

# List-level control in the flanker task

Julie M Bugg<sup>1</sup>  and Corentin Gonthier<sup>2</sup> 

Quarterly Journal of Experimental Psychology  
2020, Vol. 73(9) 1444–1459  
© Experimental Psychology Society 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1747021820912477  
qjep.sagepub.com



## Abstract

Current theories posit multiple levels of cognitive control for resolving conflict, including list-level control: the global or proactive biasing of attention across a list of trials. However, to date, evidence for pure list-level control has largely been confined to the Stroop task. Our goals were twofold: (a) test the generality of theoretical accounts by seeking evidence for list-level control in the letter flanker task, using an established method involving diagnostic items, and investigating the conditions under which list-level control may and may not be observed and (b) develop and test a potential solution to the challenge of isolating list-level control in tasks with a relatively limited set of stimuli and responses such as arrow flanker. Our key findings were that list-level control was observed for the first time in a letter flanker task on diagnostic items (Experiment 1), and it was not observed when the design was altered to encourage learning and use of simple stimulus–response associations (Experiment 2). These findings support the generalisability of current theoretical accounts positing dual-mechanisms or multiple levels of control, and the associations as antagonists to control account positing that list-level control may be a last resort, to conflict tasks besides Stroop. List-level control was also observed in the arrow flanker task using a modified design (Experiment 3), which could be extended to other conflict tasks with limited sets of stimuli (four or fewer), although this solution is not entirely free of confounds.

## Keywords

Cognitive control; flanker task; proportion congruency effects; proactive control

Received: 26 April 2019; revised: 18 February 2020; accepted: 19 February 2020

There is a growing appreciation in the cognitive control literature of the intimate relationship between learning and cognitive control. Humans learn the statistical regularities within their environment and adapt attention accordingly (e.g., Abrahamse et al., 2016; Egner, 2014). As an example, consider a student who is enrolled in two lecture courses. In one course, the student's classmates frequently chat in a distracting way; this happens more rarely in the other course. Over time and likely outside of their awareness (Blais et al., 2012), the extent to which the student attends to the chatter of their classmates may vary from one course to the other. If this real-life example follows what is observed in laboratory tasks, the student should be less distracted by chatter in the course in which it occurs frequently. Why that is the case has been explained by various mechanisms, which can be reactive (e.g., item-specific control: the student learns which classmates [items] tend to chit-chat and pays less attention to these individuals) or proactive (list-wide control: the student learns the likelihood of encountering distracting chatter in each course and minimises attention to neighbouring classmates in courses with a high probability of encountering chatter, so as to prevent being distracted).

With the emergence of theories positing dual mechanisms (reactive and proactive) or multiple levels of cognitive control (e.g., item level and list level) for resolving conflicts in information processing (e.g., Braver et al., 2007; Bugg, 2012), a challenge researchers have faced in the lab is designing task variants that effectively isolate each level, thereby allowing for examination of potential dissociations. Arguably, the most progress has been made in the context of the Stroop task, where participants name the ink colour of colour words that are congruent (e.g., RED in red ink) or incongruent (e.g., BLUE in red ink). Variants have been developed that isolate item-level (item-specific) and list-level (list-wide) control of Stroop interference (for reviews, see Bugg, 2017; Bugg & Crump,

<sup>1</sup>Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

<sup>2</sup>Department of Psychology, University of Rennes, Rennes, France

### Corresponding author:

Julie M Bugg, Department of Psychological and Brain Sciences, Washington University in St. Louis, Campus Box 1125, St. Louis, MO 63130, USA.

Email: jbugg@wustl.edu

2012; see also Crump & Milliken, 2009, for a variant that isolates context-level control). Contrasting performance between these two variants (e.g., Bugg, 2014a, 2014b; Gonthier et al., 2016) has enabled researchers to further theoretical understanding of the characteristics of reactive and proactive control, respectively. For example, relative to reactive control, proactive control operates more generally, affecting all stimuli in each context including those that may differ from the stimuli that promoted learning within a context (e.g., Gonthier et al., 2016). Consider the example offered earlier—what this means is that a student may learn that there tends to be a high probability of encountering chatter based on the consistent chatter of certain classmates, but the adaptation in attention that occurs based on this learning (i.e., minimising attention to neighbouring classmates) would benefit performance (i.e., help keep one focused on the lecture) not just when encountering those classmates but additionally when encountering any other classmate and their distracting speech.

However, such theorising about mechanisms or levels of control and their defining characteristics has been based largely on performance of the Stroop task, raising the question of whether existing accounts are task-specific or task-general accounts of cognitive control. For example, it is not yet known whether these accounts generalise to tasks involving spatial conflict (e.g., filtering out the distraction in the periphery). Thus, it is important to examine whether the mechanisms of cognitive control posited to exist within the Stroop task are also observable and operate similarly within other conflict tasks. Achieving this aim critically depends on the existence of task variants that can isolate list-level and item-level control, and for reasons we will explain momentarily these variants have been elusive in conflict tasks other than the Stroop.

With this overarching aim in mind, this study had two objectives: (a) test the generality of theoretical accounts by seeking evidence for list-level control in a letter flanker task (where participants indicate the identity of a central letter while ignoring the identity of flanking letters; e.g., Eriksen & Eriksen, 1974), another commonly used conflict task, and investigating the conditions under which list-level control may and may not be observed and (b) develop and test a potential method for isolating list-level control in conflict tasks which are composed of fewer unique stimuli than the Stroop task, such as the arrow flanker task (where participants indicate the identity of a central arrow while ignoring the identity of flanking arrows; e.g., Fan et al., 2002; Kopp et al., 1996).

### **Use of proportion congruence manipulations to isolate levels of control**

A relatively straightforward manipulation that has effectively isolated list-level and item-level control in the Stroop

task is the proportion congruence manipulation, which varies the relative proportion of congruent to incongruent trials (see Bugg, 2012, 2017; Bugg & Crump, 2012). The term *isolate*, here, refers to the ability to measure cognitive control independent of confounds or alternative mechanisms that can otherwise explain the effect (e.g., item-specific contingency learning; Schmidt & Besner, 2008).<sup>1</sup> The proportion congruence manipulation results in a mostly congruent condition (words and colours match in most trials) and a mostly incongruent condition (words and colours conflict in most trials). This manipulation can be applied at the level of a list (a given block of trials is mostly congruent or mostly incongruent), or at the level of items within a list (certain stimuli, as defined by a particular feature such as the colour of the item, are mostly congruent or mostly incongruent: for example, the word blue is usually presented with a colour other than blue). Comparing Stroop effects (incongruent–congruent) between mostly congruent and mostly incongruent conditions reveals smaller Stroop effects in the mostly incongruent condition, regardless of whether the manipulation is applied to the list or item level. However, the interpretation of this difference varies depending on the level targeted by the manipulation.

The presence of a “pure”<sup>2</sup> list-wide proportion congruence effect (LWPC; i.e., smaller Stroop effect in mostly incongruent lists compared with mostly congruent lists) has been taken as evidence for a global control mechanism that modulates attention to the distracting dimension, based on the global probability of conflict within a list. In other words, control would be implemented to decrease the processing of words when the probability of conflict is high, as in a mostly incongruent list (see Egner & Hirsch, 2005, for a parallel view that colour processing may be amplified). In terms of the dual mechanisms of control account (Braver et al., 2007), such a mechanism is considered proactive in that it is thought to exert a preparatory or sustained influence on performance, being engaged even prior to stimulus onset (DePisapia & Braver, 2006; Gonthier et al., 2016; for a pathway level control model, see Botvinick et al., 2001). By contrast, the presence of an item-specific proportion congruence effect (ISPC; i.e., smaller Stroop effect for items that are mostly incongruent compared with items that are mostly congruent) has been taken as evidence for an item-level control mechanism that modulates attention to the word dimension on a trial-by-trial basis, depending on the likelihood of conflict associated with a given item (e.g., Bugg & Dey, 2018; Bugg et al., 2011; Bugg & Hutchison, 2013; Chiu et al., 2017; for computational models, see Blais et al., 2007; Verguts & Notebaert, 2008). Such a mechanism is considered reactive because it exerts its influence poststimulus onset, once the item has revealed its identity as a mostly congruent or mostly incongruent item (see Gonthier et al., 2016).

In the flanker task literature, there is currently evidence for item-level control based on the use of select ISPC

manipulations (Bugg, 2015): mostly incompatible items elicit a smaller flanker compatibility effect (incompatible-compatible) than mostly compatible items,<sup>3</sup> even within a list that is globally unbiased (50% compatible). However, researchers have yet to demonstrate an LWPC effect in flanker performance that can be attributed to list-level control, independently of item-specific mechanisms (see Bugg, 2012). Many studies have shown reduced conflict in mostly incompatible lists when compared with mostly compatible lists (e.g., Gratton et al., 1992; Lehle & Hübner, 2008; Taylor, 1977; Wendt et al., 2012; Wendt & Luna-Rodriguez, 2009), but these designs used biased items within biased lists (e.g., 25% compatible items within 25% compatible lists), leading to ambiguous results. Indeed, reduced conflict in a mostly incompatible list comprising mostly incompatible items can equally be driven by a reactive control mechanism operating at the level of items (triggered after the presentation of each item) or by item-specific contingency learning (producing responses that are highly contingent on the distracting feature of each item), or by a proactive control mechanism operating at the level of the list (applied preemptively to all items).

The question of the existence of the LWPC effect in flanker tasks has important consequences for cognitive control research: If an LWPC effect cannot be observed, the implication is that accounts positing dual mechanisms or multiple levels of control may not be representative of the flanker task, one of the most frequently used paradigms for the study of conflict. Like the Stroop task, the flanker task requires selection of relevant over irrelevant information. However, in the flanker task, selection is based on spatial location such that participants respond to the central target while ignoring the flanking stimuli. This core difference in attentional deployment could affect the way cognitive control is implemented in the task (see, for example, Spieler et al., 2000, for evidence that the effects of conflict may differ between Stroop colour naming and flanker tasks). In addition, certain flanker tasks may encourage reliance on contingency learning (for evidence and a discussion, see Bugg, 2015), which could preclude use of list-level control (Bugg, 2014a). Therefore, it remains unclear whether list-level control is a legitimate mechanism used to resolve flanker conflict, or alternatively if item-level mechanisms govern performance on this task.

Encouraging results for the existence of a pure LWPC effect were provided by the findings of Wendt et al. (2012) who demonstrated a pattern that fits with the operation of a list-level control mechanism. They combined an LWPC manipulation in a two-choice letter flanker task (respond to central letter) with an interleaved search task (indicate position of a target number in a three-digit string). The key finding was worse performance on the search task when the position of the search target corresponded to the position of the flankers in the flanker task, and this effect was

especially pronounced in the mostly incompatible list. Wendt et al. attributed this pattern to a “conflict-induced filter” that attenuated processing of the flankers to a greater degree in the mostly incompatible list. This evidence is consistent with the use of list-level control, but the design of this study makes it unclear whether the results are generalisable to more traditional flanker tasks. For example, the presence of a secondary search task could elicit subtle dual-task effects. Moreover, indexing list-level control via performance on a secondary task makes it difficult to perform direct comparisons to existing indices of item-level control (e.g., as assessed by the ISPC effect in the flanker task; Bugg, 2015). It is thus desirable to determine whether a pure LWPC effect can be observed in flanker performance itself, independent of item-specific mechanisms. Given the role of potentially confounded item-specific mechanisms, doing so required a careful choice of experimental paradigm.

### Isolating list-level control: the ABS design

In the Stroop task, list-level control has been successfully isolated using a design inspired by the associations as antagonists to control (AATC) account (Bugg, 2014a), which we refer to hereafter as the AATC-based Stroop design (*ABS design*) for short. The ABS design comprises two lists (blocks) of trials: one that is mostly congruent and one that is mostly incongruent. Critically, there are two sets of items in each list: the *inducer* items are biased and serve to induce an attentional bias within each list, whereas the *diagnostic* (transfer) items are unbiased and frequency-matched across lists. For example, the words and colours RED, BLUE, PURPLE, and WHITE may play the role of inducer items (e.g., 75% congruent in the mostly congruent list and 25% congruent in the mostly incongruent list), whereas the words and colours YELLOW and GREEN play the role of diagnostic items (50% congruent in both lists). Critically, these two sets of items are randomly intermixed in each list, creating lists that are mostly congruent (67% congruent) or mostly incongruent (33% congruent). As an additional control, diagnostic items share no overlapping features with the inducer items: for example, the word RED only appears written in red, blue, purple, or white (and not in yellow or green). Consequently, there is no possibility that an attentional setting that became associated with features of the inducer items (e.g., decreased word processing for red in mostly incongruent lists) could be primed in a bottom-up fashion on trials comprising diagnostic items. The diagnostic items are then used to examine the LWPC effect, independently of item-specific mechanisms and bottom-up priming. Observing a reduced Stroop effect on the diagnostic items in the mostly incongruent list when compared with the diagnostic items in the mostly congruent list (i.e., an LWPC effect for

diagnostic items) is necessary for claiming the operation of pure list-level control (for a recent consensus paper on this topic, see Braem et al., 2019).

In the Stroop task, an LWPC effect for diagnostic items has been observed consistently using the above design (Bugg, 2014a; Bugg & Chanani, 2011; Gonthier et al., 2016; Hutchison, 2011). However, the earliest attempts at examining transfer failed to demonstrate list-level control (Blais & Bunge, 2010; Bugg et al., 2008); importantly, these first attempts differed in a subtle but important way from the design just described in that they utilised a two-item set for the inducer items (e.g., only the words and colours RED and BLUE were used). As Bugg (2014a) demonstrated, using a two-item set may have biased the system to capitalise on stimulus–response associations that enabled high levels of performance and precluded use of list-level control. That is, because a given word (e.g., RED) appeared in a single, incongruent colour (blue) on most trials in mostly incongruent lists, participants may have simply learned to say blue in response to the word RED, leading to fast response times (RTs) on incongruent trials and smaller Stroop effects selectively for inducer items without actually implementing cognitive control. According to the AATC account (Bugg, 2014a), list-level control may be a last resort that participants engage when they cannot rely on simple stimulus–response learning (i.e., predicting highly contingent responses) to achieve task goals. As anticipated by AATC, shifting to the ABS design described above where the inducer set was composed of four items instead of two has consistently produced list-level control in the Stroop task (Bugg, 2014a; Gonthier et al., 2016; see also Bugg & Chanani, 2011; Hutchison, 2011). In this case, participants cannot predict the correct response on incongruent trials because a given word (e.g., RED) is paired equally often with each of the three other colours (blue, purple, and white).

## Current study

The major objective of our study was to seek evidence for pure list-level control in the flanker task. As such, it may seem that an obvious approach would be to isolate this mechanism by applying the ABS design used previously in the Stroop task. However, doing so is not as simple as it may seem in conflict tasks like the flanker task, given that the ABS design requires a minimum of five or six items (two diagnostic items, and three or more inducer items to limit associative learning). The first challenge is evident when considering one popular version of the flanker task that employs strings of letters (e.g., HHHHHHH; SSSSHSS; Eriksen & Eriksen, 1974). Participants must learn stimulus–response mappings for six different stimuli, and this may create a memory load that could impede list-level control, for all participants or disproportionately for

certain groups (cf. Kane & Engle, 2003). In the Stroop task, participants usually respond vocally, which circumvents the issue.

The second challenge is evident when considering a second popular version that employs strings of arrows that are compatible (e.g., <<<<<<<) or incompatible (e.g., >>><>>>) (cf. Fan et al., 2002; Kopp et al., 1996). The arrow flanker task comprises a maximum of four stimuli (left, right, up, and down arrows) and associated responses (see, for example, Bugg, 2015, for a four- rather than two-choice version). This makes it impossible to create two independent sets of items including an inducer set that is composed of more than two stimuli, as the ABS design requires.

Our approach was therefore as follows.<sup>4</sup> In Experiment 1, we applied the ABS design to a letter flanker task. To minimise any putative effects of load on list-level control, we used nonarbitrary stimulus–response mappings. To foreshadow, we observed the first evidence for a pure LWPC effect on diagnostic items in a flanker task, confirming that list-level control is indeed used in this task. Experiment 2 was then conducted to test the theoretical prediction of the AATC account that list-level control may not be observed if the design enables participants to instead rely on simple stimulus–response associations. Replicating the pattern observed previously with Stroop and consistent with the AATC account (Bugg, 2014a), we did not find an LWPC effect for diagnostic items when altering the design of Experiment 1 to facilitate associative learning. This suggested that reliable stimulus–response associations also preclude list-level control in the letter flanker task. This finding raised the possibility that list-level control could be difficult to observe in the other popular variant of the flanker task, which uses arrows: The relatively small stimulus/response set of four items (arrows) necessarily means that if nonoverlapping inducer and diagnostic sets were created as in the ABS design, reliable stimulus–response associations would be present in the lists (as the inducer set would include only two items). In Experiment 3, we developed a modified ABS design to examine list-level control in the arrow flanker task.

## Experiment 1

The purpose of Experiment 1 was to examine whether an LWPC effect would be found for diagnostic items in a letter flanker task under conditions previously shown to produce list-level control in the Stroop task (Bugg, 2014a). As described earlier, this experiment used the ABS design, with four letters serving the role of inducer items that set the bias of the list to be mostly compatible or mostly incompatible and two distinct letters serving the role of diagnostic items that are 50% congruent, allowing for assessment of pure list-level control. Inducer items occurred disproportionately more frequently than

**Table 1.** Frequency of trial types as a function of task block, for Experiments 1 and 2.

Experiment	Task block	Flankers	Target						
			S	D	J	K	F	L	
Experiment 1	LWmc	S	36	2	2	2			
		D	2	36	2	2			
		J	2	2	36	2			
		K	2	2	2	36			
		F					24	24	
		L					24	24	
	LWmi	S	6	12	12	12			
		D	12	6	12	12			
		J	12	12	6	12			
		K	12	12	12	6			
		F					24	24	
		L					24	24	
	Experiment 2	LWmc	S	36	6				
			D	6	36				
J					36	6			
K					6	36			
F							24	24	
L							24	24	
LWmi		S	6	36					
		D	36	6					
		J			6	36			
		K			36	6			
		F					24	24	
		L					24	24	

LWmc: list-wide mostly compatible (73% compatible); LWmi: list-wide mostly incompatible (27% compatible).

Four letters played the role of the inducer items (86% compatible or 14% compatible, respectively) and two other letters played the role of the diagnostic (50% compatible) items. Displayed is one possible counterbalance in which targets F and L are the diagnostic set (in boldface in this example).

unbiased items (as is typical in the ABS design; Bugg, 2014a), so that when intermixed the list-wide proportion congruence was still sufficiently biased (meaning mostly compatible [73%] or mostly incompatible [27%] in this experiment; see Table 1). Importantly, stimulus–response mappings were natural rather than arbitrary (participants pressed the keyboard letter corresponding to the target letter presented on-screen) to decrease the potential working memory load of maintaining six different stimulus–response mappings, which could have interfered with list-level control.

## Method

**Participants.** Sample size was defined based on an a priori power analysis (see <https://osf.io/9afk8>). The lowest reported effect size in similar paradigms that produced an LWPC effect is .190 (Bugg, 2014a; Experiment 1); a power analysis using G\*Power 3.1.9.4 indicated a required

sample size of 48 subjects to attain a statistical power of .90 for this effect size. A sample of  $N=48$  participants (34 females, 14 males) completed the study. All were college students aged 18–25 (mean age = 19.17,  $SD=1.17$ ) with normal or corrected vision.

**Design and stimuli.** The letters S, D, F, J, K, and L were used to create stimuli. Compatible trials were composed of letter strings in which the central target letter matched the flanker letters (e.g., SSSSS), whereas incompatible trials comprised a central target letter that conflicted with the identity of the flanker letters (e.g., SSJSS). LWPC was manipulated within subjects such that each participant performed the flanker task in an LWmc (73% compatible) list and an LWmi (27% compatible) list. Order was counterbalanced between subjects.

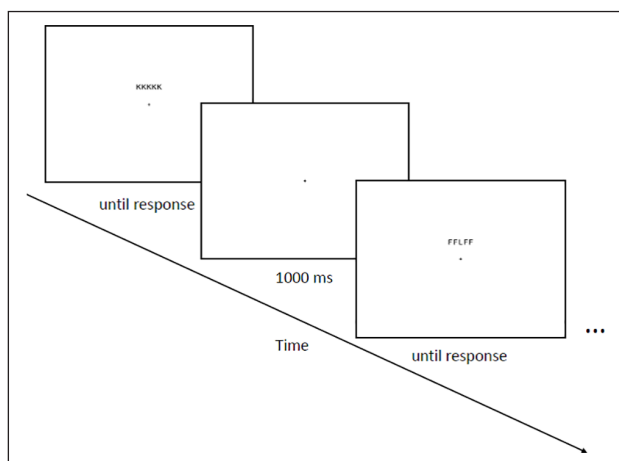
The set of inducer items was composed of four target letters and the diagnostic set was composed of the remaining two letters. For example, if the letters S, D, J, and K were assigned to be inducer items (86% compatible in the mostly compatible [LWmc] list and 14% compatible in the mostly incompatible [LWmi] list), then F and L were assigned to be diagnostic items (50% compatible regardless of the list; see Table 1 for the frequency of trials for each item set). As shown in Table 1, these sets did not overlap: Targets from the inducer set could appear with any of the three other letters from the inducer set, but not with a letter from the diagnostic set (and vice versa). Assignment of letters to the inducer and diagnostic sets was counterbalanced across participants.

**Procedure.** The experiment comprised a stimulus–response mapping practice phase, a flanker practice phase, and two blocks (i.e., LWmc and LWmi lists) of the flanker task. During the mapping practice phase, participants viewed a single letter on screen and were told to press the key that corresponded to the letter. Response keys were naturally mapped on the keyboard (e.g., participants pressed the “S” key to respond to the target letter S, the “J” key for the target letter J). Participants positioned their hands on the keyboard just as if they were typing (e.g., the ring finger of their left hand on the “S” key, the index finger of their right hand on the “J” key). Each of the six letters was presented eight times for a total of 48 mapping practice trials. Participants were given corrective feedback on incorrect trials during the mapping phase. Participants then began the flanker practice phase. Participants were instructed to press the key that corresponded to the central letter as quickly and accurately as possible. As shown in Figure 1, the letter strings were composed of five letters (one central target and four flankers, two on each side) and were presented centrally until a response was detected. On each trial, a letter string appeared with a fixation cross below the central target letter until a response was made, and then the fixation cross remained on an otherwise blank screen

for 1,000 ms following each response. The next stimulus appeared immediately thereafter. Following 12 practice trials that were 50% congruent, participants completed the two blocks of the flanker task, with each block comprising 264 trials. Participants were given a short break after 132 trials in each block. Trials were presented in a random order without replacement (see Table 1 for the frequency of each trial type). RT and accuracy were recorded on each trial.

## Results

Average RTs were computed on correct trials only. All trials with RTs lower than 200 ms or higher than 2,000 ms were dropped from the analysis (see, for example, Bugg,



**Figure 1.** Sample compatible and incompatible displays used in Experiments 1 and 2. Each set of letters was displayed on screen until a response was detected. A central fixation cross appeared below the flanker stimuli and during the otherwise blank 1,000 ms response-to-stimulus interval.

2015); this eliminated less than 1% of trials in all conditions. We report the analyses of the inducer set first and then separately the critical analyses for the diagnostic set (e.g., Bugg, 2014a; Gonthier et al., 2016). One subject failed to comply with the task (error rate >90% on incompatible trials) and was removed, yielding a final sample size of  $N=47$ . Descriptive statistics are presented in Table 2. Data files can be accessed via the Open Science Framework platform at <https://osf.io/9afk8>.

**RT.** For inducer items, a 2 (LWPC)  $\times$  2 (trial type) within-subjects analysis of variance (ANOVA) yielded a large main effect of trial type, confirming the compatibility effect,  $F(1, 46)=119.08$ , mean square error (MSE)=2,497,  $p < .001$ ,  $\eta_p^2 = .72$ . This main effect was qualified by a significant two-way interaction,  $F(1, 46)=66.49$ , MSE=1,410,  $p < .001$ ,  $\eta_p^2 = .59$ , indicating a reduced compatibility effect in the LWmi block.

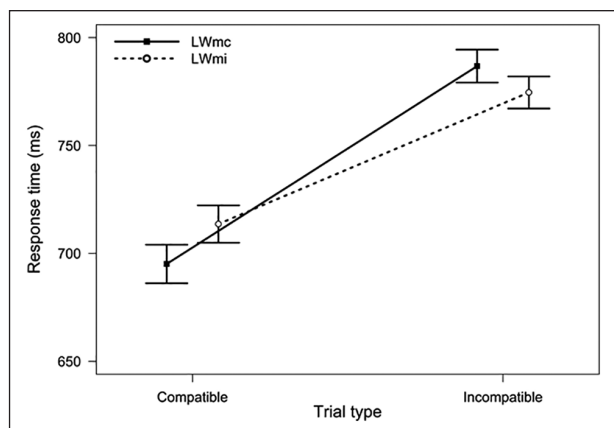
The key analysis was performed on the diagnostic items. The 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA again yielded a compatibility effect,  $F(1, 46)=150.87$ , MSE=1,816,  $p < .001$ ,  $\eta_p^2 = .77$ , as well as a significant two-way interaction,  $F(1, 46)=6.84$ , MSE=1,624,  $p = .012$ ,  $\eta_p^2 = .13$ , indicating as predicted a smaller compatibility effect in the LWmi block ( $M=61$  ms,  $SD=56$  ms) than in the LWmc block ( $M=92$  ms,  $SD=161$  ms) (see Figure 2).

**Error rate.** A 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA for inducer items revealed a marginal main effect of trial type,  $F(1, 46)=3.67$ , MSE=0.002,  $p = .061$ ,  $\eta_p^2 = .07$ , as well as a marginally smaller compatibility effect in the LWmi block,  $F(1, 46)=3.88$ , MSE=0.001,  $p = .055$ ,  $\eta_p^2 = .08$ . The 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA on the diagnostic items revealed a significant main effect of trial type,  $F(1, 46)=18.21$ , MSE=0.001,  $p < .001$ ,  $\eta_p^2 = .28$ , but a nonsignificant

**Table 2.** Descriptive statistics for Experiment 1: response times and error rates as a function of task block (LWPC), item type, and compatibility.

Item type	Task block					
	LWmc			LWmi		
	COM	INC	CE	COM	INC	CE
<b>RTs</b>						
Inducer items	714 (160)	838 (177)	124 (60)	766 (176)	801 (166)	35 (65)
Diagnostic items	695 (155)	787 (163)	92 (161)	714 (158)	775 (157)	61 (56)
<b>Error rates</b>						
Inducer items	0.031 (0.029)	0.051 (0.060)	0.020 (0.048)	0.043 (0.053)	0.047 (0.042)	0.004 (0.053)
Diagnostic items	0.033 (0.033)	0.056 (0.051)	0.024 (0.053)	0.028 (0.043)	0.046 (0.054)	0.018 (0.033)

LWPC: list wide proportion congruence; LWmc: list-wide mostly compatible (73% compatible); LWmi: list-wide mostly incompatible (27% compatible); COM: compatible trials; INC: incompatible trials; CE: compatibility effect computed as incompatible-compatible; RTs: response times. Average values with standard deviations in parentheses. Inducer items were 86% compatible in the LWmc block and 14% compatible in the LWmi block; diagnostic items were 50% compatible in both blocks.



**Figure 2.** Mean response time in Experiment 1 for diagnostic items, as a function of task block and trial type. Error bars represent within-subjects standard errors of the mean (Morey, 2008).

two-way interaction,  $F(1, 46) = 0.39$ ,  $MSE = 0.001$ ,  $p = .536$ ,  $\eta_p^2 = .01$ .

## Discussion

The key finding of Experiment 1 was an LWPC effect for diagnostic items in a letter flanker task. The compatibility effect was significantly reduced in mostly incompatible compared with mostly compatible lists. This represents the first evidence of list-level control in a flanker task using diagnostic items, which makes it possible to rule out item-specific mechanisms and bottom-up priming of attention as explanations for better performance in mostly incompatible lists (e.g., Braem et al., 2019; cf. Wendt et al., 2012). This evidence demonstrates that list-level control is a viable mechanism that facilitates performance in tasks with spatial conflict (flanker), suggesting that theoretical accounts positing dual mechanisms (Braver et al., 2007) or multiple levels of control (Bugg, 2012) may be general and not task-specific.

## Experiment 2

The findings of Experiment 1 converged with patterns observed previously in the Stroop task using the ABS design (Bugg, 2014a). The ABS design was inspired by the AATC account (Bugg, 2014a), which also predicts that participants may not engage in list-level control if they can instead utilise simpler, stimulus–response learning to achieve high levels of performance (e.g., responding quickly to incompatible items in mostly incompatible lists). Interestingly, however, some findings suggest that the AATC account may not apply to all paradigms: In one prior experiment, list-level control was observed on diagnostic items in a Simon task even when participants could predict responses on inducer items, which comprised only two items thereby enabling participants to utilise stimulus–response learning (Wühr et al.,

2015). The same finding was observed in another experiment using a similar design in a prime-probe task (Schmidt, 2016). It is uncertain what could create such a discrepancy with the Stroop task, but these results suggest that preventing use of stimulus–response associations to guide responding on incongruent trials (by using an inducer set of four rather than two items) may not be strictly necessary to allow for list-level control in some tasks. In turn, this has consequences for the flanker task, and especially for arrow-based versions of the task which do not include the six different stimuli necessary to implement the ABS design. In Experiment 2, we decided to directly test whether the AATC account applies to the flanker task and in so doing, better characterise the boundary conditions under which list-level control may or may not be observed.

## Method

**Participants.** Sample size was defined based on the same a priori power analysis performed for Experiment 1, which indicated a required sample size of 48 subjects to attain a statistical power of .90 (see <https://osf.io/9afk8>). A sample of  $N = 48$  participants (34 females, 14 males) completed the study. All were college students aged 18–25 (mean age = 19.75,  $SD = 1.63$ ) with normal or corrected vision.

**Design and stimuli.** The design and stimuli were identical to Experiment 1, with one exception: Inducer items comprised two nonoverlapping sets of two letters, whereas diagnostic items comprised the two remaining letters as in Experiment 1. In other words, unlike in Experiment 1, each of the four letters in the inducer set was paired with only one other letter (see Table 1).

**Procedure.** The procedure was identical to Experiment 1 (see Figure 1).

## Results

Preprocessing was identical to Experiment 1; no subject was excluded in Experiment 2. Descriptive statistics are presented in Table 3. Data files can be accessed via the Open Science Framework platform at <https://osf.io/9afk8>.

**RT.** For inducer items, a 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA yielded a large main effect of trial type, confirming the compatibility effect,  $F(1, 47) = 198.60$ ,  $MSE = 1,420$ ,  $p < .001$ ,  $\eta_p^2 = .81$ . This main effect was qualified by a significant two-way interaction,  $F(1, 47) = 38.66$ ,  $MSE = 1,430$ ,  $p < .001$ ,  $\eta_p^2 = .45$ , indicating a reduced compatibility effect in the LWmi block.

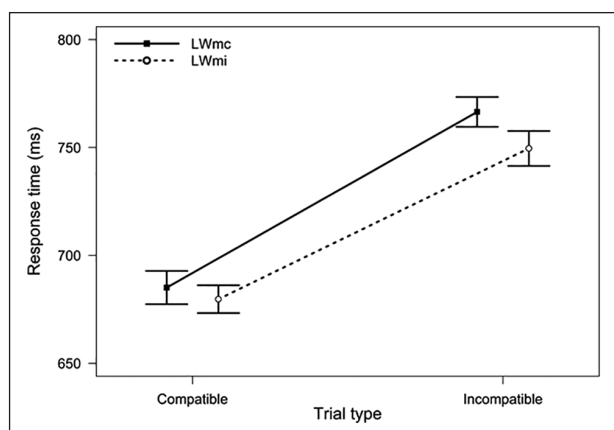
The key analysis was again performed on diagnostic items. The 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA again yielded a compatibility effect,  $F(1, 47) = 163.66$ ,  $MSE = 1,678$ ,  $p < .001$ ,  $\eta_p^2 = .78$ , but a nonsignificant

**Table 3.** Descriptive statistics for Experiment 2: response times and error rates as a function of task block (LWPC), item type, and compatibility.

Item type	Task block					
	LWmc			LWmi		
	COM	INC	CE	COM	INC	CE
<b>RTs</b>						
Inducer items	693 (124)	804 (152)	111 (51)	716 (112)	758 (122)	42 (56)
Diagnostic items	685 (123)	767 (124)	82 (49)	680 (108)	750 (126)	70 (50)
<b>Error rates</b>						
Inducer items	0.048 (0.048)	0.067 (0.074)	0.019 (0.052)	0.038 (0.049)	0.064 (0.052)	0.026 (0.050)
Diagnostic items	0.047 (0.051)	0.060 (0.066)	0.013 (0.052)	0.041 (0.048)	0.045 (0.051)	0.004 (0.046)

LWPC: list-wide proportion congruence; LWmc: list-wide mostly compatible (73% compatible); LWmi: list-wide mostly incompatible (27% compatible); COM: compatible trials; INC: incompatible trials; CE: compatibility effect computed as incompatible–compatible; RTs: response times.

Average values with standard deviations in parentheses. Inducer items were 86% compatible in the LWmc block and 14% compatible in the LWmi block; diagnostic items were 50% compatible in both blocks.

**Figure 3.** Mean response time in Experiment 2 for diagnostic items, as a function of task block and trial type. Error bars represent within-subjects standard errors of the mean (Morey, 2008).

two-way interaction,  $F(1, 47)=2.06$ ,  $MSE=780$ ,  $p=.158$ ,  $\eta_p^2=.04$ , indicating no significant difference between the compatibility effect in the LWmi block ( $M=70$ ms,  $SD=50$ ms) and the LWmc block ( $M=82$ ms,  $SD=49$ ms) (see Figure 3). The corresponding Bayes factor was  $BF_{01}=3.637$  against the interaction, indicating moderate (Lee & Wagenmakers, 2013) evidence in favour of the null. In other words, there was no LWPC effect for diagnostic items in this experiment.

**Error rate.** A 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA for inducer items revealed a main effect of trial type,  $F(1, 47)=20.31$ ,  $MSE=0.001$ ,  $p<.001$ ,  $\eta_p^2=.30$ , but the compatibility effect was not significantly smaller in the LWmi block,  $F(1, 47)=0.42$ ,  $MSE=0.001$ ,  $p=.520$ ,  $\eta_p^2=.01$ . The 2 (LWPC)  $\times$  2 (trial type) within-subjects ANOVA on the subset of diagnostic items revealed neither a main effect of trial type,  $F(1, 47)=2.76$ ,  $MSE=0.001$ ,

$p=.103$ ,  $\eta_p^2=.06$ , nor a two-way interaction,  $F(1, 47)=0.81$ ,  $MSE=0.001$ ,  $p=.373$ ,  $\eta_p^2=.02$ .

### Discussion

The purpose of Experiment 2 was to determine whether an LWPC effect would be observed for diagnostic items in the letter flanker task if the design were modified to enable participants to use simple stimulus–response learning to predict responses on most trials (inducer items). Consistent with our predictions based on the AATC account and with prior findings in the Stroop task (Bugg, 2014a), we did not observe an LWPC effect for diagnostic items and the Bayesian analysis suggested moderate evidence in favour of the null. These findings confirm that it is important to consider the potential for stimulus–response learning when evaluating evidence for list-level control in the flanker task (contrary to other paradigms such as the Simon and prime-probe tasks, where such a consideration might be less critical; Schmidt, 2016; Wühr et al., 2015). When the situation encourages associative learning as in the current experiment, list-level control may not be observed even though it is a viable mechanism that participants otherwise use to facilitate performance in the flanker task (Experiment 1).

The findings of Experiment 2 have important implications for investigating list-level control in the arrow flanker task, the other popular flanker task variant. Directly implementing the ABS design in the task, with two diagnostic items and four inducer items, is not possible given that the arrow flanker task has maximally four items (arrows). Making the set of 50% diagnostic items entirely separate from the inducer items (so that a target arrow from the diagnostic set can never appear with flankers from the inducer set, and vice versa) necessarily requires introducing reliable stimulus–response associations (such that



responses on inducer items could be predicted based on the flankers). As demonstrated by Experiment 2, such associations may preclude use of list-level control. In Experiment 3, we explored whether an alternative design may enable the observation of list-level control in the arrow flanker task.

### Experiment 3

Experiments 1 and 2 demonstrated that list-level control can be observed in a flanker task, and that its use may be limited to a condition in which use of simpler, stimulus–response learning is not a viable alternative strategy for achieving high levels of performance. This conclusion has important implications for pursuing evidence for list-level control in the arrow flanker task which is composed of only four different stimuli. The standard ABS design cannot be used because each set (inducer and diagnostic) would comprise just two items, and participants could therefore capitalise on stimulus–response associations in the inducer set. In Experiment 3, we propose and test a potential solution: use of a *modified* ABS design.

As in the original ABS design (Bugg, 2014a), the modified ABS design proposed here for the arrow flanker task comprises sets of inducer and diagnostic items. However, unlike the original design, there are only two instead of four items in the inducer set. Assignment of items to sets is based on the identity of the relevant dimension (i.e., target; as in item-specific proportion congruence studies where this is done to discourage reliance on stimulus–response learning; Bugg & Dey, 2018; Bugg et al., 2011; Bugg & Hutchison, 2013). For example, if left and right target arrows served the role of inducer items (signalling 75% and 25% proportion congruence for items comprising this feature in the mostly compatible and mostly incompatible lists, respectively), then up and down target arrows served the role of diagnostic items (signalling 50% proportion congruence regardless of the list).

Critically, to circumvent the tendency for participants to rely on simple associative learning instead of using list-level control, as occurs in the letter flanker task (Experiment 2) and in the Stroop task when the inducer set is composed of only two items (e.g., Blais & Bunge, 2010; Bugg, 2014a; Bugg et al., 2008), this modified design allows the inducer and diagnostic sets to overlap. For example, incompatible trials in the inducer set include left and right target arrows not only with right and left flankers, respectively, but also with up and down flankers. As a consequence, participants cannot minimise conflict on most trials by predicting highly contingent responses (because there are three equally likely incompatible response options).

This modified design also dictates an alternative analysis strategy. To ensure that the results are not biased by overlap between the two sets, analysis of the LWPC effect

has to be performed selectively on the subset of diagnostic items that does not include a biased feature from the inducer set (i.e., feature that is predictive of the probability of encountering conflict for these items). Using the above example, if left and right target arrows serve as inducer items (i.e., are biased) and up and down target arrows as diagnostic items, then the analysis examining transfer of the LWPC effect would be performed on the *critical subset of diagnostic items* comprising solely up and down arrows (see bolded and underlined items in Table 4). That is, the analysis would *not* include any diagnostic items that have a left or right arrow in any position including flankers (e.g., <<<<^<<<, ^^^^<^^^<) to prevent any possible feature-based priming of the attentional setting associated with the inducer items (e.g., possibility that a left or right flanker arrow or target arrow could prime the biased attentional setting associated with left or right target arrows, thereby leading to a relaxation of attention for these items in the mostly compatible list and a heightening of attention in the mostly incompatible list, and thus a spurious LWPC effect for diagnostic items driven by features from the inducer set).

In other words, in a conflict task with only four stimuli and response possibilities, the modified ABS design proposed here retains key elements of the ABS design that are important for drawing conclusions about list-level control: (a) use of diagnostic items that are matched in PC and frequency across mostly compatible and mostly incompatible lists, (b) analysis of diagnostic items performed on PC and frequency-matched items that do not include a biased feature from the inducer set (in either the flanker or target position), (c) at least two equally contingent response possibilities on incompatible trials in the inducer set. The first two elements are necessary for valid inference (attributing LWPC effect for diagnostic items to list-level control). The third should ideally be retained because contingencies may preclude list-level control in the flanker task, as demonstrated in Experiment 2 of this study and in the Stroop task (Bugg, 2014a; see also Blais & Bunge, 2010; Bugg et al., 2008). An advantage of retaining the third element regardless of the task relates to interpretation of a null LWPC effect for diagnostic items. If at least two equally contingent responses are possible for incongruent trials in the inducer set, one can rule out that reliance on contingency learning precluded list-level control and entertain alternative interpretations (i.e., that list-level control may not be a mechanism used to guide performance in certain tasks or under certain conditions). Thus, the present experiment provided an opportunity to test a potential design that could be used not just in the flanker task, but that could make it easier to isolate list-level control in any conflict task that has a necessarily limited stimulus (and response) set (e.g., other flanker tasks; spatial Stroop tasks, Logan & Zbrodoff, 1979; certain Simon tasks).

**Method**

**Participants.** All participants ( $N=20$ ) were right-handed college students aged 18–25 years with normal or corrected vision. Sample size was defined based on Bugg (2014a), which included 16–18 subjects per *between-subjects* LWPC condition when demonstrating list-level control in the Stroop task. Because LWPC was manipulated within subjects in the current experiment and because the LWPC manipulation was stronger than in Bugg’s study (75% vs. 25% congruent here compared with 67 vs. 33% congruent in the prior study; see the “Design and stimuli” section), a sample size of 20 was expected to yield sufficient power to observe list-level control.

**Design and stimuli.** Compatible trials were composed of arrow strings in which the central target arrow matched the flanker arrows (e.g., <<<<<<<<), whereas incompatible trials comprised a central target arrow that conflicted with the identity of the flanker arrows (e.g., >>><>>>). There were seven arrows in each string with the three peripheral arrows on each side considered the flankers. LWPC was manipulated within subjects such that each participant performed the flanker task in an LWmc (75% compatible) list and an LWmi (25% compatible) list. Order was counterbalanced between subjects.

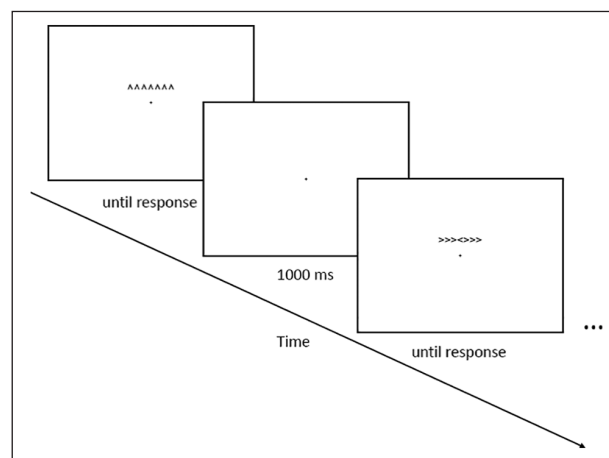
The relevant dimension was used to define sets of items as inducer items and diagnostic items (i.e., to signal PC): one set was composed of left and right targets and the other was composed of up and down targets.<sup>5</sup> In other words, if left and right targets were assigned to be inducer items (86% compatible in the mostly compatible [LWmc] list and 14% compatible in the mostly incompatible [LWmi] list), then up and down targets were assigned to be diagnostic items (50% compatible regardless of the list; see Table 4 for the frequency of trials for each item set). As shown in Table 4, these sets overlapped such that, for example, a left or right target could appear with an up or down flanker. Assignment of targets to the role of inducer or diagnostic items was counterbalanced across participants.

**Procedure.** The experiment comprised two blocks (i.e., LWmc and LWmi lists) of the flanker task with each block comprising 240 trials. Participants were instructed to press the key that corresponded to the direction the central target arrow was pointing while ignoring the flanking arrows, and to make their responses as quickly and accurately as possible. Participants used the 2, 4, 6, and 8 keys on the number pad to indicate a down, left, right, or up central arrow, respectively. (Given the natural mapping between stimuli and responses, no practice phase was included in this experiment.) Participants used the index finger of their right hand to press one of the four response keys and rested this finger on the 5 key between responses. As shown in Figure 4, the arrow strings were composed of seven arrows

**Table 4.** Frequency of trial types as a function of task block (LWPC) for Experiment 3.

Task block	Flankers	Target			
		<	>	^	v
LWmc	<	72	4	<b>6</b>	<b>6</b>
	>	4	72	<b>6</b>	<b>6</b>
	^	4	4	<u>18</u>	<u>6</u>
	v	4	4	<u>6</u>	<u>18</u>
LWmi	<	12	24	<b>6</b>	<b>6</b>
	>	24	12	<b>6</b>	<b>6</b>
	^	24	24	<u>18</u>	<u>6</u>
	v	24	24	<u>6</u>	<u>18</u>

LWPC: list-wide proportion congruence; LWmc: list-wide mostly compatible (75% compatible); LWmi: list-wide mostly incompatible (25% compatible). The distinction between the inducer (86% compatible or 14% compatible, respectively) and diagnostic (50% compatible) set of stimuli was defined the target identity. Diagnostic items are in boldface in this example, and the critical subset of diagnostic items that does not include the inducer item feature (in the flanker or target position) is underlined. Displayed is one possible counterbalance in which left/right targets are inducer items and up/down targets are diagnostic items.



**Figure 4.** Sample compatible and incompatible displays used in Experiment 3. Each set of arrows was displayed on screen until a response was detected. A central fixation cross appeared below the arrows and during the otherwise blank 1,000 ms response-to-stimulus interval.

(one central target and six flankers, three on each side) and were presented centrally until a response was detected. On each trial, an arrow string appeared with a fixation cross below the central target arrow until a response was made, and then the fixation cross remained on an otherwise blank screen for 1,000 ms following each response. The next stimulus appeared immediately thereafter. Participants were given a short break after 120 trials in each block. Trials were presented in a random order without replacement (see Table 4 for the frequency of each trial type). RT and accuracy were recorded on each trial.

**Table 5.** Descriptive statistics for Experiment 3: response times and error rates as a function of task block (LWPC), item type, and compatibility.

Item type	Task block							
	LWmc				LWmi			
	COM	INC (all trials)	INC (same set)	CE	COM	INC (all trials)	INC (same set)	CE
<b>RTs</b>								
Inducer items	461 (53)	537 (57)	567 (67)	106 (49)	475 (66)	517 (62)	547 (60)	72 (43)
Diagnostic items	471 (53)	535 (53)	599 (98)	128 (88)	486 (59)	536 (54)	569 (65)	83 (52)
<b>Error rates</b>								
Inducer items	0.001 (0.004)	0.015 (0.020)	0.038 (0.059)	0.037 (0.057)	0.002 (0.009)	0.010 (0.014)	0.022 (0.033)	0.020 (0.035)
Diagnostic items	0.003 (0.009)	0.015 (0.023)	0.029 (0.062)	0.026 (0.064)	0.003 (0.009)	0.008 (0.013)	0.012 (0.031)	0.009 (0.033)

LWPC: list-wide proportion congruence; LWmc: list-wide mostly compatible (73% compatible); LWmi: list-wide mostly incompatible (27% compatible); COM: compatible trials; INC (all trials): all incompatible trials including trials comprising targets and flankers from different arrow sets (e.g., up target with left flankers); INC (same set): incompatible trials including only trials with flankers belonging to the same arrow set as the target (i.e., if target is left/right, flankers are right/left); CE: compatibility effect computed as incompatible (same set)–compatible; RTs: response times. Average values with standard deviations in parentheses. Inducer items were 86% compatible in the LWmc block and 14% compatible in the LWmi block; transfer items were 50% compatible in both blocks.

**Analytical approach.** As explained in the introduction to this experiment, to assess transfer of the LWPC effect, we restricted the analysis to the critical subset of diagnostic items that did not include a biased feature from the inducer set (in either the flanker or target position). Accordingly, if items with up/down target arrows served as diagnostic items, mean performance on incompatible trials was calculated exclusively for trials that comprised up targets with down flankers and down targets with up flankers. Mean performance on compatible trials was necessarily derived from trials comprised entirely of up or down arrows. Note that because the diagnostic set is 50% compatible (see Table 4 bolded cells) and because this analysis excludes incompatible trials with left or right arrows, the critical subset of diagnostic items includes more compatible trials than incompatible trials, equally in each list (see Table 4 bolded and underlined cells; cf. Hutchison, 2011). This analytic decision is invisible to the participant: From their point of view, items in the diagnostic set are 50% compatible, and when incompatible these items are associated equally often with the three other responses.

For consistency, analyses of inducer items were also restricted to incompatible trials from the same arrow set (e.g., if items with left/right target arrows served as inducer items, mean performance on incompatible trials was calculated exclusively for trials that comprised left targets with right flankers and right targets with left flankers). Compatible trials for inducer items were necessarily limited to trials comprised entirely of left or right arrows, in this example.

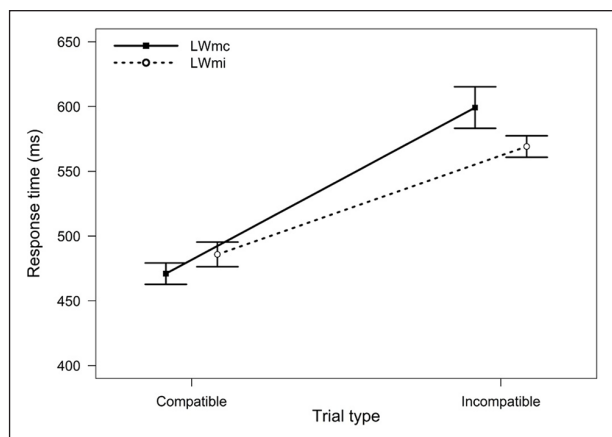
## Results

Average RTs were computed on correct trials only. As in Experiments 1 and 2, all trials with RTs lower than 200 ms or higher than 2,000 ms were dropped from the analysis; this eliminated less than 1% of trials in all conditions. Descriptive statistics are presented in Table 5. We report the restricted analyses of the inducer set first, and then separately the critical analyses for the diagnostic set. Data files can be accessed via the Open Science Framework platform at <https://osf.io/9afk8>.

**RT.** For inducer items, the main effect of trial type was significant,  $F(1, 19)=112.44$ ,  $MSE=1,417$ ,  $p<.001$ ,  $\eta_p^2=.86$ , confirming slowing on incompatible compared with compatible trials (i.e., compatibility effect). This main effect was qualified by a significant two-way interaction,  $F(1, 19)=8.09$ ,  $MSE=678$ ,  $p=.010$ ,  $\eta_p^2=.30$ , indicating a reduced compatibility effect in the LWmi block.

The key analysis investigated the critical subset of diagnostic items. The main effect of trial type was significant,  $F(1, 19)=57.75$ ,  $MSE=3,876$ ,  $p<.001$ ,  $\eta_p^2=.75$ ; most importantly, the two-way interaction was also significant,  $F(1, 19)=7.58$ ,  $MSE=1,333$ ,  $p=.013$ ,  $\eta_p^2=.29$ . This interaction indicated a reduced compatibility effect in the LWmi block ( $M=83$  ms,  $SD=52$ ) when compared with the LWmc block ( $M=128$  ms,  $SD=88$ ), in line with the hypothesis of list-level control. This pattern is represented in Figure 5.

**Error rate.** The main effect of trial type was significant for inducer items,  $F(1, 19)=11.20$ ,  $MSE=0.001$ ,  $p=.003$ ,



**Figure 5.** Mean response time in Experiment 3 for the critical subset of diagnostic items, as a function of task block and trial type. Error bars represent within-subjects standard errors of the mean (Morey, 2008).

$\eta_p^2 = .37$ , and for the critical subset of diagnostic items,  $F(1, 19) = 5.09$ ,  $MSE = 0.001$ ,  $p = .036$ ,  $\eta_p^2 = .21$ , indicating compatibility effects in all cases. However, the compatibility effect was not modulated by PC for either inducer items,  $F(1, 19) = 1.51$ ,  $MSE = 0.001$ ,  $p = .235$ ,  $\eta_p^2 = .07$ , or the critical subset of diagnostic items,  $F(1, 19) = 1.06$ ,  $MSE = 0.001$ ,  $p = .316$ ,  $\eta_p^2 = .05$ .

## Discussion

The key finding in Experiment 3 was the observation of a significant LWPC effect for the subset of diagnostic items that did not share a biased feature from the inducer set (including in a flanker position). This finding demonstrates the feasibility of observing an LWPC effect for unbiased diagnostic items in an arrow-based flanker task with a relatively limited number of stimuli and response options. Allowing the two sets of two items to overlap and restricting data analysis to the subset of diagnostic trials where the inducer and diagnostic items did not overlap appeared to be the most viable strategy for assessing list-level control in the context of the AATC framework in this type of task. This approach adequately controls for differences in frequency and PC at the item level (the subset of diagnostic items is presented equally frequently in the mostly incompatible and mostly compatible lists, and the trials within this subset are matched in PC across the two lists), and for the possibility that mechanisms other than list-level control (e.g., feature-based priming) affect performance on trials where features from the inducer set appear alongside features from the diagnostic set.

However, as a reviewer also pointed out, this approach does not quite control for all possible confounds. This modified design confounds the LWPC manipulation with the proportion congruency of the stimulus dimension that does not

define the item sets (the flankers). In other words, as can be seen in the example in Table 4, the two target arrows forming the set of diagnostic items (up and down arrows) also appear as incompatible flankers for items in the inducer set (e.g., a left target surrounded by up flankers), and this happens more frequently in the mostly incompatible block (24 instances) than in the mostly compatible block (four instances). This could conceivably introduce a bias: The frequency of encountering up and down arrows from the diagnostic set as incongruent flankers on inducer trials may affect performance on the critical subset of diagnostic items where the up and down arrows appear as targets. The confound is clearly not one of contingency, as the up and down flankers are not more predictive of the correct answer in the mostly incompatible block relative to the mostly compatible block; rather, the confound may facilitate the reduction in the compatibility effect in the mostly incompatible block for the diagnostic subset, by allowing participants to reactively use information about the higher likelihood that up and down arrows appear in an incompatible trial requiring cognitive control.

Due to this confound, it remains an open question whether a pure list-level control mechanism can be observed in paradigms like the arrow flanker task that include only four different stimuli. Unfortunately, it does not seem possible to achieve a better design. This modified form of the ABS design thus currently represents the most adequate way to investigate list-level control in tasks with a limited number of stimuli.

## General discussion

The present set of experiments aimed to determine whether existing theoretical accounts (e.g., dual mechanisms, Braver et al., 2007; multiple levels, Bugg, 2012; associations as antagonists to top-down control [AATC], Bugg, 2014a) generalise to tasks other than the Stroop, by examining the conditions under which list-level control may be found in flanker tasks based on an established LWPC manipulation involving diagnostic items (i.e., the ABS design). A secondary aim was to evaluate the fruitfulness of a modified ABS design and analytical approach for revealing list-level control in a conflict task limited to four different stimuli (arrow flanker task).

The first key finding was observed in Experiment 1, where an LWPC effect was found for diagnostic items in the letter flanker task. The compatibility effect, which indexes the extent to which participants selectively focus on the central target and ignore the flanking distractors in the periphery, was reduced in the mostly incompatible block compared with the mostly compatible block. Critically, this was the case even for diagnostic items, which were matched in frequency and proportion congruency across blocks and did not share features (target or flanker identity) with the inducer items that created the

overall bias (mostly incompatible or mostly compatible) of each block. This finding supports the view that control over flanker interference is not governed exclusively by item-specific mechanisms (item-level control or contingency learning), nor by bottom-up priming of associated attentional sets (Braem et al., 2019). Rather, list-level control appears to be a viable mechanism for minimising interference in the flanker task. In other words, attention towards the spatially separated distractors (flankers) can be modulated based on the global probability of conflict within a list, affecting performance on all items.

Experiment 1 represents the first study to report evidence for list-level control in the flanker task that comes directly from evaluating performance on the flanker task. This evidence complements that provided by Wendt et al. (2012) in the form of search task performance on search trials interspersed within a letter-based flanker paradigm. Together these studies demonstrate that list-level control is a viable mechanism that influences the magnitude of compatibility effects in flanker tasks, thereby supporting the generalisability of current theoretical accounts to the flanker task. Taken together with the evidence across several Stroop studies, the present findings are consistent with accounts of cognitive control that acknowledge that control operates via multiple mechanisms including proactive control (Braver et al., 2007) or at multiple levels (Bugg, 2012), including the list level. In conjunction with the designs used to demonstrate item-level control in the letter flanker task (Bugg, 2015), the present ABS design makes it possible to pit list-level and item-level control head-to-head, as has been done in the Stroop task to determine their relative costs and benefits (Gonthier et al., 2016). This would offer yet another opportunity to contrast patterns across tasks and refine theoretical accounts.

The second key finding emerged from Experiment 2. List-level control was not observed in the letter flanker task when the design was modified to enable participants to rely on stimulus–response associations to guide performance. This was done by using a variant of the ABS design previously employed in the Stroop task, where the inducer set was broken into two pairs of items that appeared only with each other (Bugg, 2014a). This meant that on the inducer trials representing the majority of trials within the list, participants could bypass control and predict the highly contingent response (e.g., in the mostly incompatible block, if S and J were one pair of inducer items, when participants encountered an S in the flanker position, they could accurately predict that J would most frequently be the correct response). Consistent with our AATC-based predictions and with findings previously observed in the Stroop task, the magnitude of the compatibility effect did not differ across mostly incompatible and mostly compatible lists. In other words, the associations appeared to discourage use of list-level control, consistent with the AATC account (Bugg, 2014a).

Experiment 2 speaks to the question of when list-level control is utilised. According to the AATC account, it may be a last resort that participants utilise only when they cannot rely on simpler, stimulus–response learning to achieve high levels of task performance (e.g., predicting responses to incompatible trials in the mostly incompatible lists). The present findings are consistent with this view and reinforce that at least in some tasks, inclusion of inducer items that permit participants to use stimulus–response associations may preclude use and observation of list-level control. An interesting question for future research that emerges based on these findings is why this is the case for the Stroop (Bugg, 2014a) and flanker tasks (Experiments 1 and 2) but it may not be the case for the Simon task (Wühr et al., 2015) or the prime-probe task (Schmidt, 2016). That is, why would the AATC account be applicable to only a subset of conflict tasks?

A methodological aim and contribution of this study was to test a potential solution to the challenge of isolating the contribution of list-level control to performance on tasks for which the presence of stimulus–response associations is difficult to avoid because they have a limited number of stimuli (and responses). One such task is the other commonly used flanker task involving up, down, left, and right arrows. The solution we tested in Experiment 3 was a modified ABS design and corresponding analytical approach that retained elements that are important for drawing conclusions about list-level control (cf. Braem et al., 2019): (a) use of diagnostic items that are matched in PC and frequency across mostly compatible and mostly incompatible lists, (b) analysis of diagnostic items is performed on PC and frequency-matched items that do not include a biased feature from the inducer set (in either the flanker or target position), and (c) there are at least two equally contingent response possibilities for incompatible trials in the inducer set.

As expected, the modified ABS design used in Experiment 3 elicited an LWPC effect for the critical subset of diagnostic items, and this effect could not be attributed to stimulus–response learning (i.e., differential contingencies across lists). This design makes it possible to investigate list-level control in tasks with smaller stimulus and response sets than the five or six options required for the standard ABS design. This can be useful in paradigms such as the arrow flanker task that do not intrinsically include more than four stimuli; it can also apply to paradigms where the stimulus set is voluntarily restricted because subjects have to respond using a keyboard (e.g., functional magnetic resonance imaging [fMRI] or electroencephalography [EEG] due to the possibility of movement artifacts), especially in populations such as young children where learning more than four different stimulus–response mappings can be challenging.

Such investigations will be critical for refining our theoretical understanding of the defining characteristics of the

various levels at which control can be implemented and dissociations among those levels across a broader range of tasks. We thus anticipate that this modified design may facilitate future investigations of list-level control in broader tasks than the Stroop paradigm in adults. However, as noted above, this design unfortunately does not rule out one particular confound, namely the possibility that participants reactively use information about the likelihood that distractors conflict with the targets on the critical subset of diagnostic trials based on information they learned from inducer trials. For this reason, the modified ABS design with four stimuli falls short of the objective of demonstrating “pure” list-level control. Absent any better solution, this modified design seems to be the best available compromise to investigate list-level control, but care will have to be taken when generalising the conclusions obtained with this approach.

### Limitations

A few limitations merit discussion. First, in Experiments 1 and 2, we devised a letter flanker task that employed non-arbitrarily mapped response keys. Although we have no reason to suspect that there is something special about the six letters/response keys we chose (and although counterbalancing which of the six letters were used for the diagnostic set, as was done, should limit this problem should it exist), we did not examine whether the same pattern would be observed for other letters. Nonetheless, if researchers seek a design and procedure that produces list-level control in a flanker task, the approach adopted in Experiment 1 is a solid option. Second, support for the AATC account and the view that Experiment 2 represents a boundary condition for list-level control was inferred based on a significant LWPC effect in Experiment 1 and a nonsignificant LWPC effect in Experiment 2. Ideally, stronger evidence would be sought by contrasting directly the designs of Experiments 1 and 2 within a single experiment. The challenge is that such a study would require more than 400 subjects to achieve .80 power, assuming a three-way interaction of the size one could expect based on the current data. Third, we followed the approach used in prior studies to index list-level control and this approach uses a design that is not optimised for examining the contribution of congruency sequence effects to the critical LWPC patterns (for reviews, see Bugg, 2017; Bugg & Crump, 2012; see also Braem et al., 2019). In the one study that evaluated such effects, Hutchison (2011) observed LWPC effects on diagnostic items in a Stroop task and found that LWPC effects occurred regardless of preceding trial type and could not be completely explained by congruency sequence effects. In addition, using an alternative approach, researchers have dissociated LWPC effects from congruency sequence effects (Torres-Quesada et al., 2013, 2014). Nonetheless, future research might modify the ABS design

to enable a direct test of the role of congruency sequence effects.

### Conclusion

The current findings support the domain generality of existing theoretical accounts of cognitive control by demonstrating that pure list-level control is observed in a letter flanker task and showing that its use may not be the default but rather depends on whether high levels of task performance can be achieved by a simpler alternative (learning associations between stimuli and responses), in line with the AATC account. A modified version of the established ABS design enabled observation of list-level control in the arrow flanker task. This design may be extended to other tasks with relatively few stimuli and responses where use of stimulus–response learning may preclude list-level control, but it cannot be implemented in a way that is entirely devoid of a confound and the results should be interpreted with care.

### Authors' note

Pre-registrations for Experiments 1 and 2 and data for all experiments can be accessed at <https://osf.io/9afk8>.

### Acknowledgements

The authors would like to thank Maxwell Coll, Mackenzie Glassner, Sonia Trojahn, and Stephanie Zhao for assistance with data collection.



### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Julie M Bugg  <https://orcid.org/0000-0002-0969-186X>  
Corentin Gonthier  <https://orcid.org/0000-0001-8573-0413>

### Notes

1. Another mechanism, temporal learning, has been posited to explain the list-wide proportion congruence effect (LWPC) effect (e.g., Schmidt, 2013); however, recent modelling and experimental evidence challenges this account (Cohen-Shikora et al., 2019; Spinelli et al., 2019).
2. We use this term to denote an LWPC effect that cannot be accounted for by item-specific mechanisms (item-level control or contingency learning) or bottom-up priming of attention. Hereafter, we simply use the term LWPC effect.
3. Compatibility is often used to refer to congruency in a flanker task; we use the term compatibility hereafter.

4. The design, hypotheses, and analytical approach for Experiments 1 and 2 were preregistered on OSF (<https://osf.io/9afk8>). Experiment 3 was conducted prior to Experiments 1 and 2. We thank two anonymous reviewers (who reviewed a prior version of this article that did not include Experiments 1 and 2) for their suggestions to examine list-level control in other flanker tasks besides the arrow task.
5. We also conducted an experiment ( $N=18$ ) in which the irrelevant dimension was used to define sets of items. Although compatibility effects were smaller for diagnostic items in the LWmi (list-wide mostly incompatible) list than the LWmc (list-wide mostly compatible) list, the list-wide PC effect (PC  $\times$  trial type interaction) was not significant ( $p=.087$ ; Bayes factor [BF]=1.23). We cannot be sure whether this is due to low power or a theoretically interesting explanation pertaining to use of the irrelevant dimension to define sets of items.

## References

- Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*, 693–728.
- Blais, C., & Bunge, S. (2010). Behavioral and neural evidence for item-specific performance monitoring. *Journal of Cognitive Neuroscience*, *22*, 2758–2767.
- Blais, C., Harris, M. B., Guerrero, J. V., & Bunge, S. (2012). Rethinking the role of automaticity in cognitive control. *Quarterly Journal of Experimental Psychology*, *65*, 268–276.
- Blais, C., Robidoux, S., Risko, E. F., & Besner, D. (2007). Item-specific adaptation and the conflict monitoring hypothesis: A computational model. *Psychological Review*, *114*, 1076–1086.
- 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., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences*, *23*, 769–783.
- Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 76–106). Oxford University Press.
- Bugg, J. M. (2012). Dissociating levels of cognitive control: The case of Stroop interference. *Current Directions in Psychological Science*, *21*, 302–309.
- Bugg, J. M. (2014a). Conflict triggered top-down control: Default mode, last resort, or no such thing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 567–587.
- Bugg, J. M. (2014b). Evidence for the sparing of reactive cognitive control with age. *Psychology and Aging*, *29*, 115–127.
- Bugg, J. M. (2015). The relative attractiveness of distractors and targets affects the coming and going of item-specific control: Evidence from flanker tasks. *Attention, Perception, & Psychophysics*, *77*(2), 373–389.
- Bugg, J. M. (2017). Context, conflict, and control. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 79–96). John Wiley.
- Bugg, J. M., & Chanani, S. (2011). List-wide control is not entirely elusive: Evidence from picture-word Stroop. *Psychonomic Bulletin & Review*, *18*, 930–936.
- Bugg, J. M., & Crump, M. J. C. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology*, *3*, Article 367.
- Bugg, J. M., & Dey, A. (2018). When stimulus-driven control settings compete: On the dominance of categories as cues for control. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 1905–1932.
- Bugg, J. M., & Hutchison, K. A. (2013). Converging evidence for control of color-word Stroop interference at the item level. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 433–449.
- Bugg, J. M., Jacoby, L. L., & Chanani, S. (2011). Why it is too early to lose control in accounts of item-specific proportion congruency effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 844–859.
- Bugg, J. M., Jacoby, L. L., & Toth, J. (2008). Multiple levels of control in the Stroop task. *Memory & Cognition*, *36*, 1484–1494.
- Chiu, Y. C., Jiang, J., & Egner, T. (2017). The caudate nucleus mediates learning of stimulus-control state associations. *Journal of Neuroscience*, *37*, 1028–1038.
- Cohen-Shikora, E. R., Suh, J., & Bugg, J. M. (2019). Assessing the temporal learning account of the list-wide proportion congruency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 1703–1723.
- Crump, M. J. C., & Milliken, B. (2009). The flexibility of context-specific control: Evidence for context-driven generalization of item-specific control. *Quarterly Journal of Experimental Psychology*, *62*, 1523–1532.
- DePisapia, N., & Braver, T. S. (2006). A model of dual-control mechanisms through anterior cingulate and prefrontal cortex interactions. *Neurocomputing*, *69*, 1322–1326.
- Egner, T. (2014). Creatures of habit (and control): A multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology*, *5*, Article 1247.
- Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*, 1784–1790.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a non-search task. *Perception & Psychophysics*, *16*, 143–149.
- Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*, 340–347.
- Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Doubly dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, *44*, 778–788.
- Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation and responses. *Journal of Experimental Psychology: General*, *121*, 480–506.
- Hutchison, K. A. (2011). The interactive effects of list-wide control, item-based control, and working memory capacity on Stroop performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 851–860.

- Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.
- Kopp, B., Rist, F., & Mattler, U. (1996). N200 in the flanker task as a neurobehavioral tool for investigating executive control. *Psychophysiology*, *33*, 282–294.
- Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge University Press.
- Lehle, C., & Hübner, R. (2008). On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review*, *15*, 814–818.
- Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*, 166–174.
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *8*, 61–64.
- Schmidt, J. R. (2013). Temporal learning and list-level proportion congruency: Conflict adaptation or learning when to respond? *PLOS ONE*, *8*, Article e82320.
- Schmidt, J. R. (2016). Time-out for conflict-monitoring theory: Preventing rhythmic biases eliminates the list-level proportion congruent effect. *Canadian Journal of Experimental Psychology*, *71*, 52–62.
- Schmidt, J. R., & Besner, D. (2008). The Stroop effect: Why proportion congruence has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 514–523.
- Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 506–526.
- Spinelli, G., Perry, J. R., & Lupker, S. J. (2019). Adaptation to conflict frequency without contingency and temporal learning: Evidence from the picture-word interference task. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 995–1014.
- Taylor, D. A. (1977). Time course of context effects. *Journal of Experimental Psychology: General*, *106*, 404–426.
- Torres-Quesada, M., Funes, M. J., & Lupiáñez, J. (2013). Dissociating proportion congruent and conflict adaptation effects in a Simon-Stroop procedure. *Acta Psychologica*, *142*, 203–210.
- Torres-Quesada, M., Milliken, B., Lupiáñez, J., & Funes, M. J. (2014). Proportion congruent effects in the absence of sequential congruent effects. *Psicológica*, *35*(1), 101–115.
- Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, *115*, 518–525.
- Wendt, M., & Luna-Rodriguez, A. (2009). Conflict frequency affects flanker interference: Role of stimulus-ensemble-specific practice and flanker-response contingencies. *Experimental Psychology*, *56*, 206–217.
- Wendt, M., Luna-Rodriguez, A., & Jacobsen, T. (2012). Conflict-induced perceptual filtering. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 675–686.
- Wühr, P., Duthoo, W., & Notebaert, W. (2015). Generalizing attentional control across dimensions and tasks: Evidence from transfer of proportion congruent effects. *Quarterly Journal of Experimental Psychology*, *68*, 779–801.