2014). Automatic motor activation by mere instruction. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 1300-1309.

Exner, S. (1879). Physiologie der Grosshirnrinde. *Handbuch der physiologie*, *2*(part 2), 189-350.

Festini, S. B., & Reuter-Lorenz, P. A. (2013). The short-and long-term consequences of directed forgetting in a working memory task. *Memory*, *21*(7), 763-777.

Festini, S. B., & Reuter-Lorenz, P. A. (2014). Cognitive control of familiarity: Directed forgetting reduces proactive interference in working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 78-89.

Formica, S., González-García, C., Senoussi, M., & Brass, M. (2021). Neural oscillations track the maintenance and proceduralization of novel instructions. *NeuroImage*, *232*, 117870.

Frings, C., Hommel, B., Koch, I., Rothermund, K., Dignath, D., Giesen, C., ... & Philipp, A. (2020). Binding and retrieval in action control (BRAC). *Trends in Cognitive Sciences*, *24*(5), 375-387.

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., & Ruz, M. (2017). Encoding, preparation and implementation of novel complex verbal instructions. *Neuroimage*, *148*, 264-273.

Grange, J. A., Kowalczyk, A. W., & O'Loughlin, R. (2017). The effect of episodic retrieval on inhibition in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1568-1583.

Hommel, B. (2009). Action control according to TEC (theory of event coding). *Psychological Research*, *73*(4), 512-526.

Hommel, B. (2015). Between persistence and flexibility: The Yin and Yang of action control. In *Advances in motivation science* (Vol. 2, pp. 33-67). Elsevier.

Hommel, B., & Wiers, R. W. (2017). Towards a unitary approach to human action control. *Trends in Cognitive Sciences*, *21*(12), 940-949.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, *42*(3), 643-650.

Koch, I., Gade, M., Schuch, S., & Philipp, A. M. (2010). The role of inhibition in task switching: A review. *Psychonomic Bulletin & Review*, *17*(1), 1-14.

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking—An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557.

Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 33-44.

Liefooghe, B. (2021, November 8). Tenacious Instructions. Retrieved from osf.io/au7pv.

Liefooghe, B. , De Houwer, J., & Wenke, D. (2013). Instruction-based response activation depends on task preparation. *Psychonomic Bulletin & Review, 20,* 481-487.

Liefooghe, B., & De Houwer, J. (2018). Automatic effects of instructions do not require the intention to execute these instructions. *Journal of Cognitive Psychology*, *30*(1), 108-121.

Liefooghe, B., & Verbruggen, F. (2019). On the assimilation of instructions: Stimulus-response associations are implemented but not stimulus-task associations. *Journal of Cognition*, 2:20.

Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38,* 1325-1335.

Liefooghe, B., Degryse, J., & Theeuwes, M. (2016). Automatic effects of No-Go instructions. *Canadian Journal of Experimental Psychology, 70,* 232-241.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Macleod C.M. (1999). The item and list methods of directed forgetting: Test differences and the role of demand characteristics. *Psychonomic Bulletin & Review*, *6*, 123–129.

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, *101*, 1-12.

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*(2), 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). Reflexive activation of newly instructed stimulus–response rules: Evidence from lateralized readiness potentials in no-go trials. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 365-373.

Meiran, N., Pereg, M., Kessler, Y., Cole, M.W., & Braver, T.S. (2015b). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 768.

Meiran, N., Liefooghe, B., & De Houwer, J. (2017). Powerful instructions: Automaticity without practice. *Current Directions in Psychological Science, 6,* 509-514.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 100-110.

Monsell, S., & Driver, J. (2000). Banishing the control homunculus. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 3–32). Cambridge, MA: MIT Press.

Moors, A., Boddez, Y., & De Houwer, J. (2017). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, *9*(4), 310-318.

Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2017). Neural coding for instruction-based task sets in human frontoparietal and visual cortex. *Cerebral Cortex*, *27*(3), 1891-1905.

Nobre, A. C., & Stokes, M. G. (2019). Premembering experience: A hierarchy of time-scales for proactive attention. *Neuron*, *104*(1), 132-146.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). New York: Plenum.

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? *Psychologica Belgica*, *50*, 277-308.

Oberauer, K., Souza, A. S., Druey, M. D., & Gade, M. (2013). Analogous mechanisms of selection and updating in declarative and procedural working memory: Experiments and a computational model. *Cognitive Psychology*, *66*(2), 157-211.

Palenciano, A. F., González-García, C., Arco, J. E., Pessoa, L., & Ruz, M. (2019). Representational organization of novel task sets during proactive encoding. *Journal of Neuroscience*, *39*(42), 8386-8397.

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601-605.

Rastle, K., Lally, C., Davis, M. H., & Taylor, J. S. H. (2021). The dramatic impact of explicit instruction on learning to read in a new writing system. *Psychological Science*, 32(4), 471-484.

Ruge, H., Schäfer, T. A., Zwosta, K., Mohr, H., & Wolfensteller, U. (2019). Neural representation of newly instructed rule identities during early implementation trials. *eLife*, *8*.

Schmidt, M., Frings, C., & Tempel, T. (2021). Selective directed forgetting of motor sequences. *Acta Psychologica*, 218: 103352.

Schmidt, J. R., De Houwer, J., & Rothermund, K. (2016). The Parallel Episodic Processing (PEP) model 2.0: A single computational model of stimulus-response binding, contingency learning, power curves, and mixing costs. *Cognitive Psychology*, *91*, 82-108.

Schmidt, J. R., Liefooghe, B., & De Houwer, J. (2020a). An episodic model of task switching effects: Erasing the homunculus from memory. *Journal of Cognition*, 3:22.

Schmidt, J. R., Liefooghe, B., & De Houwer, J. (2020b). Erasing the homunculus as an ongoing mission: A reply to the commentaries. *Journal of Cognition*, 3:28.

Schmidt, J. R., & Liefooghe, B. (2016). Feature integration and task switching: Diminished switch costs after controlling for stimulus, response, and cue repetitions. *PLOS ONE*, 11: e0151188.

Schuch, S., & Koch, I. (2003). The role of response selection for inhibition of task sets in task shifting. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 92-105.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*.

Smith, R. E., Hunt, R. R., McVay, J. C., & McConnell, M. D. (2007). The cost of event-based prospective memory: salient target events. *Journal of Experimental Psychology: Learning,*

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: AC library for programming cognitive experiments on the MS Windows platform. *Behavior research methods*, *38*(2), 280-286.

Theeuwes, M., De Houwer, J., Eder, A., & Liefooghe, B. (2015). Congruency effects on the basis of instructed response-effect contingencies. *Acta psychologica*, *158*, 43-50.

Theeuwes, M., Liefooghe, B., & De Houwer, J. (2014). Eliminating the Simon effect by instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1470.

Theeuwes, M., Liefooghe, B., De Schryver, M., & De Houwer, J. (2018). The role of motor imagery in learning via instruction. *Acta Psychologica, 184,* 110-123.

Townsend, J.T., & Ashby, F.G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), Cognitive theory. Vol. 3. (pp. 200-239). Hillsdale, N.J.: Erlbaum.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Vandierendonck, A., Liefooghe, B. , & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference. *Psychological Bulletin, 136,* 601-626.

Vandierendonck, A., De Vooght, G., & Van der Goten, K. (1998). Interfering with the central executive by means of a random interval repetition task. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 197-218.

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*(11), 418-424.

Verbruggen, F., McLaren, I. P., & Chambers, C. D. (2014). Banishing the control homunculi in studies of action control and behavior change. *Perspectives on Psychological Science*, *9*(5), 497-524.

Wenke, D., Gaschler, R., & Nattkemper, D. (2007). Instruction-induced feature binding. *Psychological Research*, *71*, 92-106.

Wenke, D., Gaschler, R., Nattkemper, D., & Frensch, P. A. (2009). Strategic influences on implementing instructions for future actions. *Psychological Research, 73*, 587-601.

Whitehead, P. S., & Egner, T. (2018a). Cognitive control over prospective task-set interference. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 741.

Whitehead, P. S., & Egner, T. (2018b). Frequency of prospective use modulates instructed task-set interference. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(12), 1970-1980.

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., . . . Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General, 147*(7), 1043-1065.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864-871.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, *67*, 289-314.

Tables

Table 1. Cell means and corresponding standard errors of the second diagnostic task in Experiment 1. The following abbreviations are used: Applied inducer stimulus (Applied), Instructed inducer stimulus (Instructed).

		Congruency			
		Congruent		Incongruent	
		M	*SE*	*M*	*SE*
RTs					
	Applied	532	12	536	12
	Instructed	530	11	537	12
PCs					
	Applied	0.95	0.01	0.93	0.01
	Instructed	0.95	0.01	0.93	0.01

Table 2. Cell means and corresponding standard errors of Experiment 2. The following abbreviations are used: No Inducer Task (NIT), Inhibit Inducer Response (IIR), Control condition (C).

		Congruency			
		Congruent		Incongruent	
	Condition	M	SE	M	SE
Filler runs					
RTs	NIT	525	19	534	20
	IIR	556	19	573	20
	C	534	19	544	20
PCs	NIT	0.95	0.01	0.94	0.01
	IIR	0.96	0.01	0.94	0.01
	C	0.96	0.01	0.95	0.01
Cued Trials Target Runs					
RTs	NIT	497	18	510	20
	IIR	528	19	541	20
	C	535	18	541	20
PCs	NIT	0.94	0.01	0.93	0.01
	IIR	0.96	0.01	0.95	0.01
	C	0.96	0.01	0.95	0.01
Corresponding Trials Filler Runs					
RTs	NIT	522	18	533	18
	IIR	552	19	563	19
	C	527	18	535	18
PCs	NIT	0.95	0.01	0.93	0.01
	IIR	0.96	0.01	0.94	0.01
	C	0.96	0.01	0.94	0.01

Table 3. Cell means and corresponding standard errors of the second diagnostic task in Experiment 3.

		Run Type	Congruency			
			Congruent		Incongruent	
			M	SE	M	SE
RTs	Absent	Complete	546	7	560	7
		Cancel	539	8	554	8
		Proceed	556	8	566	8
	Present	Complete	552	7	561	8
		Cancel	550	8	555	8
		Proceed	557	8	574	9
PCs	Absent	Complete	0.93	0.01	0.90	0.01
		Cancel	0.94	0.01	0.91	0.01
		Proceed	0.94	0.01	0.91	0.01
	Present	Complete	0.94	0.01	0.90	0.01
		Cancel	0.94	0.01	0.91	0.01
		Proceed	0.94	0.01	0.90	0.01

Table 4. Cell means and corresponding standard errors of the second diagnostic task in Experiment 4.

		Run Type	Congruency			
			Congruent		Incongruent	
			M	SE	M	SE
RTs	Absent	Proceed	574	9	582	11
		Replace	585	12	589	11
	Present	Proceed	573	10	590	11
		Replace	581	10	583	10
PCs	Absent	Proceed	0.92	0.01	0.89	0.01
		Replace	0.92	0.01	0.91	0.01
	Present	Proceed	0.93	0.01	0.90	0.01
		Replace	0.91	0.01	0.91	0.01

Figure Captions

Figure 1. Schematic outline of the inducer-diagnostic paradigm. In the inducer-diagnostic paradigm, participants are presented with a series of runs that each exists of various trials (multiple runs together make up for a block). Across all runs, the diagnostic task remains the same: Participants decide whether a stimulus is presented in italic or upright (e.g., upright, press left; italic, press right). At the start of each run, two new arbitrary stimulus-response (S-R) mappings for the inducer task are introduced in a declarative format (i.e., e.g., If "cat", press left; if "dog", press right). Before participants can execute these mappings on a probe trial of the inducer task presented towards the end of the run (e.g., the word 'dog' presented in green), they are first presented with a number of trials of the diagnostic task. The main target of the inducer-diagnostic paradigm is to explore the impact of the inducer task stimulus-response mappings (that are kept in mind for later use) on performance during the diagnostic task trials. Specifically, the combination of inducer and diagnostic stimulus-response mappings results in a congruency variable: Congruent trials require a diagnostic task response that matches with the stimulus-response mappings of the inducer task (e.g., "cat" presented upright or "dog" presented in italic), whereas incongruent trials require a diagnostic task response that mismatches with the stimulus-response mappings of the inducer task (e.g., "cat" presented in italic or "dog" presented upright). Performance is typically better on congruent as compared to incongruent trials, and this is referred to as the Instruction-Based Congruency (IBC) effect.

Figure 2. Schematic outline of the different run types used in Experiment 1 and their corresponding time parameters. The runs were separated by a 500ms interval. The critical run type (Target Runs) allows for testing the impact on the second diagnostic task of a specific inducer task S-R mapping that was not executed for the completion of the inducer task trials.

Figure 3. Mean RTs of Experiment 1 as a function of Congruency (Congruent, Incongruent), Stimulus Type (Applied, Unapplied), and Diagnostic Task (First, Second). Error bars denote standard errors.

Figure 4. Schematic outline of the different run types used in Experiment 2 and their corresponding time parameters. The runs were separated by a 500ms interval. The critical run type (Target Runs) allows for testing the impact on the post-cue diagnostic task of inducer task S-R mapping that were cancelled before ever being executed.

Figure 5. Mean RTs of Experiment 2 as a function of Condition (Inhibit inducer response, No inducer task, Control), Run Type (Filler Run, Target Run) and Congruency (Congruent, Incongruent). Error bars denote standard errors.

Figure 6. Schematic outline of the different run types used in Experiment 3 and their corresponding time parameters. The runs were separated by a 1500ms interval in which 5 exclamation marks (‘!!!!!’) were presented in the screen center.

Figure 7. Mean RTs of Experiment 3 as a function of Congruency, Run Type, and First Diagnostic Task. Error bars denote standard errors.

Figure 8. Schematic outline of the different run types used in Experiment 4 and their corresponding time parameters. The runs were separated by a 1500ms interval in which 5 exclamation marks (‘!!!!!’) were presented in the screen center.

Figure 9. Mean RTs of Experiment 4 as a function of Congruency, Run Type, and First Diagnostic Task. Error bars denote standard errors.

Figures

Figure 1.

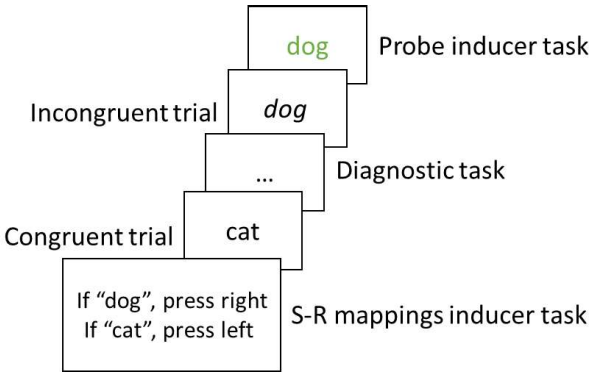


Figure 2.

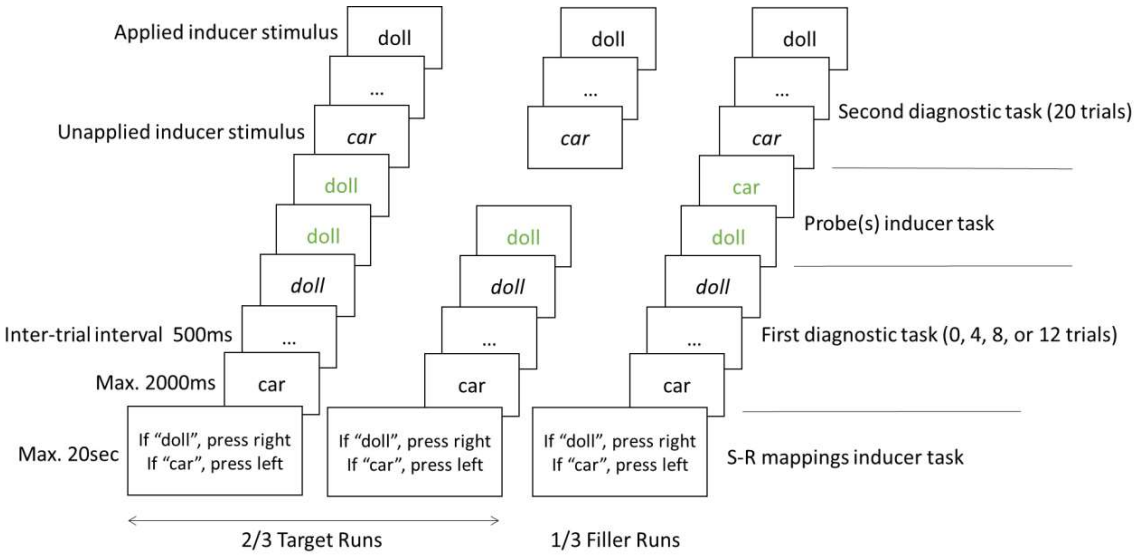


Figure 3.

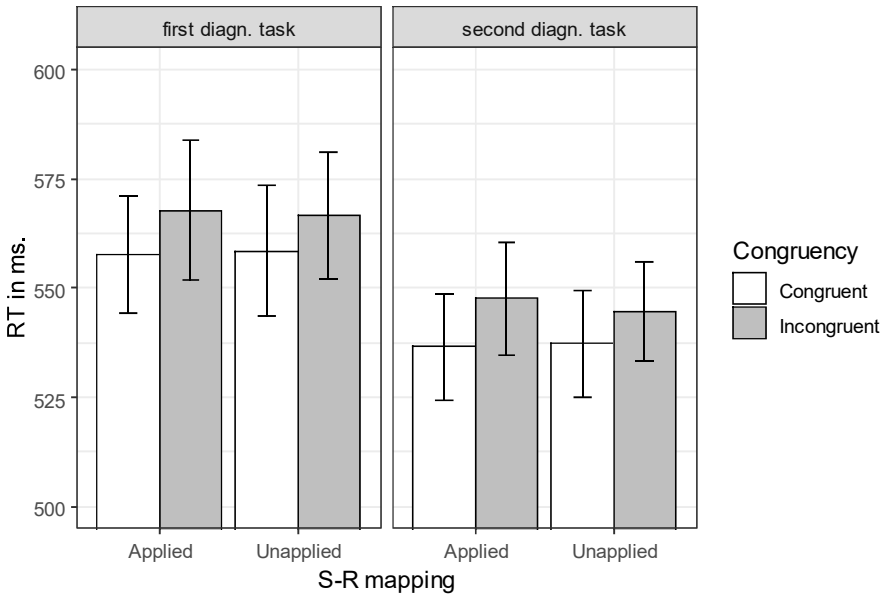


Figure 4.

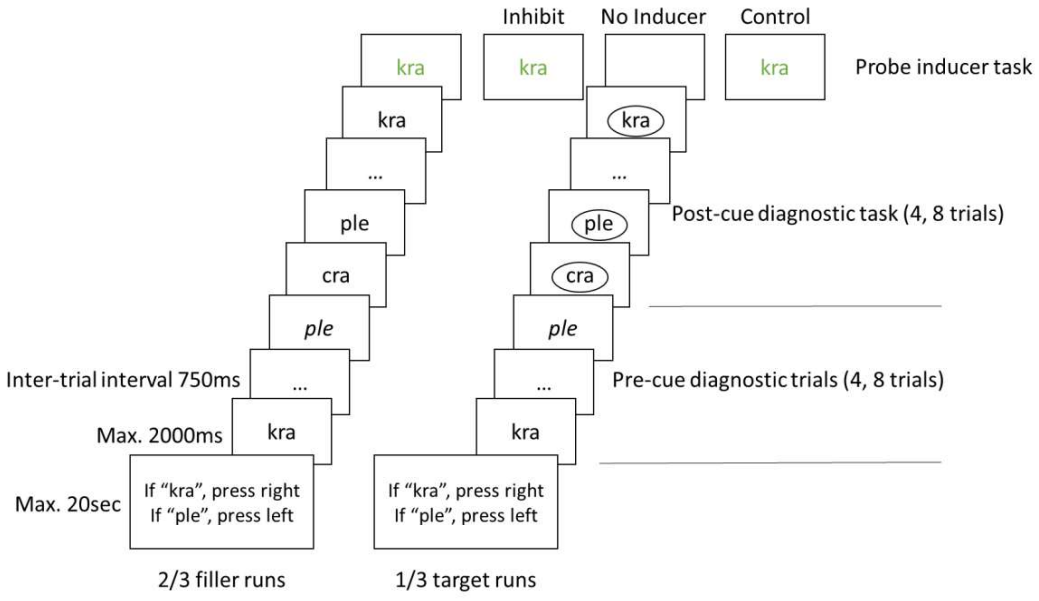


Figure 5.

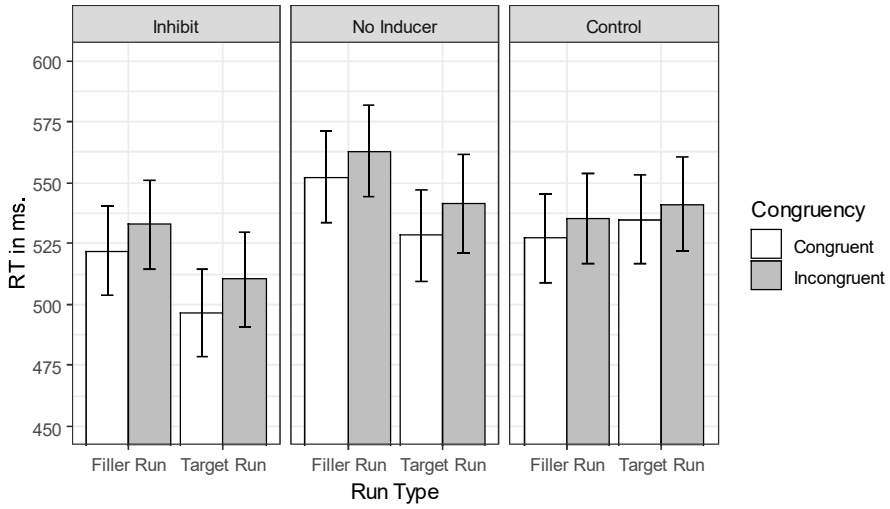


Figure 6.

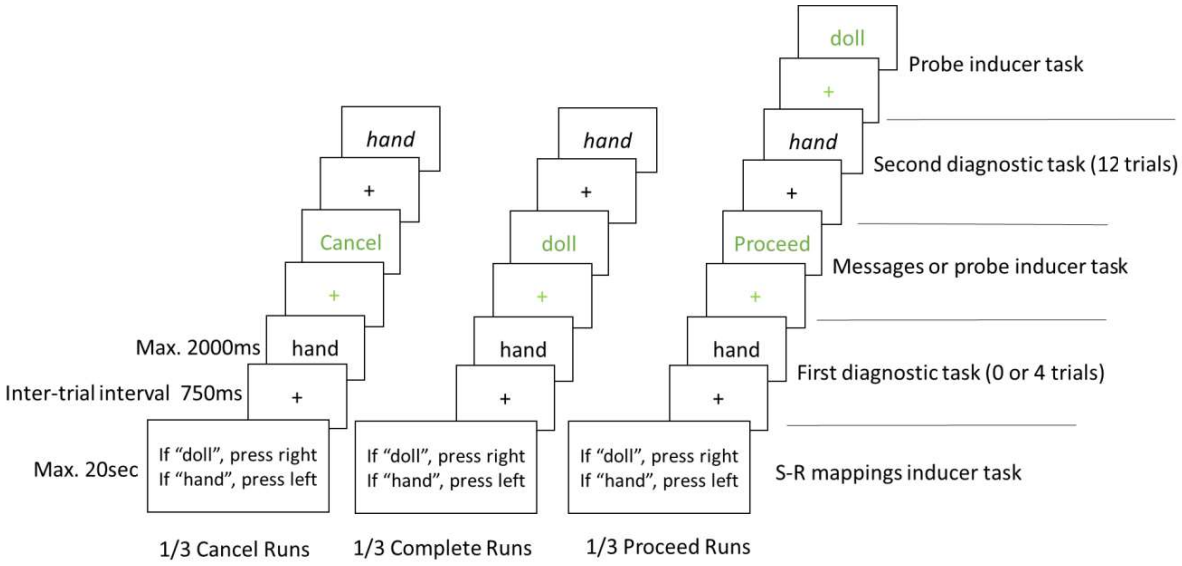


Figure 7.

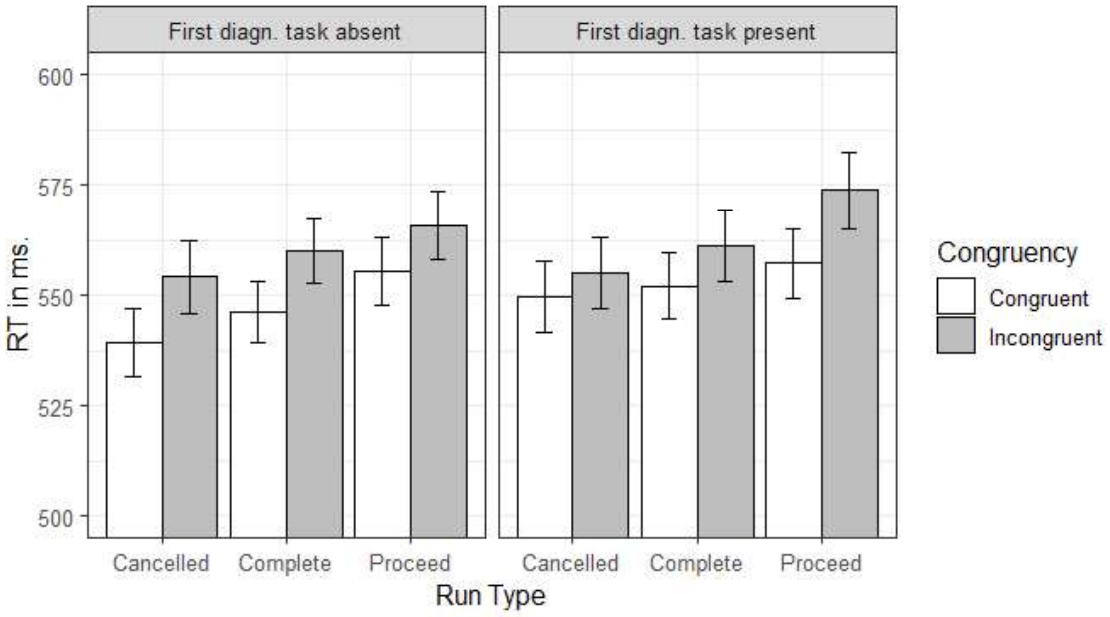


Figure 8.

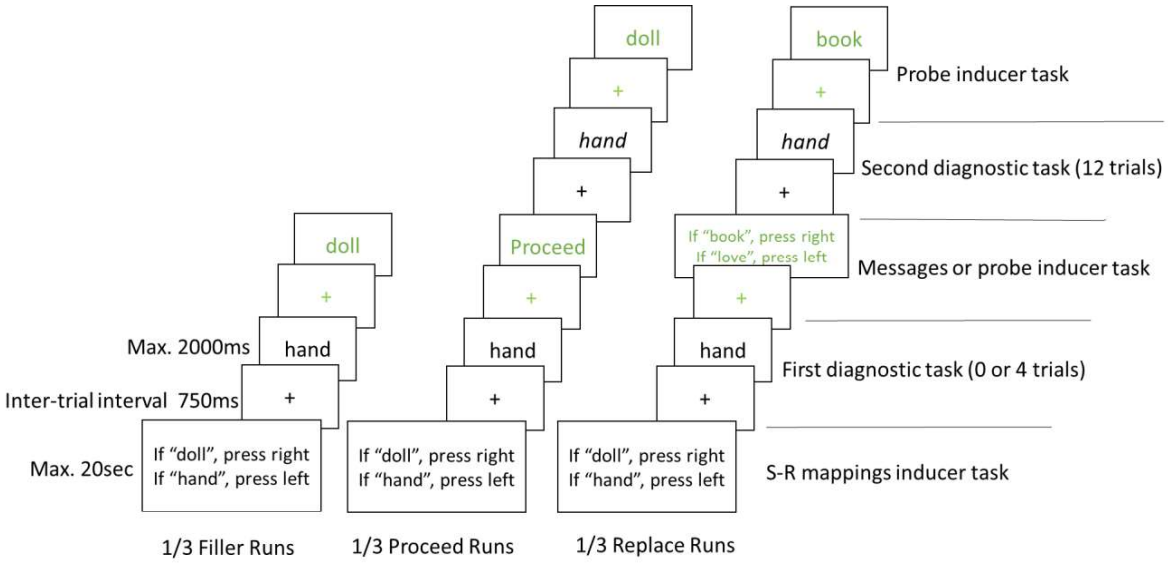
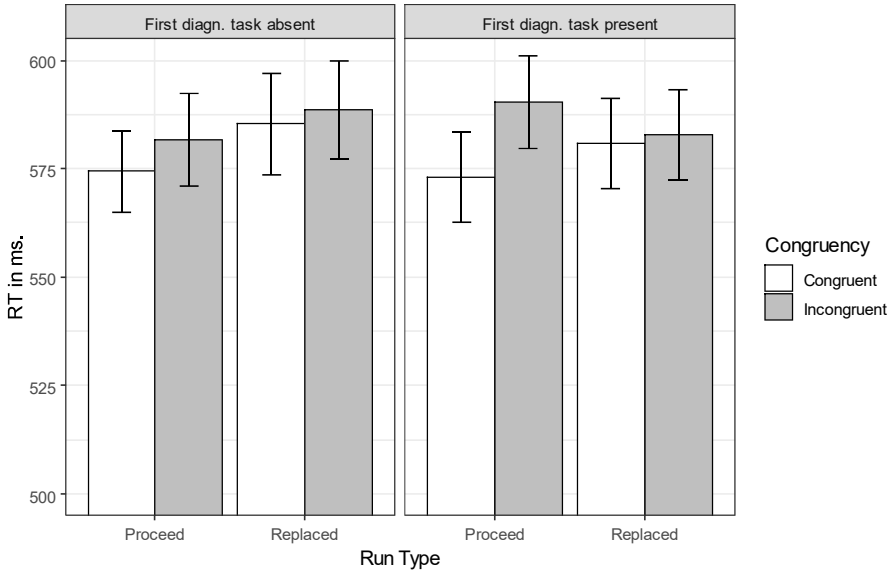


Figure 9.


First diagn. task absent
First diagn. task present
600
575
550
525
500
RT in ms.
Proceed
Replaced
Proceed
Replaced
Run Type
Congruency
Congruent
Incongruent