Memory & Cognition (2019) 47:1582–1591
https://doi.org/10.3758/s13421-019-00951-3

The instruction-based congruency effect predicts task execution efficiency: Evidence from inter- and intra-individual differences

Senne Braem [1,2] · Berre Deltomme [2] · Baptist Liefooghe [3]

Published online: 18 June 2019

Abstract
In contrast to traditional conflict paradigms, which measure interference from (over)trained associations, recent paradigms have been introduced that investigate automatic interference from newly instructed, but never executed, associations. In these prospective-instruction paradigms, participants receive new task instructions (e.g., if cat press left, if dog press right), but before they have to apply the instructions, they are first presented with another task that measures the automatic interference from the instructed task information. The resulting instruction-based congruency (IBC) effect is assumed to reflect the strength with which instructions are encoded and maintained in view of their future application. If this assumption holds true, the IBC effect should be inversely related to the speed with which the task instructions are eventually executed. To test this hypothesis, we administered a prospective-instruction paradigm to a large sample of 184 participants and observed a negative correlation between the IBC effect and mean reaction time on the instructed task. Similarly, an analysis looking at within-subject variations in the IBC effect and instructed task reaction times showed the same negative relation. Finally, we also present additional analyses suggesting this effect is independent from standard (experience-based) interference effects, and report explorative analyses that tested possible correlations with personality trait questionaires. Together, these findings confirm a key assumption of the IBC effect in prospective-instruction paradigms, and further support the use of this paradigm in instruction research.

Keywords Learning via instructions · Cognitive control · Task sets · Automaticity · Automatic processing · Task switching

Introduction

Learning via verbal instructions is a unique human ability through which new behavior can emerge almost instantaneously, without needing trial-and-error learning (Deacon, 1997). For example, we use instructions to learn how to handle our new smartphone, to find our way in a new city, or to prepare a new recipe. Despite the primordial importance of learning via instructions, our understanding of its underlying (neuro)cognitive dynamics is still in its infancy. Nevertheless, in recent years considerable advancements have been made in an exponentially growing number of studies, which focus on the *automatic effects of instructions* (e.g., Cohen-Kdoshay & Meiran, 2007, 2009; Cole, Laurent, & Stocco, 2013; De Houwer, Beckers, Vandorpe, & Custers, 2005; Liefooghe, Wenke, & De Houwer, 2012; Meiran, Pereg, Givon, Danielli, & Shahar, 2016; Wenke, De Houwer, De Winne, & Liefooghe, 2015).

For some time it has been assumed that overt practice or training is the sole pathway to automaticity (Anderson, 1992; Laberge & Samuels, 1974; Logan, 1985, 1988; Schneider & Shiffrin, 1977). Such automaticity has been extensively investigated using conflict paradigms, such as the Stroop (1935) or the Simon (1969) task, in which the automatic interference of over-trained stimulus-response (S-R) associations is measured (e.g., word reading in the Stroop task or location categorization in the Simon task). Recent studies on instructions, however, demonstrated that such conflict effects can also be

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13421-019-00951-3) contains supplementary material, which is available to authorized users.

✉ Senne Braem
Senne.Braem@vub.be

[1] Department of Experimental and Applied Psychology, Vrije Universiteit Brussel, Pleinlaan 2, B – 1050 Brussels, Belgium

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

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

obtained for merely instructed S-R associations, which were not even executed once (e.g. Braem, Liefooghe, De Houwer, Brass, & Abrahamse, 2017; Cohen-Kdoshay & Meiran, 2007, 2009; De Houwer et al., 2005; Everaert, Theeuwes, Liefooghe, & De Houwer, 2014; Liefooghe, De Houwer, & Wenke, 2013; Liefooghe et al., 2012; Meiran & Cohen-Kdoshay, 2012; Meiran, Pereg, Kessler, Cole, & Braver, 2015a, 2015b; Wenke et al., 2015; Whitehead & Egner, 2018a, 2018b). To this end, *Prospective-Instruction* (PI) paradigms were used in which a set of S-R mappings is instructed, which serve an inducer task that has to be executed in the near future. In the time period between the instruction phase and the actual application phase, automatic effects of these mappings are measured in a diagnostic task (e.g., Liefooghe et al., 2012, 2013; Meiran, Cole, & Braver, 2012; Meiran et al., 2015a; for a review, see Meiran, Liefooghe, & De Houwer, 2017). Two PI paradigms have been especially helpful in understanding automatic effects of instructions.

First, in the PI paradigm by Liefooghe and colleagues (2012; see also Liefooghe et al., 2013), participants are presented with multiple runs of trials, where each run begins with the instruction of two novel inducer mappings (e.g., if cat press left; if dog press right) that they have to perform by the end of the run. In between the instruction and application of the inducer mappings, participants are presented with a series of diagnostic trials in which the inducer stimuli (i.e., the word "cat" or "dog") are printed either upright or italic, and participants have to press the left or right button for upright or italic (e.g., if upright press left; if italic press right). This way, incongruent (e.g., "cat" printed in italic) or congruent (e.g., "dog" printed in italic) trials can be created that require either a different or the same response on the inducer and diagnostic task, respectively. The automatic effects of instructions are quantified by the difference in performance between congruent and incongruent diagnostic trials (Instruction-Based Congruency or IBC effect).

Second, a similar PI task was introduced by Meiran and colleagues (2015a). In this task, participants are instructed with two new inducer mappings on each run, which have to be applied when stimuli are presented in a green square. This is similar to the PI paradigm by Liefooghe and colleagues, but the main difference lies in the diagnostic task that is used. Specifically, during the diagnostic phase inducer stimuli are presented in a red square. Participants are asked to press a single response key (i.e., the NEXT response) in order to move on to the next trial and do so until a green square appears. The crucial manipulation is that the NEXT response is part of one of the inducer mappings. Based on this response overlap and the presence of inducer stimuli in the diagnostic phase, Meiran et al. (2015a, 2015b) also observed robust IBC effects.

Whereas more common conflict effects (e.g., Stroop, Simon) are thought to reflect the retrieval of associations from long-term memory (e.g., Laberge & Samuels, 1974), IBC effects supposedly originate from functional associations actively maintained in working memory (Brass, Liefooghe, Braem, & De Houwer, 2017), which are implemented on the basis of verbal instructions (e.g., Liefooghe et al., 2012; Meiran et al., 2015a). As a consequence, instructions that are represented more strongly in working memory should induce a stronger IBC effect. Similarly, instructions that are represented more strongly in working memory should result in faster task execution (Monsell, 1978; Sternberg, 1969). For example, the stronger you maintain the instruction "press left when I see the word cat" in working memory, the faster you will respond left upon seeing the word "cat". Following the above two ideas, the hypothesis follows almost naturally that shorter reaction times (RTs) on the inducer task (where the instructions have to be executed) should be associated with larger IBC effects. Interestingly, this was also the reasoning and approach of a recent correlation study by Meiran and colleagues (2016), who, surprisingly, observed the opposite pattern than would be expected a finding we will further discuss below. Here, we set out to test this basic assumption using the PI paradigm of Liefooghe and colleagues (2012, 2013).

A second aim was to systematically compare a participant's performance on this IBC task and a closely matched paradigm that measures interference from experienced or overtrained association, as in the Simon or Stroop task. Therefore, a second part of the experiment consisted of the same task as the diagnostic task (i.e., categorizing the font type of a word with left vs. right responses), but the words were now "left" or "right," instead of the instructed stimuli. The content of the words was still irrelevant, and hence induced a congruency effect when the meaning of the word did not match with the response side, which we will refer to as a semantic Simon effect (De Houwer, 1998). This allowed us to assess a closely matched congruency effect that directly measured the interference from the experienced associations with the words left and right, versus the newly instructed associations as in the IBC task. Our hypothesis was that if, indeed, the IBC effect measures interference from a different, independent source (e.g., procedural working memory) than traditional interference effects, we should not see a correlation between the two effects.

A third and final, more exploratory, goal of this study was to examine the relation between the IBC effect and questionnaire measures of personality traits that could be hypothesized to correlate with this effect. Importantly, this was not the main goal of our study, but, when testing such a large sample, we deemed the inclusion of these questionaires a very cost-effective way to get first insights into whether or not these traits could predict the degree to which people implement instructions. Detecting these relations at this stage could help motivate and inform future research endeavors along those lines. Specifically, we distributed a suggestibility questionnaire (Kotov, Bellman, & Watson, 2004), which we

speculated to relate positively to the implementation of new instructions. Highly suggestible people might be more sensitive to new instructions, and therefore have stronger instructed task representations, which should result in larger IBC effects. Next, we also assessed two clinical questionaires: Beck's Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and the Autism Quotient questionnaire (Baron-Cohen, Hoekstra, Knickmeyer, & Wheelwright, 2006). Depression has been related to cognitive rigidity (Meiran, Diamond, Toder, & Nemets, 2011) and a lesser motivation to engage in effortful control processes (Grahek, Everaert, Krebs, & Koster, 2018), and autism spectrum disorder to an insistence on sameness and repetitive behaviors (Gabriels, Cuccaro, Hill, Ivers, & Goldson, 2005) as well as language and communication problems (Mundy, Sigman, Ungerer, & Sherman, 1986). Together, these impairments can all be hypothesized to have a negative effect on the ability to implement and retain new task instructions (i.e., negative correlations with the IBC effect). Last, the BIS/BAS questionnaire (Carver & White, 1994) was used to see whether general motivational differences can explain interindividual differences in the IBC effect.

Method

Participants

One hundred and eighty-four undergraduate students from Ghent University (156 female, mean age = 18.78 years, between 17 and 37 years old, 156 right-handed, all proficient in Dutch) participated in the experiment in return for student credit. A power analysis showed that to detect a correlation of .2 or higher with a one-tailed significance test at $p = 0.05$ and a power of 80%, we would need 153 participants. Each participant provided signed informed consent at the beginning of the experiment and was debriefed after the experiment.

Materials

Apparatus The task was programmed with Tscope5 (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006), and performed on five desktop computers (Dell Optiplex 3020 minitower, Intel Core i5-4570; screen: Benq XL2411Z LED monitor: 1,920 × 1,080).

Stimuli The stimuli consisted of 56 high-frequent Dutch four-letter nouns selected from the SUBTLEX-NL database (Keuleers, Brysbaert, & New, 2010; the same selection was used in Braem et al., 2017). The stimuli were centrally presented in Arial bold font, size 36.

Procedure The first part of the experiment consisted of four blocks, of which the first block was considered a practice block and excluded from the analysis. This was necessary because multiple participants needed this first block to become familiar with the paradigm, as is also evident from the significantly larger RTs and error rates on this first bock. Each block was separated by a self-paced break, and contained seven runs (four runs with four diagnostic trials; two with eight diagnostic trials and one run with 16 diagnostic trials), which totaled 48 diagnostic trials per block. Each run ended with one inducer trial (i.e., seven inducer trials per block). The words were randomly paired, and each pair was randomly assigned to one of the runs.

The procedure is depicted in Fig. 1. Each run started with the presentation of two S-R mappings (e.g., "if lamp press left, if bird press right") for a maximum of 20 s, or until the participant pressed the spacebar. Following the inducer instructions, the diagnostic task started. Stimuli were either printed roman or italic in black on a white background, and participants had to press left when the word was printed in roman, and right when the word was printed in italic. Each stimulus was presented in both font types an equal number of times in a random order, resulting in a balanced number of congruent and incongruent words. After the diagnostic task, one of the two words was printed in green, upon which the inducer instructions had to be applied. The inter-trial-interval was 750 ms, and the inter-run-interval was 1,500 ms. The response deadline was 2,000 ms for both tasks. The screen turned red for 200 ms following incorrect or late responses.

The second part of the experiment consisted of another four blocks, only this time, no inducer task was used. Instead, only diagnostic trials were presented where the words were either "left" or "right" in Dutch (i.e., "links" or "rechts"). Similar to the first part, each block counted 48 trials and had the same inter-trial-interval and response deadline. To ensure comparability, we also excluded the first block of this phase from the analysis. This phase was set up to assess the semantic Simon effect. Importantly, due to an initial programming error in this task (i.e., the order of congruency conditions was not randomized), we had to exclude the first 40 participants from all analyses that included the semantic Simon effect.

Finally, participants filled in four questionnaires. Specifically, they were asked to complete a suggestibility questionnaire (Kotov et al., 2004), Beck's depression Inventory (Beck et al., 1961), the Autism Questionnaire (Baron-Cohen et al., 2006), and, finally, the BIS/BAS questionnaire (Carver & White, 1994).

All data, experiment files, and analysis scripts can be found on the open science framework (https://osf.io/5y4gh/)

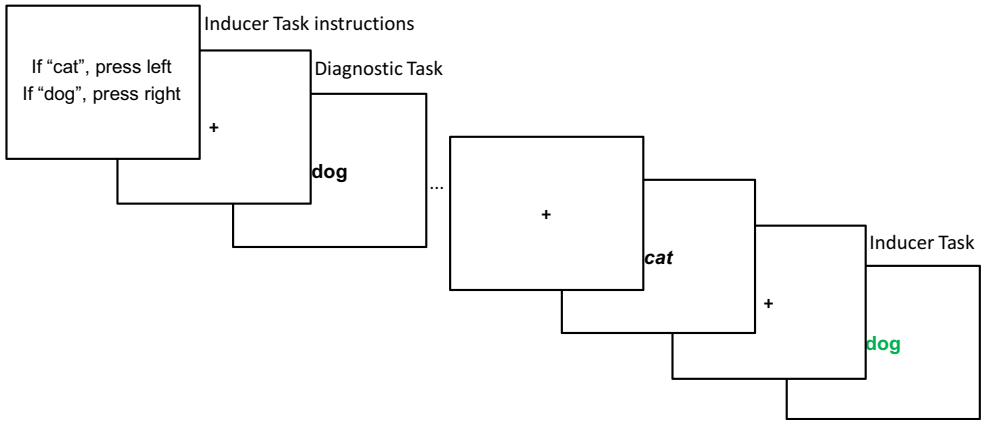

Fig. 1 General paradigm and trial procedure. Each run started with the presentation of the inducer task instructions, which participants were instructed to execute upon presentation of the task stimulus in green. In between, participants were presented with four, eight, or 16 trials of the diagnostic task where the task stimuli were presented in black and the task was to respond to their font type (upright vs. italic). Diagnostic trials can be further categorized into congruent or incongruent trials, depending on whether they require the same or different response as the to-be-performed instructions

Results

Participants were removed from the analysis when their accuracy on either the diagnostic or the inducer task was more than two SDs below the mean. However, because this threshold was below chance level for the inducer task, a total of 15 participants were excluded from the analysis because they did not perform accurately on more than 11 out of 21 trials of the inducer task (52.38%). Another three participants were removed from analysis because of scoring less than two SDs from the mean on the diagnostic task. Notably, when we used more conservative criteria for excluding subjects (resulting in the exclusion of ten more subjects), we reached the same statistical conclusions. The remaining 165 participants (139 female, mean age = 18.55 years, 141 right-handed) had a mean accuracy on the diagnostic task of 95.19% (SD = 3.33%), and a mean accuracy on the inducer task of 86.81% (SD = 9.94%). For all analyses, diagnostic trials that were part of a run where an error was made on the inducer trial, or diagnostic trials following an error on a preceding diagnostic trial, were removed from the analyses. For the RT analyses, only correct RTs were considered, and RTs shorter than 100 ms or longer than 2.5 SDs than the mean for (congruent or incongruent) diagnostic or inducer trials were excluded. Although we had specific hypotheses about the direction of our effects, we report two-tailed tests to also statistically assess the opposite finding (as observed in Meiran et al., 2016).

In line with previous studies, participants showed a significant IBC effect in both RTs, $t(164) = 6.232$, $p < .001$, and error rates, $t(164) = 5.827$, $p < .001$, which was stable across diagnostic trials (see Supplementary Material; see also Deltomme, Mertens, Tibboel, & Braem, 2018; Meiran et al., 2015a). More specifically, subjects performed slower (679 ms, SD = 90 ms) and made more errors (5.74 %, SD = 4.89 %) on incongruent trials than on congruent trials (659 ms, SD = 87 ms; 3.55 %, SD = 3.21 %). Similarly, participants showed a congruency effect when presented with the words left and right (instead of words instructed to be associated with left or right) in the RTs, $t(133) = 4.687$, $p < .001$, and error rates, $t(133) = 3.712$, $p < .001$. They responded slower (601 ms, SD = 72 ms), and made more errors (6.03 %, SD = 4.40 %), on incongruent trials, than congruent trials (585 ms, SD = 62 ms; 4.45 %, SD = 3.50 %). The semantic Simon effect was numerically smaller than the IBC effect, but not significantly, $t(133) = 1.348$, $p = .180$. One might have predicted a reverse pattern, as the semantic Simon effect supposedly measures interference from more "hardwired" associations. However, we believe one should take into account that word meaning was always irrelevant in the semantic Simon task, and the same two words were used throughout the task. This might have made it easier for participants to inhibit word meaning in this task (relative to the IBC paradigm).

Introduction to the correlational analyses In what follows, we report the analyses of our three hypotheses outlined in the introduction. Specifically, first and most importantly, we tested our hypothesis that the IBC effect should be associated with faster inducer task performance (*inter- and intra-individual differences in the IBC task*). Second, we examined if the IBC effect was indeed "independent" from the semantic Simon effect (*inter-individual differences across tasks*). Finally, we explored whether the IBC correlated with certain personality traits of interest (*inter-individual differences across tasks and questionnaires*).

To this end, we computed standardized measurements of the IBC effect (we believe these standardized measures are superior, but non-standardized measures gave highly similar results, reaching the same levels of significance). Specifically, the IBC effect was calculated using standardized RTs per subject (i.e., z-scores). For each subject separately, we substracted

the mean RT from the RTs that were eligible for analysis (see above), after which we divided the remaining values by their standard deviation (zRT = (RT – mean(RT)) / SD(RT)). These standerdized RT values were used to compute the IBC effect per subject ($meanzRT_{incongruent} - meanzRT_{congruent}$). This way, each individual's IBC effect was corrected for overall differences in RT. The IBC effect in the error rates was quantified by subtracting mean error percentage on congruent trials from that on incongruent trials, which was further multiplied by mean accuracy (dividing by mean error rate or the SD was impossible due to a few subjects showing no errors). In other words, this score indicated the difference in percentage errors attributable to incongruent versus congruent trials. Upon closer investigation of these congruency effects and the initial scatter plots, we further removed one more subject from the correlation analyses as this person had a standardized IBC effect of more than 3 SDs below the mean (-0.71). Importantly, if anything, this participant actually skewed the correlations very much "in favor" of our first hypothesis as his or her mean inducer RT was also the highest (but not an outlying value: 1,392 ms). Finally, to ensure that our analysis were minimally sensitive to outliers, we only report non-parametric rank-ordered Spearman correlations, but Pearson r-values were highly similar and mostly reached the same statistical conclusions. We still report Peason r-values in addition to the Spearman correlations, when their conclusions diverged, as this can be informative about the role of more extreme values in driving a correlation.

Internal reliability of the interference effects Before studying the correlations between our interference effects and inducer RTs, we wanted to determine the internal reliability of our measures. To that end, we used the Spearman-Brown prediction formula, after splitting our data in odd versus even *quadruplets* (note that contrasting odd versus even *trials* or *pairs of trials* would result in an uneven distribution of trials that are at the beginning or end of a diagnostic phase), or odd versus even runs for the inducer RT. These analyses indicated that the Spearman-Brown corrected split-half reliability was r_s = .333 for the IBC effect, r_s = .599 for the semantic Simon effect, and r_s = .836 for the inducer RT. This fits well with the observation that single mean measures tend to show higher reliabilities than difference score measures (Paap & Sawi, 2016), or the relatively low reliability of cognitive control tasks more generally (Hedge, Powell, & Sumner, 2018). Importantly for the present purposes, it should be kept in mind that potential correlations between these measures will be restricted by the reliability of those two measures to begin with. Therefore, the absence of a correlation should be interpreted with caution, as it could also be driven by the relatively low reliability of these measurements (i.e., especially the IBC effect).

Inter-individual differences in the IBC task To investigate whether the IBC effect can indeed be interpreted as a measure of encoding strength of instructed associations, we measured its correlation with task performance on the inducer task. Based on the basic assumption that the IBC effect reflects the degree of activation of the instructed task set in working memory, we hypothesized a negative relation between the IBC effect and the inducer RTs since a high activation (high IBC effect) should result in faster application of the instructed task set (lower inducer RTs). In line with this prediction, the mean RT on the inducer task correlated negatively with the IBC effect in the RTs, r_s = -.248, p = .001 (n = 164), showing that participants who had a larger IBC effect were faster in performing the inducer trial (see Fig. 2, left panel). Bayesian analyses indicated a Bayes factor B_{10} of 18.948, suggesting that these data were 18.984 times more likely to be observed under the alternative hypothesis (i.e., that they would correlate negatively) than the null hypothesis. This is typically considered strong evidence for the alternative hypothesis (Jeffreys, 1961).

To further ensure that this correlation was specific to the inducer trial RT, we performed a number of *post hoc* follow-up correlations (n = 164, unless indicated otherwise). First, interestingly, the correlational analysis between the IBC effect and mean diagnostic RT did not show this negative correlation, r_s = -.017, p = .828 (while mean diagnostic RT and mean inducer RT do correlate positively with each other, r_s = .283, p = .001). Therefore, the correlation between the IBC effect and inducer trial RT cannot be considered a side effect of a more general relation between the IBC effect and general RT. In fact, the above-reported correlation between the IBC effect and mean inducer RT also reached significance when controlling for mean diagnostic RT by subtracting it from mean inducer RT, r_s = -.214, p = .006, or when subtracting the RT from the last diagnostic trial from that of the inducer trial (which can be considered a "switch cost" from the diagnostic task to the inducer task), r_s = -.226, p = .004. Next, we also wanted to see whether this correlation with the inducer RT was not related to a speed-accuracy trade-off. However, while people were indeed less accurate when being faster on the inducer task, r_s = .173, p = .026 (as in most speeded RT tasks), the error rate in the inducer task was not related to the IBC effect in the diagnostic task, r_s = -.027, p = .721. Moreover, our main correlation between the IBC effect and inducer RT was not influenced by the error rate on the inducer task, as evidenced by a partial correlation between the IBC effect and inducer RT, controlling for error rate on the inducer task, $r_{partial}$ = -.248, p = .001. Finally, to ensure that this correlation was specific to the IBC effect, and did not reflect a general relation between inducer trial RT and interference effects, we also tested the correlation between the inducer RT and the semantic Simon effect, r_s = .006, p = .941 (n = 133). Please note that the sample size for this correlation was only 133, instead of 164.

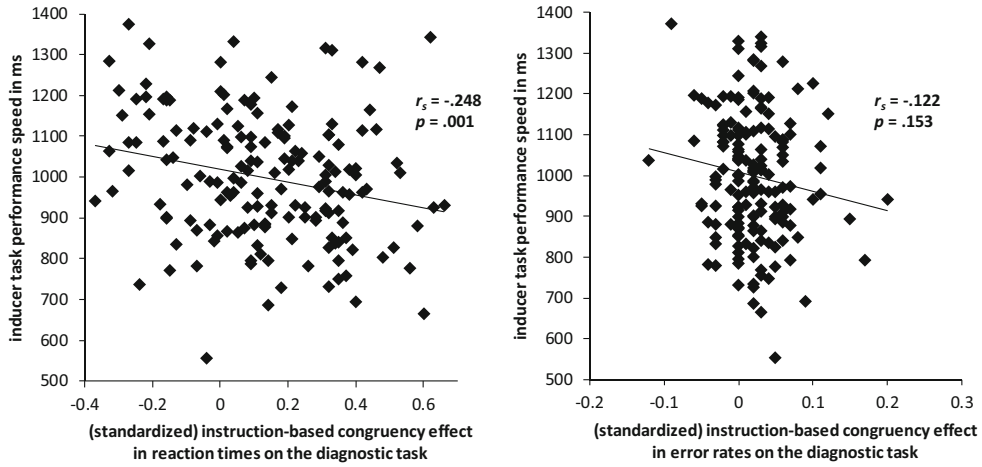

Fig. 2 Scatter plots (with linear trendline) depicting the general relation between the (standardized) instruction-based congruency effect in reaction times (RTs) (**left panel**) or error rates (**right panel**) and the (instructed) inducer time RT. Together, these analyses demonstrate that participants who performed the instructed task more efficiently (i.e., faster) showed a bigger instruction-based interference effect. The r_s-values indicate the Spearman correlation between the two variables depicted in the scatter plot

Nonetheless, Bayesian analyses indicated that this was strong evidence for the null hypothesis, with a Bayes factor B_{01} of 9.219, suggesting these data were 9.219 times more likely under the null hypothesis than the alternative hypothesis (i.e., that it would correlate positively or negatively).

Last, we also analyzed whether the IBC effect computed in the error rates might show a similar relation with inducer task RT, also allowing us to counter an interpretation in terms of a speed-accuracy trade-off. The IBC effect in the error rates seemed to show a marginally significant correlation with inducer RT when measured with Pearson correlation, r = -.131, p = .094, but, importantly, not when analyzed with Spearman correlation, r_s = -.112, p = .153, suggesting that this hint at a correlation was mostly driven by the more extreme values (see Fig. 2, right panel).

In the Supplementary Material, we report additional analyses investigating correlations with the time needed to implement instructions (participants could choose when to proceed to the task), which indicated a trend for a positive correlation with the IBC effect in the error rates (consistent with recent findings by Cole, Patrick, Meiran, & Braver, 2018).

Intra-individual differences in the IBC task While the above results demonstrate that subjects who showed a faster application of task instructions results also showed a larger IBC effect, we also wanted to investigate whether the same relation could be observed within subjects. That is, we analyzed within subjects whether participants showed faster performance on the inducer task in runs where they showed larger IBC effects. To that end, we calculated the IBC effect per run, based on the first four trials only (to ensure that the size of the IBC effect was not determined by the length of the run), excluding runs where performance on one of the first four diagnostic trials (or the inducer trial) was incorrect. Following these extra exclusion criteria, three more participants were removed from the analyses because of having less than seven runs/observations (out of 21, which equaled to less than 2 SD below the mean number of observations) where the IBC effect could be calculated.

Next, Pearson's r correlation coefficients were calculated for each subject separately, measuring the correlation between the IBC effects and their respective inducer trial RTs. These Pearson r values were Fisher transformed, and subjected to a one-sample t-test, showing that the mean Fisher transformed Pearson r (mean = -.065, SD = .357; untransformed mean Pearson r = -.057) was significantly smaller than zero, $t(160)$ = -2.290, p = .023. This indicates that there was also a negative relation between the IBC effect and inducer RTs within subjects. A similar result was observed using a more conservative threshold of at least ten observations per subject (naturally resulting in a larger exclusion of subjects), $t(139)$ = -2.094, p = .038. One might note that a mean correlation of -.057 is small, but we believe it is important to emphasize that this is based on the relation between separate congruency effects each computed using only four observations, and single-trial RTs from the inducer task. Therefore, given the very noisy nature of these measures, we believe this is a realistic effect size.

Inter-individual differences across tasks In a second step, we also examined whether the congruency effect in the IBC task correlated with the congruency effect in the semantic Simon task, where the interference stemmed from known associations with the words “left” or “right.” To test the correlation between the two, we used the same standardized measures of the congruency effect as computed for the IBC task.

Neither congruency effects correlated with one another, r_s = -.002, p = .982 (n = 133). Similarly, the congruency

effects as measured in error rates showed no correlation r_s = .099, p = .256. Bayesian analyses indicated a Bayes factor B_{01} of 10.287, suggesting that these data were 10.287 times more likely to be observed under the null hypothesis than the alternative hypothesis (i.e., that they would correlate positively). This is considered strong evidence for the null hypothesis (Jeffreys, 1961). Nonetheless, we want to warn the reader that this result should be treated with caution, given the low reliability of both measurements. Moreover, while the IBC effect in RTs did correlate postively with the IBC effect in error rates, r_s = .236, p = .002, as was the case for the correlation between the RT and accuracy measure of the semantic Simon effect, r_s = .479, p < .001, no such across-measurement correlations were observed across tasks, both ps > .1.

Inter-individual differences across tasks and questionnaires

Finally, as part of a more exploratory set of analyses, we set out to examine whether certain personality traits could predict differences in the IBC effect versus experience-based congruency effect. To that end, we also computed a third measure (for RTs and error rates separately) by substracting the semantic Simon effect from the IBC effect (for which the Spearman-Brown corrected split-half reliability was r_s = .421). Arguably, this measure could offer a more sensitive index as to how much more a subject shows interference from implemented instructions, relative to how much they show interference from well-known associations in long-term memory. As with previous analyses, the analyses that included the semantic Simon effect (or the difference score with it) had a sample size of 133, the others 164.

In short, none of the four questionnaires showed correlations with the congruency effects or the difference between them, even when ignoring corrections for multiple comparisons. Similarly, regression analyses with the IBC effect (or difference score) as dependent variable, and all four questionnaire scores, gender, and age as predictors, showed no significant effects. Only one questionnaire correlated marginally significantly with the difference score between the IBC task and the semantic Simon effect: the autism quotient questionnaire, r_s = -.156, p = .073. Because this questionnaire also has subscales, which have been argued to measure important subtraits, we further explored whether this difference was more pronounced for one of the five subscales. Out of the five subscales (social skills, attention switching, attention to detail, communication, and imagination), only attention switching and communication correlated significantly with this difference score, - r_s = .175, p = .044, and - r_s = .202, p = .020. The other three subscales showed no significant correlation, all ps >.2. These two subscales are thought to measure the problems people with autism experience in switching their attention and communicating, such as in social interactions or other environments where a lot of new information is being presented, and have been linked to other symptoms of autism such as perseveration and insistence on sameness. Possibly, these autism traits could thus be related to a reduced capacity to hold newly instructed information in working memory. However, these results should be treated with caution, and are only reported for future reference.

Finally, because within-subject RT variance has been found to correlate with other important constructs such as working memory and intelligence (Meiran & Shahar, 2018; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007), we also tested these correlations using non-standardized IBC or semantic Simon effects, but this did not reveal new correlations. The complete set of correlations can be found on the Open Science Framework.

Discussion

With this study we wanted to test whether the IBC effect increases as a function of efficient task execution. If this instruction-based reflexivity does reflect the presence of a representation in working memory that allows for fast task execution upon demand, we can expect the IBC effect to be bigger, the stronger the representation. To this end, we analyzed the correlation between the IBC effect and the inducer task RTs. As predicted, a negative correlation was observed, both between and within subjects: a bigger IBC effect was associated with faster inducer task performance.

This result relates well to earlier findings in classical conflict paradigms. For example, it has been observed that (over)training color word reading, thought to result in stronger memory representations, induces bigger Stroop effects (Cohen, Dunbar, & McClelland, 1990). In a similar vein, Yeung and Monsell (2003) demonstrated how task sets that received more practice caused more interference than those that did not. However, for obvious reasons, training cannot be used to strengthen the representations of *merely instructed* task sets. Therefore, we took a different approach by studying natural variations between and within different individuals in another index of memory representation intensity, the reaction time speed (on the very first trial), widely thought to be a valuable index of (working) memory strength (Monsell, 1978; Sternberg, 1969).

Our study also suggests that the IBC effect taps into other types of memory representations than more typical experience-based congruency tasks do. Specifically, we found no correlation between the IBC effect and a closely matched version that measures interference from more long-term associations with the sides left and right, namely by presenting the words “left” and “right.” These findings are interesting in light of other studies that did show positive correlations between more traditional congruency effects (e.g., Keye,

Wilhelm, Oberauer, & Van Ravenzwaaij, 2009; Miyake et al., 2000; but see, Feldman & Freitas, 2016; Hedge et al., 2018). Therefore, this result suggests that the IBC effect measures an interference effect from an independent source more than other congruency tasks do. However, given that we only tested this idea with one experience-based congruency task (i.e., semantic Simon effect), future research is warranted. Moreover, given the relatively low reliability of interference effects, as observed both here and elsewhere (e.g., Hedge et al., 2018; Paap & Sawi, 2016; Whitehead, Brewer, & Blais, 2018), we should be aware of the possibility that this lack of correlation could also be due to an insensitivity of our measures. Finally, it is also important to emphasize that this first evidence for a relative independence does not allow for the conclusion that both types of pathways cannot interact. For example, previous studies have shown that congruency effects resulting from overtrained associations can be very sensitive to working-memory maniplations (Wühr & Biebl, 2011), and can even be eliminated through the use of simple instructions (e.g., Bardi, Bundt, Notebaert, & Brass, 2015; Theeuwes, Liefooghe, & De Houwer, 2014).

The present findings are also largely consistent with previous studies that manipulated different degrees of task difficulty to promote different degrees of task preparation. For example, Liefooghe et al. (2013; Experiment 2) manipulated the response deadline of the inducer task, assuming that more stringent response deadlines would go along with more task preparation, as was demonstrated by the larger IBC effects. In a similar vein, Meiran and colleagues (2015a) manipulated the number of inducer task trials, assuming that people would put more effort in preparing the inducer task if fewer inducer task trials were available, which was indeed associated with larger IBC effects. Finally, a recent study by Whitehead and Egner (2018b) demonstrated how the IBC effect increases when the general likelihood that subjects will actually have to perform the inducer task increases. Arguably, these experiments also showed a relation between task preparedness and the IBC effect. Our study can be considered complementary to their approach, as we, rather than trying to create different degrees of task difficulty, studied natural inter- and intra-individual differences in "task preparedness" to investigate its relation to the IBC effect.

Interestingly, as already hinted at in the introduction, Meiran et al. (2016) used a similar approach to the present study, but observed a positive correlation between their congruency effect in their diagnostic task and reaction times on the inducer task. We believe this is likely due to the fact that, in the PI task used by Meiran and colleagues (see introduction), participants might consider the inducer task and the diagnostic task as one integrated task (see also Meiran et al., 2015a, p. 780 for a discussion). In contrast, in the PI task used here (Liefooghe et al., 2012, 2013), the inducer and diagnostic task were more likely to be considered two independent choice-reaction tasks, because they are also introduced to the participants as two different tasks. As such, the positive correlation observed by Meiran and colleagues might indicate that participants showing a better performance on one task component (e.g., fast responses to the inducer task), also perform better on the other task component (e.g., less interference in the diagnostic trials). Specifically, the PI task of Meiran et al. (2016) requires subjects to press a "proceed" button (which double functions as one of the two response buttons in the inducer task) as long as the stimuli are surrounded by a differently colored shape. This type of button pressing can be driven by different motives and strategies to the categorization task by Liefooghe and colleagues we used. Another possibility for our different results could lie in the fact that Meiran et al. (2016) used a slightly different dependent measure of inducer task performance. Namely, Meiran and colleagues studied the RT difference between the first inducer trial and second inducer trial (which we could not compute since we only had one inducer trial per run). This RT difference is directly related to the RT of the first inducer trial, and is therefore unlikely to completely explain our opposing results. However, it is possible that the mere foresight of having more than one inducer trial might have influenced the way in which participants implemented instructions. Future studies that do systematic comparisons between both types of paradigms should be able to further unravel whether both paradigms measure the same or (slightly) different aspects of instruction-based reflexivity.

Last, we also explored correlations between the IBC effect and self-reported personality trait questionnaires. These results clearly showed that the IBC effect shows little to no correlations with the investigated traits. That is, we found no evidence for a correlation between self-reported suggestibility traits, motivational traits, or symptoms of depression. While traits associated with autism spectrum disorder did hint at a relation with the IBC effect, we want to emphasize that these correlations were weak and possibly spurious. Therefore, at this point, we see no evidence for convincing relations with established self-report questionnaires, and would suggest for future studies to first try studying relations with more closely related functions, such as self-reported efficiency in instruction learning, or the degree to which people claim to use instructions in their everyday lives, to further assess the external validity of the IBC effect.

In sum, we conclude that our findings provide much-needed evidence for one of the core assumptions behind the instruction-based congruency task as a measure for task preparation, and further consolidates its role as a promising tool to study instruction learning. Interesting future avenues would be to see which neural correlates (e.g., more fine-grained neural pattern representations in the frontal cortex; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2016) drive, or further mediate, this (cor)relation between the IBC effect and the

initial application of an instructed task set, and to examine its relation with more everday measures of instruction implementation efficiency.

Acknowledgements We would like to thank Judith Goris, Iring Koch, Yoav Kessler, Ian McLaren, Nachshon Meiran, and two anonymous reviewers for useful comments on an earlier version of this manuscript or on our analyses. S.B. is supported by FWO – Research Foundation Flanders (12K6316N). The research reported in this paper was funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (IUAPVII/33) and by the Ghent University Methusalem Grant (BOF09/01M00209).

References

Anderson, J. (1992). Automaticity and the ACT theory. *American Journal of Psychology, 105(2),* 165-180.

Bardi, L., Bundt, C., Notebaert, W., & Brass, M. (2015). Eliminating mirror responses by instructions. *Cortex, 70,* 128-136.

Baron-Cohen, S., Hoekstra, R. A., Knickmeyer, R., & Wheelwright, S. (2006). The autism-spectrum quotient (AQ)—adolescent version. *Journal of Autism and Developmental Disorders, 36(3),* 343.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4(6),* 561-571.

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.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67(2),* 319.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97(3)*, 332-361.

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,* 1140-1154

Cohen-Kdoshay, O., & Meiran, N. (2009). The representation of instructions operates like a prepared reflex: Flanker compatibility effects found in first trial following S–R instructions. *Experimental Psychology, 56(2),* 128-133

Cole, M.W., Laurent, P. & Stocco, A.(2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience, 13(1),* 1-22.

Cole, M. W., Patrick, L. M., Meiran, N., & Braver, T. S. (2018). A role for proactive control in rapid instructed task learning. *Acta Psychologica, 184*, 20-30.

De Houwer, J. (1998). The semantic Simon effect. *The Quarterly Journal of Experimental Psychology: Section A, 51(3),* 683-688.

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

Deacon, T. W. (1997), *The symbolic species.* Norton, New York.

Deltomme, B., Mertens, G., Tibboel, H., & Braem, S. (2018). Instructed fear stimuli bias visual attention. *Acta Psychologica, 184,* 31-38.

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

Feldman, J. L., & Freitas, A. L. (2016). An investigation of the reliability and self-regulatory correlates of conflict adaptation. *Experimental Psychology, 63*, 237-247.

Gabriels, R. L., Cuccaro, M. L., Hill, D. E., Ivers, B. J., & Goldson, E. (2005). Repetitive behaviors in autism: Relationships with associated clinical features. *Research in Developmental Disabilities, 26(2),* 169-181.

Grahek, I., Everaert, J., Krebs, R. M., & Koster, E. H. (2018). Cognitive control in depression: Toward clinical models informed by cognitive neuroscience. *Clinical Psychological Science, 6(4)*, 464-480.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50(3),* 1166-1186.

Jeffreys, H. (1961). Theory of probability. Oxford, UK: Oxford University Press.

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.

Keye, D., Wilhelm, O., Oberauer, K., & Van Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: Testing means and covariance hypothesis about the Simon and the Eriksen Flanker task. *Psychological Research, 73(6),* 762-776.

Kotov, R. I., Bellman, S. B., &Watson, D. B. (2004). *MISS: Multidimensional Iowa suggestibility scale brief manual* [Internet]. Stony Brook: Stony Brook University (cited 2016 Apr 26, Available from)

Laberge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

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

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

Logan, G. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology, 39(2),* 367-386.

Logan, G. (1988). Toward an instance theory of automatization. *Psychological Review, 95(4),* 492-527.

Meiran, N., & Cohen-Kdoshay, O. (2012). Working memory load but not multitasking eliminates the prepared reflex: Further evidence from the adapted flanker paradigm. *Acta Psychologica, 139,* 309-313.

Meiran, N., Cole, M. W., & Braver, T. S. (2012). When planning results in loss of control: Intention-based reflexivity and working-memory. *Frontiers in Human Neuroscience 6*, 104.

Meiran, N., Diamond, G. M., Toder, D., & Nemets, B. (2011). Cognitive rigidity in unipolar depression and obsessive compulsive disorder: Examination of task switching, Stroop, working memory updating and post-conflict adaptation. *Psychiatry Research, 185(1-2),* 149-156.

Meiran, N., Liefooghe, B., & De Houwer, J. (2017). Powerful instructions: Automaticity without practice. *Current Directions in Psychological Science, 26(6),* 509-514.

Meiran, N., Pereg, M., Givon, E., Danielli, G., & Shahar, N., (2016). The role of working memory in rapid instructed task learning and intention-based reflexivity: An individual differences examination. *Neuropsychologia, 90,* 180-189.

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

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015b). Reflexive activation of newly instructed stimulus–response rules: Evidence from lateralized readiness potentials in no-go trials. *Cognitive, Affective, & Behavioral Neuroscience, 15,* 365-373.

Meiran, N., & Shahar, N. (2018). Working memory involvement in reaction time and its contribution to fluid intelligence: An examination of individual differences in reaction-time distributions. *Intelligence, 69,* 176-185.

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.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10(4),* 465-501.

Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2016). Neural coding for instruction-based task sets in human frontoparietal and visual cortex. *Cerebral Cortex, 27(3)*, 1891-1905.

Mundy, P., Sigman, M., Ungerer, J., & Sherman, T. (1986). Defining the social deficits of autism: The contribution of non-verbal communication measures. *Journal of Child Psychology and Psychiatry, 27(5),* 657-669.

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods, 274,* 81-93.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136(3),* 414.

Schneider, W. & Shiffrin, R. (1977). Controlled and automatic human information processing. *Psychological Review, 84*(2), 128-190.

Simon, J. R. (1969). Reactions towards the source of stimulation. *Journal of Experimental Psychology, 81,* 174-176.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57(4),* 421-457.

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,* 280- 286.

Stroop J. (1935). Studies of interference in serial verbal reaction. *Journal of Experimental Psychology, 18,* 643-662.

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.

Wenke, D., De Houwer, J., De Winne, J., & Liefooghe, B. (2015). Learning through instructions vs. learning through practice: Flanker congruency effects from instructed and applied S-R mappings. *Psychological Research, 79,* 899-912.

Whitehead, P. S., Brewer, G. A., & Blais, C. (2019). Are cognitive control processes reliable?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(5)*, 765-778.

Whitehead, P.S., Egner, T. (2018a). Cognitive control over prospective task-set interference. *Journal of Experimental Psychology. Human Perception & Performance, 44,* 741-755.

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

Wühr, P., & Biebl, R. (2011). The role of working memory in spatial SR correspondence effects. *Journal of Experimental Psychology: Human Perception and Performance, 37(2),* 442.

Yeung, N., & Monsell, S. (2003). The effects of recent practice on task switching. *Journal of Experimental Psychology: Human Perception and Performance, 29(5)*, 919-936.