The Quarterly Journal of Experimental Psychology

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: http://www.tandfonline.com/loi/pqje20

Reproducing the location-based context-specific proportion congruent effect for frequency unbiased items: A reply to Hutcheon and Spieler (2016)

Matthew J. C. Crump, Nicholaus P. Brosowsky & Bruce Milliken

To cite this article: Matthew J. C. Crump, Nicholaus P. Brosowsky & Bruce Milliken (2016): Reproducing the location-based context-specific proportion congruent effect for frequency unbiased items: A reply to Hutcheon and Spieler (2016), The Quarterly Journal of Experimental Psychology, DOI: 10.1080/17470218.2016.1206130

To link to this article: http://dx.doi.org/10.1080/17470218.2016.1206130

View supplementary material

Accepted author version posted online: 24 Jun 2016.
Published online: 12 Jul 2016.

Submit your article to this journal

Article views: 20

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=pqje20

Download by: [Brooklyn College] **Date:** 23 August 2016, At: 10:29

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2016
http://dx.doi.org/10.1080/17470218.2016.1206130

Routledge
Taylor & Francis Group

Reproducing the location-based context-specific proportion congruent effect for frequency unbiased items: A reply to Hutcheon and Spieler (2016)

Matthew J. C. Crump[a], Nicholaus P. Brosowsky[b] and Bruce Milliken[c]

[a]Department of Psychology, Brooklyn College and the Graduate Center of the City University of New York, Brooklyn, NY, USA; [b]Department of Psychology, Graduate Center of the City University of New York, Brooklyn, NY, USA; [c]Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, ON, Canada

ABSTRACT

Stroop effects can be modulated by context-specific cues associated with different levels of proportion congruent, even for items that appear equally frequently in each context. This result has important theoretical implications, because it rules out frequency-driven learning explanations of context-specific proportion congruent (CSPC) effects and leaves open the possibility that a cue-driven retrieval process can reinstate attentional control settings in a rapid online fashion. The purpose of the present work was to address reproducibility concerns that have been raised about this finding. We conducted several reproductions and novel extensions using Amazon's mechanical Turk in both Stroop and flanker tasks. We successfully replicated the central finding that CSPC effects can be observed for frequency-unbiased items. We also provide new Monte Carlo simulation analyses to estimate reproducibility of the phenomena that show important limitations on these designs for measuring contextual control.

ARTICLE HISTORY
Received 24 November 2015
Accepted 13 June 2016

KEYWORDS
Attention; Automaticity; Contextual control; Proportion congruent; Replication

Several findings in the attention literature provide converging evidence that online processing of contextual cues in the environment can trigger adjustments to attention processes (for reviews see Bugg, 2012; Bugg & Crump, 2012; Chun & Turk-Browne, 2007; Egner, 2008; Vecera, Cosman, Vatterott, & Roper, 2014). One line of evidence for contextual cueing of attentional control settings consists of demonstrations in Stroop (Stroop, 1935) and flanker (Eriksen & Eriksen, 1974) tasks that interference effects are larger for items appearing in a context associated with a higher proportion of congruent items than in a context associated with a lower proportion of congruent items (for reviews see Bugg, 2012; Bugg & Crump, 2012). This evidence is consistent with a contextual control account where contextual cues trigger the reinstatement of attentional control settings typically used in those contexts in the past. The contextual control interpretation has been debated because some designs confound the context-specific proportion congruent (CSPC) manipulation with item frequency, leaving open the possibility that CSPC effects reflect item-specific practice (Logan, 1988) rather than online, context-triggered adjustments to control settings.

A critical set of experiments by Crump and Milliken (2009) addressed the frequency-driven learning account of CSPC effects by showing that CSPC effects can be obtained for frequency unbiased items. However, Hutcheon and Spieler (2016) recently reported one failed straight replication of Crump and Milliken's second experiment, and two additional failed attempts to show CSPC effects for frequency unbiased items in different but similar designs, suggesting that the original Crump and Milliken results were spurious. Given the theoretical

CONTACT Matthew J. C. Crump mcrump@brooklyn.cuny.edu Department of Psychology, Brooklyn College of the City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210, USA.
Supplemental data for this article can be accessed here. doi:10.1080/17470218.2016.1206130
© 2016 The Experimental Psychology Society

importance of determining whether CSPC effects can be obtained for frequency unbiased items, we conducted additional replications of both experiments reported by Crump and Milliken using a Stroop task, and for the first time using a flanker task. To foreshadow the main results, consistent with our previous findings, we report that CSPC effects can be obtained for frequency unbiased items in both Stroop and flanker tasks. We now briefly review the relevant background issues that motivate the need to assess the reproducibility of CSPC effects for frequency unbiased items and refer the reader to Crump and Milliken (2009) as well as to Hutcheon and Spieler (2016) who have provided additional background. We reserve the term *replication* to refer to experiments that repeat the same design as that in prior experiments and *reproduction* to refer to experiments that repeat closely related designs measuring the same general phenomena.

Early CSPC designs used location as a contextual cue in Stroop (Crump, Gong, & Milliken, 2006; Crump, Vaquero, & Milliken, 2008) and flanker tasks (Corballis & Gratton, 2003). In those designs, Stroop or flanker items were presented in one of two locations on a screen in a random intermixed fashion across trials, with one location associated with a higher proportion of congruent than incongruent items, and the other location associated with a lower proportion of congruent than incongruent items. Those designs produced larger congruency effects in the high than in the low proportion congruent locations, a difference referred to as the context-specific proportion congruent effect.

There have been and continue to be multiple accounts of CSPC effects. Corballis and Gratton (2003) used left and right location contexts and proposed that distinct attentional control settings may be acquired by each hemisphere and controlled by contextual information presented in different hemifields. Crump et al. (2006) suggested that cues may operate more generally to retrieve and apply attentional control settings preserved in memory associated with a present context. King, Korb, and Egner (2012) showed that CSPC effects in the flanker task depend on context repetitions between trials and proposed that context cues prime associated control settings, perhaps in a less online manner than previously thought. King, Donkin, Korb, and Egner (2012) used a linear ballistic accumulator model (Brown & Heathcote, 2008) to fit their data and showed that changes in threshold provided a better explanation of CSPC effects than changes in drift rate. They suggested that CSPC effects do not reflect cue-driven changes to attentional weights and instead reflect cue-driven biases to decision criteria or response caution (see also Schmidt, Lemercier, & De Houwer, 2014, who propose a rhythmic bias account).

A common assumption among the above accounts of CSPC effects is that contextual cues change aspects of control processes, such as attention weights or decision-thresholds. However, some CSPC designs have confounded the proportion congruent manipulation with item frequency, leaving open the possibility that CSPC effects reflect additive influences of congruency on the one hand and a frequency-sensitive learning process on the other hand (Logan, 1988; Schmidt & Besner, 2008; Schmidt, Crump, Cheesman, & Besner, 2007). For example, Crump et al. (2006) defined one location as 75% congruent and the other as 25% congruent, but presented particular combinations of word/colour/location stimuli with different frequencies in each location. In a given block, congruent compounds (e.g. red/red/above) appeared nine times, and incongruent compounds (e.g. red/green/above) appeared one time each in the 75% congruent location, whereas each congruent and incongruent compound appeared three times in the 25% congruent location. A frequency-sensitive learning process would be expected to speed reaction times (RTs) for more frequent than infrequent compounds—the relevant predictions of this simple principle are illustrated in Figure 1. If subjects are sensitive to the frequency of event compounds, then RTs should be faster for the more frequent congruent compounds in the 75% than in the 25% congruent locations, and faster for the more frequent incongruent compounds in the 25% than in the 75% congruent locations. Such an influence would produce larger congruency effects in the 75% than in the 25% locations and would provide a parsimonious explanation of CSPC effects in terms of item-specific learning.

The results of Crump and Milliken (2009) were theoretically important because they provided a critical test of a frequency-driven learning account of CSPC effects. In both experiments they used a prime–probe version of the Stroop procedure, where a word prime was presented briefly for 100 ms followed by a to-be-named coloured rectangle probe that appeared above or below fixation. In both experiments, one location was associated with a higher proportion of congruent items and the other with a lower proportion of congruent items. There were two sets of

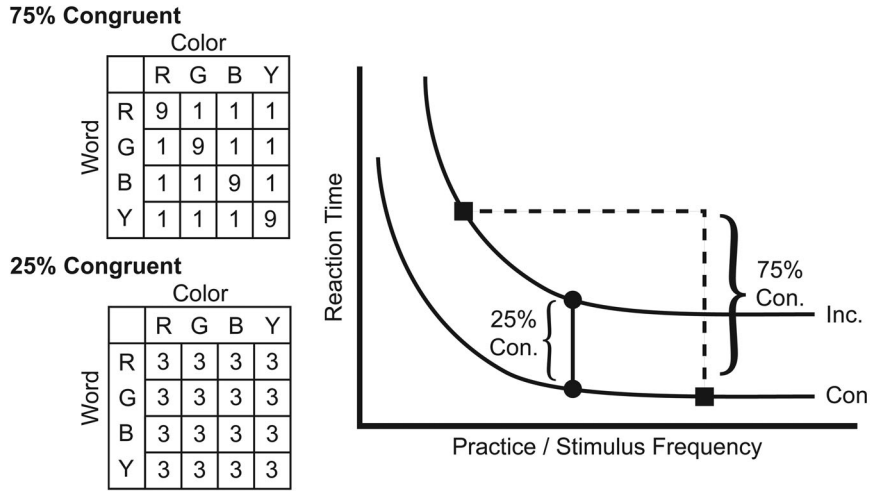

Figure 1. The tables and graph illustrate how a frequency-sensitive learning process could produce a context-specific proportion congruent effect for frequency biased items. The tables show frequencies for word/colour pairs in 75% and 25% congruent locations from a typical context-specific proportion congruent design. The graph shows predicted reaction times for congruent and incongruent items in the 75% and 25% congruent locations as a function of practice. The braces show larger Stroop effects for the 75% than for the 25% congruent location because the differences between congruent and incongruent items are being compared at different points on the learning curve. Con. = congruent; inc. = incongruent; R = red; G = green; B = blue; Y = yellow.

items, here termed frequency biased and unbiased sets. The frequency biased sets involved Stroop stimuli made up from two colours (e.g. red and blue), and the unbiased sets were made up from two different colours (e.g. green and yellow).

In Experiment 1, the frequency biased items appeared in their congruent forms (e.g. red in RED, and blue in BLUE) 100% of the time in one location, and appeared in their incongruent forms (e.g. red in BLUE and blue in RED) 100% of the time in the other location. The frequency unbiased items appeared in their congruent and incongruent forms with equal probability in both locations. Both sets of items were intermixed and presented in a randomized fashion from trial to trial. Thus, the list-wide proportion congruent was .50, whereas the location-specific proportion congruent was .75 in one location and .25 in the other. The results showed a significant CSPC effect for the frequency unbiased items that appeared to emerge with practice. Specifically, the Stroop effect was larger for the frequency unbiased items when they appeared in the 75% congruent location than when they appeared in the 25% congruent location. This difference was not observed in the first two blocks of the experiment, but was observed in the second two blocks of the experiment.

The design of Experiment 2 was similar except that the frequency biased items were allowed to be congruent and incongruent in both locations. Specifically, in one location 92% of those items were congruent, and 8% were incongruent, and the reverse was true for the other location. Thus, the overall list-wide proportion congruent remained .50, and the location-specific proportion congruent was reduced to .71 in one location and .29 in the other. This experiment measured and compared CSPC effects both for frequency biased and unbiased items, and CSPC effects were obtained for both sets of items, with larger CSPC effects for frequency biased than for unbiased items according to a one-tailed test.

Hutcheon and Spieler (2016) reported a failed attempt to replicate the second experiment reported by Crump and Milliken (2009). Their straight replication did not show CSPC effects for frequency biased or unbiased items. They conducted two additional experiments that also failed to show CSPC effects for frequency unbiased items. Their second experiment increased the set size of frequency biased and unbiased items from two to four and found no evidence for CSPC effects for either frequency biased or unbiased sets. Their third experiment involved a two-phase design. In the first phase, frequency biased items (set size 4) were used to establish the CSPC manipulation across four blocks of trials, and the new frequency unbiased items were introduced (in addition to the existing items) in a following test

phase involving two blocks. That experiment showed significant CSPC effects for frequency biased items in the first phase and no CSPC effects for the frequency unbiased items in the test phase. Given their three failed attempts, they concluded that the results of Crump and Milliken may have been spurious, reflecting a Type I error.

Present aims

Given the theoretical importance of the presence or absence of CSPC effects for frequency unbiased items, we conducted our own set of reproductions of both experiments reported by Crump and Milliken (2009). We conducted all of the reproductions through the web using Amazon's Mechanical Turk (AMT). Each task was programmed using JavaScript and HTML and ran locally in web browsers. Web-based methods for conducting attention and performance experiments requiring precise temporal control for stimulus presentation and response recording have been validated by several studies (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015; Crump, McDonnell, & Gureckis, 2013; Reimers & Maylor, 2005; Reimers & Stewart, 2015; Schubert, Murteira, Collins, & Lopes, 2013; Simcox & Fiez, 2014). As well, the web-based approach allowed us to conduct the replications in a timely manner, with a larger sample size, and with a subject pool that extends beyond the typical undergraduate psychology pool. Conducting experiments online in this manner does not offer the same kind of stringent control made possible by laboratory studies, so demonstrating positive findings in the face of variable testing conditions (e.g. screen size, location, distraction, etc.) would speak to the robustness of any findings.

The first set of reproductions asked whether CSPC effects can be obtained for frequency unbiased items in a Stroop task. We used the same prime–probe variant of the Stroop task as Crump and Milliken (2009), but changed how subjects made identification responses to the colour. The original studies involved vocal naming, and reaction times were collected with a voice-key, which was not possible with our web-based program. Our subjects either pressed a single key (the first letter of the colour), or typed the entire colour to respond on each trial. The second set of experiments consisted of extensions of the Crump and Milliken (2009) designs from the Stroop to the flanker task. Each of the following experimental sections report designs and analyses for determining the presence of CSPC effects for frequency unbiased items. After these sections we also report a broader Monte Carlo simulation based analysis of power to detect CSPC effects using the present designs.

Experiment 1

Experiment 1 determined whether CSPC effects for frequency unbiased items can be reproduced using the design of the first experiment reported by Crump and Milliken (2009). The design involved a prime–probe variant of the Stroop task where a colour word (red, green, blue, or yellow) is presented briefly in the centre of the screen, followed by a coloured rectangle presented above or below fixation. The CSPC manipulation was introduced by varying proportion congruent between locations. Two sets of items could appear in both high and low proportion congruent locations. The frequency biased set appeared as congruent items on 100% of trials in one location and as incongruent items on 100% of trials in the other location. The frequency unbiased items appeared in both locations as congruent items on 50% of trials and as incongruent items on 50% of trials. All items were mixed together and presented randomly, such that the overall list-wide proportion congruent was .50, and the CSPC was .75 in one location and .25 in the other. As already mentioned, because this reproduction effort was conducted online, it was not possible to use a vocal naming response. Instead, we tested two groups of subjects that differed only in how they recorded their responses. One group recorded their responses with a single key press response on each trial, whereas the other group recorded their responses with a typewritten response. Typewritten responses have been shown previously to produce large Stroop effects (Logan & Zbrodoff, 1998).

Method

Subjects

All subjects were recruited from AMT and were compensated $1.50 for participating. For each experiment the number of HITs (human intelligence tasks, an Amazon term for a work unit) refers to the number of subjects who initiated the study. Subjects were included in the study if they completed all trials. A total of 100 HITS were posted, and 95 subjects completed all trials (49 for button response, 46 for typing response). Demographic information was collected

and is reported in Supplemental Material A on the article's online page. We calculated the number of subjects needed to achieve power of .8 to detect the three-way interaction between learning phase, proportion congruent, and congruency reported by Crump and Milliken (2009, Experiment 1) for frequency unbiased items as $n = 16$. We assumed that a minimum of 16 subjects would be a conservative estimate when generalizing to online populations where we expected increased variability in performance, so we chose to include substantially more subjects than required by the power analysis in each of the response type conditions.

Apparatus and stimuli

The experiment was programmed in house using JavaScript and HTML. The program allowed subjects to complete the task only if they were running Safari, Google Chrome, or Firefox web browsers.

Each experiment ran as a pop-up window that filled the entire screen. The background was black. The word stimuli appeared in the centre of the screen in white, 30-point, Helvetica font. The colour-patch stimuli were rectangles 100 pixels in height by 300 pixels in width. The screen-size of the web-browser was computed when the page was loaded, and the vertical location of the patches was determined as a proportion of the height of the screen. Patches appearing in the top location were centred horizontally, and were presented vertically at a position that was a quarter of the distance from the top of the display to the bottom of the display. Patches appearing in the bottom location were centred horizontally, and were presented vertically three quarters of the distance from the top of the display to the bottom of the display. The words were red, green, blue, and yellow, and the colours were the equivalent web-standard hues associated with those labels.

Design

The design followed Experiment 1 reported by Crump and Milliken (2009). Two non-overlapping sets of Stroop items were created for each subject, termed the frequency biased and unbiased sets. Each set involved two colour words (e.g. red and green) and their corresponding congruent (red in RED, green in GREEN) and incongruent pairs (red in GREEN, green in RED). The other set involved the congruent and incongruent pairs made up from the remaining colour words (e.g. blue and yellow). In the current version, for each subject, we randomly assigned two colour words to the frequency biased set and the remaining two colour words to the frequency unbiased set.

For each subject, one location (either above or below fixation) was defined as the high proportion congruent location, and the other location was defined as the low proportion congruent location. Assignment of location to the two proportion congruent conditions was randomly determined. The frequency biased items appeared in their congruent form 100% of the time in the high proportion congruent location, and in their incongruent form 100% of the time in the low proportion congruent location. The frequency unbiased items appeared in both congruent and incongruent forms 50% of the time in both locations. There were four blocks of 96 trials, with 48 trials involving frequency biased items and 48 trials involving frequency unbiased items in each block. For the purpose of analyses, the first two blocks of 96 trials constituted the first learning phase while the last two blocks constituted the second learning phase. Tables showing example item frequencies in each location can be found in Crump and Milliken (2009). Within each block, all trials were intermixed and were presented in a randomized fashion.

Procedure

All subjects were AMT workers who found the experiment using the AMT system. The subject recruitment procedure and tasks were approved by the Brooklyn College Institutional Review Board. Each subject read a short description of the task and gave consent by pressing a button acknowledging they had read the displayed consent form. Subjects then completed a short demographic survey and proceeded to the main task, which was displayed as a pop-up window.

At the beginning of each trial, participants were presented with a fixation cross displayed in white against a black background for 1000 ms, followed by a blank interval of 250 ms. Next, a prime word was presented centrally for 100 ms, followed immediately by a colour-patch probe displayed above or below fixation. The probe remained on the screen until the participant made a response. Subjects in the button response condition gave a single key-press response by pressing the keys r, g, b, or y. Subjects in the typing condition typed the entire colour name as their response. Feedback was immediately displayed and was presented on screen randomly between 1000 and 1300 ms, then feedback was removed, and the next trial began in 500 ms. Feedback indicated

whether the answer was correct or incorrect. If the response time was greater than 1500 ms, then the message "respond faster" appeared to encourage speeded responding.

Results

Subjects with mean error rates (i.e. proportions) collapsed across all conditions greater than .2 were excluded from further analysis. This subject exclusion criterion was applied to all remaining analyses for all experiments. Four subjects in the single-button response type condition were removed, and no subjects in the typing response type condition were removed. For all remaining subjects, RTs greater than 100 ms from correct trials in each condition were submitted to an outlier rejection procedure (Van Selst & Jolicoeur, 1994) that eliminated an average of 3% of the observations in each condition. The same non-recursive version of the outlier procedure was applied to all RT analyses reported in this article. The resulting mean RTs and error rates were submitted to the following analyses, and an alpha criterion of .05 was adopted for all statistical tests.

The mean RTs and error rates for the frequency unbiased items were submitted to separate 2 (response type: single-button vs. typed) × 2 (learning phase: first half vs. last half) × 2 (proportion congruent: high vs. low) × 2 (congruency: congruent vs. incongruent) mixed analyses of variance (ANOVAs) with response type as the sole between-subjects factor and the remaining factors as repeated measures. Mean RTs and error rates for all conditions are displayed in Table 1, and the full table of inferential statistics is presented in Supplemental Material B. For brevity we focus on the RT results most relevant to determining the presence or absence of a CSPC effect for frequency unbiased items.

Most important, we found clear evidence of a CSPC effect for frequency unbiased items. Specifically, the interaction between proportion congruent and congruency was significant, $F(1, 89) = 10.11$, $MSE = 1760.58$, $p = .002$, $\eta_p^2 = .10$. The Stroop effect was 20 ms larger in the high (93 ms) than in the low (73 ms) proportion congruent locations.

We found no evidence that the CSPC effect for frequency unbiased items differed between single-button and typing responses. Specifically, the three-way interaction between response type, proportion congruent, and congruency was not significant, $F(1, 89) = 0.003$, $MSE = 1760.58$, $p = .957$, $\eta_p^2 = .00$.

Similarly, the CSPC effect for frequency unbiased items did not depend on the learning phase. The interaction between learning phase, proportion congruent, and congruency was not significant, $F(1, 89) = 1.64$, $MSE = 1529.45$, $p = .20$, $\eta_p^2 = .02$. The absence of a three-way interaction fails to reproduce Crump and Milliken's (2009) result that the CSPC effect for frequency unbiased items was larger in the last than in the first half of the experiment. Although we did not find statistical support for this difference, we do note that the pattern of CSPC effects observed here is consistent with the direction of previous findings. Specifically, the CSPC effect was 12 ms in the first half and 28 ms in the last half of this reproduction experiment.

Discussion

We successfully reproduced the CSPC effect for frequency unbiased items using the first design reported by Crump and Milliken (2009). The present design was not a straight replication because our tasks required

Table 1. Mean correct colour identification response latencies, standard errors, error rates, Stroop effects, and context-specific proportion congruent effects for Experiment 1.

			RT		*SE*		ER		Stroop		CSPC	
Resp.	Half	PC	Con	Inc	Con	Inc	Con	Inc	I – C	SE	H – L	SE
Button	First	High	796	862	21	22	.03	.04	66	12		
		Low	806	857	23	24	.02	.03	51	11	15	11
	Last	High	768	837	31	34	.02	.03	69	11		
		Low	773	818	27	24	.02	.03	45	7	24	11
Typing	First	High	731	859	16	17	.01	.03	128	10		
		Low	740	859	17	21	.01	.02	119	16	9	16
	Last	High	716	823	19	17	.00	.02	107	10		
		Low	732	807	20	15	.02	.01	75	10	32	9

Note: RT = reaction time; *SE* = standard error; ER = error rate; Resp. = response type; PC = proportion congruent; Con = congruent; Inc = incongruent; I – C = incongruent minus congruent; CSPC = context-specific proportion congruent; H – L = high proportion congruency effect minus low proportion congruency effect.

single-button and typewritten responses, rather than vocal naming. So, our findings also provide an extension showing generalization across response requirements. Similarly, our online approach shows that CSPC effects for frequency unbiased items can be measured in a more general population, and under more variable testing conditions than would be the case in the laboratory. Last, Crump and Milliken reported that CSPC effects for frequency unbiased items developed over the course of the experiment. Although the pattern of RTs was numerically consistent with an effect of learning phase, this effect was not statistically significant.

Experiment 2

The purpose of Experiment 2 was to determine whether CSPC effects for frequency unbiased items can be reproduced using the design of the second experiment reported by Crump and Milliken (2009). Hutcheon and Spieler (2016) have reported one failed attempt to reproduce the findings from this design. The major difference between the first and second experiments reported by Crump and Milliken was that the second experiment included congruent and incongruent frequency biased items in both high and low proportion congruent contexts. In line with this method, in the present experiment the frequency biased set appeared as congruent items on 92% of trials and as incongruent items on 8% of trials in one location, and as congruent items on 8% of trials and as incongruent items on 92% of trials in the other location. The frequency unbiased items appeared in both locations as congruent and incongruent items 50% of the time. All items were mixed together and were presented randomly, such that the overall list-wide proportion congruent was .50, and the CSPC was .71 in one location and .29 in the other. The inclusion of congruent and incongruent items for the frequency biased items allowed a comparison of the size of the CSPC effect for frequency biased versus unbiased items. As with our first reproduction attempt, two groups of subjects completed this experiment. One group used single-button press responses, while the other group used typewritten responses.

Method

Subjects

All subjects were recruited from AMT and were compensated $3.00 for participating. Again, the number of HITs refers to the number of subjects who initiated the study. Subjects were included in the study if they completed all trials. A total of 50 HITS were posted, and 49 subjects completed all trials (24 for button response, 25 for typing response). Demographic information was collected and is reported in Supplemental Material A. We calculated the number of subjects needed to achieve power of .8 to detect the interaction between proportion congruent and congruency reported by Crump and Milliken (2009, Experiment 2) for frequency unbiased items as $n = 32$. Again, we include more subjects than estimated (collapsing across response type, which did not interact with the CSPC effect in Experiment 1) by the power analysis.

Apparatus and stimuli

Experiment 2 used the same apparatus and stimuli as those in Experiment 1.

Design

The design followed Experiment 2 reported by Crump and Milliken (2009) and was very similar to that of Experiment 1 with the exception that frequency biased items appeared in congruent and incongruent forms in both locations. Specifically, within each block of trials, frequency biased items were 92% congruent and 8% incongruent in the high proportion congruent location, and 8% congruent and 92% incongruent in the low proportion congruent location. The frequency unbiased items remained as 50% congruent and 50% incongruent in both locations.

Procedure

The procedure was identical to that of Experiment 1.

Results

Two subjects in the single-button response condition and one subject in the typing response condition were excluded for error rates higher than .2. The outlier elimination procedure removed an average of 3% of the observations in each condition.

The mean RTs for correct trials and error rates were submitted to separate 2 (response type: single-button vs. typed) × 2 (item type: frequency biased vs. unbiased) × 2 (learning phase: first half vs. last half) × 2 (proportion congruent: high vs. low) × 2 (congruency: congruent vs. incongruent) mixed ANOVAs with response type as the sole between-subjects factor and the remaining factors as repeated

measures. Mean RTs and error rates for all conditions are displayed in Table 2. The inferential statistics for the RT analysis are shown in Supplemental Material C1, and the error rate analysis is shown in Supplemental Material C2.

We did not observe a significant CSPC effect either for frequency biased or for unbiased items. First, the Proportion Congruent × Congruency interaction was not significant, $F(1, 44) = 1.23$, $MSE = 3980.62$, $p = .295$, $\eta_p^2 = .03$. Second, there were no significant higher order interactions modulating the Proportion Congruent × Congruency interaction. And third, the pattern of CSPC effects (see Table 2) did not generally resemble the findings of Crump and Milliken (2009), who reported positive CSPC effects for frequency biased and unbiased items.

Discussion

As with Hutcheon and Spieler's (2016) failed replication, our reproduction attempts of the second experimental design from Crump and Milliken (2009) failed to produce CSPC effects for frequency biased and unbiased items. The mean CSPC effects reported in Table 2 show non-significant trends that are opposite in direction to those in prior studies (i.e. negative CSPC effects), with the exception that subjects who gave typewritten responses showed a +20-ms CSPC effect in the last half of the experiment.

The two failed attempts to reproduce CSPC effects speak to the reliability of the second design to measure CSPC effects. The design was originally adopted to compare the size of CSPC effects between frequency biased and unbiased items. CSPC effects for frequency biased items were measured by including congruent and incongruent items in both locations, which also reduced the overall difference in proportion congruent between locations compared to that in the first design. As a result, the weaker CSPC manipulation would be expected to produce smaller CSPC effects that are more difficult to detect, and perhaps less statistically reliable. Additionally, there are clear concerns with cell size for measuring CSPC effects for frequency biased items. For example, across the entire experiment in the high proportion congruent location there are 72 total congruent items but only eight incongruent items, and any incongruent errors would reduce cell size further. As a result, the Stroop effect in the high proportion congruent location should be much more variable for the frequency biased than for the unbiased items. We return to this issue following the experimental sections where we use Monte Carlo simulations to estimate the probability of reproducing CSPC effects using the designs from Experiments 1 and 2.

Experiment 3

The purpose of Experiment 3 was to determine whether CSPC effects for frequency unbiased items can be reproduced using the design of the first experiment reported by Crump and Milliken (2009) in a

Table 2. Mean correct colour identification response latencies, standard errors, error rates, Stroop effects, and context-specific proportion congruent effects for Experiment 2.

				RT		*SE*		ER		Stroop		CSPC	
Resp.	IT	Half	PC	Con	Inc	Con	Inc	Con	Inc	I – C	SE	H – L	SE
Button	B	First	High	749	802	25	33	.03	.05	53	25		
			Low	741	814	21	24	.08	.03	73	21	−20	35
		Last	High	708	804	28	30	.02	.02	96	23		
			Low	679	780	24	28	.02	.03	101	15	−5	30
	U	First	High	751	822	27	25	.03	.05	71	12		
			Low	747	826	27	27	.02	.06	79	13	−8	11
		Last	High	739	798	33	26	.03	.03	59	15		
			Low	722	789	30	31	.03	.04	67	11	−8	17
Typing	B	First	High	725	855	23	31	.00	.03	130	26		
			Low	708	850	25	21	.01	.02	142	13	−12	26
		Last	High	718	799	24	21	.01	.03	81	18		
			Low	698	821	27	24	.03	.02	123	21	−42	19
	U	First	High	723	846	28	24	.02	.02	123	11		
			Low	732	859	26	27	.01	.02	127	9	−4	15
		Last	High	710	824	26	25	.01	.01	114	13		
			Low	731	824	31	26	.02	.01	93	16	21	14

Note: RT = reaction time; *SE* = standard error; ER = error rate; Con = congruent; Inc = incongruent; Resp. = response type; IT = item type; B = frequency biased; U = frequency unbiased; PC = proportion congruent; I – C = incongruent minus congruent; CSPC = context-specific proportion congruent; H – L = high proportion congruency effect minus low proportion congruency effect.

flanker task. The flanker stimuli were constructed from the letters s, d, j, and k. The task was to identify the centrally presented letter, which was flanked by identical distractor letters on congruent trials and by non-identical distractor letters on incongruent trials. Other than replacing the Stroop items with flanker items, the remaining aspects of the design were the same as those in Experiment 1.

Method

Subjects

All subjects were recruited from AMT and were compensated $1.50 for participating. A total of 150 HITS were posted, and 146 subjects completed all trials. Demographic information was collected and is reported in Supplemental Material A. We again included substantially more subjects than required by the same estimates for number of subjects to achieve power .8 from Experiment 1.

Apparatus and stimuli

Experiment 3 used the same general apparatus as Experiment 1, but used flanker stimuli rather than Stroop stimuli. Flanker stimuli were made from the following letter set: S, D, J, K. Each stimulus consisted of a centrally presented target letter flanked on the left and right by two congruent (DDDDD) or incongruent (SSDSS) letters. Letters were presented in white, 25-point Helvetica font, with 0-point spacing between letters. Frequency biased and unbiased sets of flanker items were constructed in the same fashion as the Stroop experiments from Experiments 1 and 2. Two letters were randomly chosen to form items for the frequency biased set, and the remaining two letters formed items for the frequency unbiased set.

Design

The design was the same as that of Experiment 1, with frequency biased items appearing on 100% of trials as congruent in the high proportion congruent location, and on 100% of trials as incongruent in the low proportion congruent location. Frequency unbiased items appeared on 50% of trials as congruent items and on 50% of trials as incongruent items in both locations.

Procedure

The general procedure followed Experiment 1, with the exception that participants made keypress responses (s, d, j, k) to identify the target letter on each trial. Additionally, the Stroop experiments employed a prime–probe variant where the distractor word was presented as a prime before the target probe colour patch was presented. For the flanker experiments the target and distractor letters were all presented simultaneously.

Results

Eight subjects were excluded for mean error rates higher than .2. The outlier elimination procedure removed an average of 3% of the observations in each condition. The mean RTs for correct frequency unbiased trials and error rates were submitted to separate 2 (learning phase: first half vs. last half) × 2 (proportion congruent: high vs. low) × 2 (congruency: congruent vs. incongruent) repeated measures ANOVAs. Mean RTs and error rates for all conditions are displayed in Table 3, and the inferential statistics for both ANOVAs are displayed in Supplemental Material D.

Most important, we found evidence of a CSPC effect for frequency unbiased items. Specifically, the Proportion Congruent × Congruency interaction was significant, $F(1, 137) = 4.31$, $MSE = 1953.97$, $p = .039$, $\eta_p^2 = .03$. The Stroop effect was larger for the high (101 ms) than low proportion (90 ms) congruent location, showing an 11-ms CSPC effect. The three-way interaction between learning phase, proportion

Table 3. Mean correct letter identification response latencies, standard errors, error rates, Flanker effects, and context-specific proportion congruent effects for Experiment 3.

		RT		*SE*		ER		Flanker		CSPC	
Half	PC	Con	Inc	Con	Inc	Con	Inc	I – C	SE	H – L	SE
First	High	783	885	12	12	.03	.04	102	6		
	Low	787	879	12	11	.03	.05	92	6	10	8
Last	High	742	841	10	11	.02	.03	99	5		
	Low	749	837	11	11	.02	.03	88	5	11	5

Note: RT = reaction time; *SE* = standard error; ER = error rate; PC = proportion congruent; Con = congruent; Inc = incongruent; I – C = incongruent minus congruent; CSPC = context-specific proportion congruent; H – L = high proportion congruency effect minus low proportion congruency effect.

congruent, and congruency was not significant, $F(1, 137) = 0.044$, $MSE = 1072.68$, $p = .833$, $\eta_p^2 = .00$.

Discussion

The major new finding was that a CSPC effect for frequency unbiased items was observed in a flanker task using the Crump and Milliken (2009) design (see also King, Korb, et al., 2012, who showed CSPC effects in a face-based flanker task with all unique items). So, again the important result of Crump and Milliken (2009) is reproducible and can be extended to other interference tasks beyond the Stroop task.

Experiment 4

The purpose of Experiment 4 was to determine whether CSPC effects for frequency unbiased items can be reproduced using the design of the second experiment reported by Crump and Milliken (2009) in a flanker task. Other than replacing the Stroop items with flanker items, the remaining aspects of the design were the same as those in Experiment 2.

Method

Subjects

All subjects were recruited from AMT and were compensated $1.50 for participating. A total of 50 HITS were posted, and 49 subjects completed all trials. Demographic information was collected and is reported in Supplemental Material A. We again included more subjects than required by the same estimates for number of subjects to achieve power .8 from Experiment 2.

Apparatus and stimuli

Experiment 4 used the same apparatus and stimuli as those in Experiment 3.

Design

The design followed was the same as that in Experiment 2.

Procedure

The procedure was the same as that in Experiment 3.

Results

Four subjects were excluded for mean error rates higher than .2. Two additional subjects were removed because of committing 100% errors for low-frequency items in the frequency biased set. The outlier elimination procedure removed an average of 3% of the observations in each condition.

The mean RTs for correct trials and error rates were submitted to separate 2 (learning phase: first half vs. last half) × 2 (item set: frequency biased vs. unbiased) × 2 (proportion congruent: high vs. low) × 2 (congruency: congruent vs. incongruent) repeated measures ANOVAs. Mean RTs and error rates for all conditions are displayed in Table 4, and the inferential statistics are shown in Supplemental Material E.

Although CSPC effects for the frequency biased and unbiased sets were numerically in the expected direction (see Table 4), the overall proportion congruent by congruency interaction was not significant, $F(1, 42) = 1.97$, $MSE = 4799.01$, $p = .168$, $\eta_p^2 = .04$. There were main effects of learning phase, $F(1, 42) = 25.99$, $MSE = 18{,}158.89$, $p = .001$, $\eta_p^2 = .38$; proportion congruent, $F(1, 42) = 5.00$, $MSE = 5831.65$, $p = .031$, $\eta_p^2 = .11$, and congruency, $F(1, 42) = 260.75$, $MSE = 6030.01$, $p = .001$, $\eta_p^2 = .86$. Responses were faster for the last

Table 4. Mean correct letter identification response latencies, standard errors, error rates, Flanker effects, and context-specific proportion congruent effects for Experiment 4.

			RT		*SE*		ER		Flanker		CSPC	
IT	Half	PC	Con	Inc	Con	Inc	Con	Inc	I – C	SE	H – L	SE
B	First	High	827	931	25	28	.03	.08	104	18		
		Low	847	949	27	28	.04	.03	102	18	2	26
	Last	High	786	873	26	26	.02	.03	87	17		
		Low	815	889	27	24	.01	.03	74	13	13	20
U	First	High	830	942	26	27	.03	.04	112	11		
		Low	845	932	25	26	.03	.03	87	8	25	10
	Last	High	771	880	22	24	.02	.03	109	8		
		Low	789	880	23	23	.02	.03	91	8	18	10

Note: RT = reaction time; *SE* = standard error; ER = error rate; IT = item type; B = frequency biased; U = frequency unbiased; PC = proportion congruent; Con = congruent; Inc = incongruent; I – C = incongruent minus congruent; CSPC = context-specific proportion congruent; H – L = high proportion congruency effect minus low proportion congruency effect.

half than for the first half of trials, for the high than for the low proportion congruent condition, and of course for congruent than for incongruent trials. No other interaction effects were significant.

At the same time, an analysis that was restricted to frequency unbiased items did reveal a significant CSPC effect. In that analysis, the Proportion Congruent × Congruency interaction was significant, $F(1, 42) = 7.95$, $MSE = 1254.31$, $p = .007$, $\eta_p^2 = .16$, and did not interact with learning phase, $F(1, 42) = 0.39$, $MSE = 1019.65$, $p = .53$, $\eta_p^2 = .01$.

Discussion

Experiment 4 produced a significant CSPC effect for frequency unbiased items, but did not produce a CSPC effect for frequency biased items. The CSPC effect found for frequency unbiased items constitutes a successful reproduction of the major finding from the second design of Crump and Milliken (2009) and contrasts with the recent unsuccessful replication attempts of this design with Stroop items. Note that we are not claiming that the results of Experiment 4 fully reproduce those of Crump and Milliken, as we did not observe a CSPC effect for frequency biased items, nor did we observe that the CSPC effect increased from the first to the last half of the experiment. The evidence for CSPC effects across reproduction attempts using this design is clearly mixed.

One interpretation of the mixed results, as suggested by Hutcheon and Spieler (2016), is that the original findings from the second experiment of Crump and Milliken (2009) are spurious, are irreproducible, and reflect a Type I error. Another possibility is that the precision with which CSPC effects can be measured using this design is low, and that the pattern of mixed results across reproductions is an expected consequence of the design. By this view, reproductions should fail and succeed at rates commensurate with the power of the design to detect effects. As part of our effort to assess the reproducibility of CSPC effects from Crump and Milliken (2009) we also conducted a series of simulation analyses to estimate the proportion of reproduction attempts that would be expected to produce significant results.

Estimating reproducibility by Monte Carlo simulation

The collection of reproduction attempts offer an instructive example of the process, value, and expected outcomes of reproduction initiatives. They show some difficulties in making inferences about the presence or absence of effects based on reproduction alone. Reproductions can fail for several reasons. The original findings may be spurious. The reproduction attempt could be flawed experimentally. And, even when phenomena are real we should expect them to fail some of the time, because reproduction failure rate for real effects is one minus the power the design has to detect the effect.

In addition to conducting experiments, we were also interested in assessing the expected rates of reproduction success for the present designs with power analyses. We have already reported power estimates using the data reported by Crump and Milliken (2009). Using this method we estimated that 16 subjects were needed to achieve power of .8 to detect the CSPC effect for frequency unbiased items in the first design. Similarly, Hutcheon and Spieler (2016), and our own analysis, estimated that 32 subjects were needed to achieve power of .8 to detect the CSPC effect for frequency unbiased items in the second design. This approach uses the observed measures of the mean CSPC effect and CSPC effect variation to estimate power, and provides useful power estimates to the extent that those measures generalize to other experiments. For example, it is possible that the sample variance of the CSPC effect from an experiment grossly overestimates or underestimates the actual population variance of the CSPC effect. As a result, power estimates can be inflated or deflated to the extent they rely on estimates of effect size that are themselves inflated or deflated.

The present experiments showed smaller effect sizes than those reported by Crump and Milliken (2009), possibly owing to the fact that the present experiments were conducted under more variable conditions online. For example, the standardized effect size for the CSPC effect for frequency unbiased items in Experiment 1 of Crump and Milliken (2009) was $d = 0.72$, but only $d = 0.3$ (Experiment 1) and 0.24 (Experiment 3) for the present replications. Our present experiment had large enough N to detect the expected CSPC effects with high power based on the prior data. We also looked at observed power for each the experiments, and, based on the number of subjects and observed effect sizes in each experiment, we calculated the smallest CSPC effect that we could have detected with power = .8. For the frequency unbiased CSPC effect we found values of 18, 22, 15, and 22 (ms) for Experiments 1–4, respectively.

For the frequency biased CSPC effect we found values of 43 and 50 (ms) for Experiments 2 and 4, respectively. These values show that we had high power to detect the expected CSPC effects for frequency unbiased items, even though the observed effect sizes were smaller than those in prior reports.

The above approach uses the observed effect sizes to estimate power, and with enough replications these estimates should converge on their hypothetical population value. A drawback of that approach is the time and effort required to run enough replications to estimate the effect size accurately. A complementary approach is to estimate the effect size by Monte Carlo simulation using a statistical model of the distributions that underlie the effect of interest. We took this approach and conducted a simulation-based power analysis (Maxwell, Kelley, & Rausch, 2008). Of particular interest was the idea that the variance of CSPC effects could be estimated by sampling reaction times from base distributions that fit actual subject distributions. The CSPC effect is a difference score between two mean reaction time difference scores, so the variance of the CSPC effect depends on sampling the underlying reaction time distributions. So, rather than estimating power based on measures of effect size from prior experiments, we instead estimated properties of the base reaction time distributions from our current data, and sampled from these distributions using a Monte Carlo simulation approach to estimate reproduction success rates across variables such as mean CSPC effect and number of subjects in each experiment. The simulation exercise is insightful for examining both expectations about reproducibility and existing issues in the present designs that need to be addressed to improve the reliability and precision of measures of contextual control like the CSPC effect.

Monte Carlo simulations were conducted using R for both experimental designs used in the Crump and Milliken (2009) study. The general approach was to sample simulated reaction time data from distributions constructed with variances representative of normal subject populations and mean differences that would produce CSPC effects of desired sizes. In this way, data for individual subjects could be simulated at the single trial level in each design, and any analyses carried out on the present experiments could be conducted on the simulations. Across simulations we varied the CSPC effect size as well as number of subjects. Then, for each pair of effect size and *N* parameters, the probability of obtaining a significant finding (e.g.$p < .05$) was determined by Monte Carlo simulation of a thousand independent replications for that pair.

To generate reaction time distributions that were representative of our subject population, we fitted ex-Gaussian functions to RT distributions from individual subject data in the present experiments. The ex-Gaussian approach estimates the mean (mu) and standard deviation (sigma) of a Gaussian distribution, as well as the tau parameter from the exponential distribution, and characterizes the skewed nature of reaction time distributions. We used the R-package *retimes* (Massidda, 2013) to estimate the parameters and to sample simulated values from ex-Gaussian distributions with programmed values. More specifically, separate ex-Gaussian distributions were fitted to the congruent and incongruent reaction time distributions from each subject in the Stroop and flanker experiments. The mean parameter values of mu, sigma, and tau collapsed across all of the subjects are shown in Table 5, which splits groups of subjects by whether they completed a single-button press Stroop, a typing Stroop, or a flanker experiment.

Data for one simulated subject from a given experiment were created as follows. Each simulated subject was assigned a unique set of ex-Gaussian parameters. Each parameter was chosen by sampling from the

Table 5. Mean and standard deviations of ex-Gaussian parameters mu, sigma, and tau, fitted to individual subject correct reaction time distributions.

		Mu		Sigma		Tau	
Experiment	Congruency	*M*	*SD*	*M*	*SD*	*M*	*SD*
Stroop button	Con	612.5	134.8	82.6	43.3	172.8	64.4
	Inc	676.6	118.4	80.7	36.6	169.7	61.0
Stroop typing	Con	583.9	96.6	56.3	31.5	161.7	72.4
	Inc	703.6	87.7	61.3	28.8	153.2	62.3
Flanker	Con	625.4	105.3	68.6	38.9	166.3	74.6
	Inc	732.4	106.3	80.7	34.2	157.7	73.4

Note: *M* = mean; *SD* = standard deviation; Con = congruent; Inc = incongruent.

empirical cumulative distribution function computed for each parameter in each experiment. We simulated Stroop and CSPC effects as differences in mu between conditions. We chose a general congruency effect of 100 ms, so we added 100 to the mu value from the congruent distribution to create the mu value for the incongruent distribution. To model CSPC effects, we decreased mu for congruent items in the high proportion congruent condition by half of the CSPC effect, and increased mu for incongruent items in the high proportion congruent condition by the other half. So, for example, the values of 400 (LPC, C), 500 (LPC, I), 390 (HPC, C), and 510 (HPC, I) for mu (where LPC = low proportion congruent, HPC = high proportion congruent, C = congruent, I = incongruent) would reflect a programmed CSPC effect of 20 ms, which is the difference between the two programmed Stroop effects of 100 ms (LPC), and 120 ms (HPC). Simulated RTs were then sampled from each ex-Gaussian distribution according to their respective cell sizes (e.g. 48 each for the frequency unbiased items in Crump & Milliken, 2009). Finally, mean RTs in each condition were computed from the simulated trial data to calculate a simulated CSPC effect.

This method allowed us to simulate single subject data, and single experiment data by running multiple simulated subjects and testing the significance of resulting CSPC effects. More important, we could simulate the practice of replicating experiments by repeating an experimental simulation a thousand times and calculating the proportion of experiments that produce a significant CSPC effect. Thus, we adopted this method and ran several simulations that varied the size of CSPC effect from 10 to 30 ms, as well as the number of subjects in each experiment from 25 to 150. We ran separate simulations using the ex-Gaussian parameter values from single-button response Stroop, typing response Stroop, and flanker experiments. Also, we ran simulations for frequency unbiased items and frequency biased items from the second experimental design of Crump and Milliken (2009). The results of the simulations are displayed in Figure 1.

The simulations show estimates of the proportion of experiments producing a significant ($p < .05$) group-level CSPC effect, and mean Cohen's d for all of the experiments as a function of programmed values for N and mean CSPC score. First, assuming an effect size of 20 ms, the CSPC effect for frequency unbiased items replicates successfully 42%, 70%, 94%, and 99% of the time with $N = 25$, 50, 100, and 150, respectively. Second, the CSPC effect for frequency biased items from the second design can be expected to replicate much less often than the CSPC effect for frequency unbiased items. Frequency biased items in the second design had substantially fewer observations per cell for low-frequency items, which produces more variable base estimates of RTs, Stroop effects, and CSPC effects (see Figure 2).

The simulation results are instructive for several reasons. First, they show that attempts to replicate CSPC effects using the present designs can be expected to produce null results with a variety of rates, so the results of single replication attempts do not necessarily allow strong inferences to be made about the presence or absence of effects. Second, they show that the present designs are unlikely to be suitable for reliably measuring CSPC effects unless large N designs are employed, and that present designs would not be suitable for precisely measuring the CSPC effect at the level of an individual subject. Third, the simulation results presented here

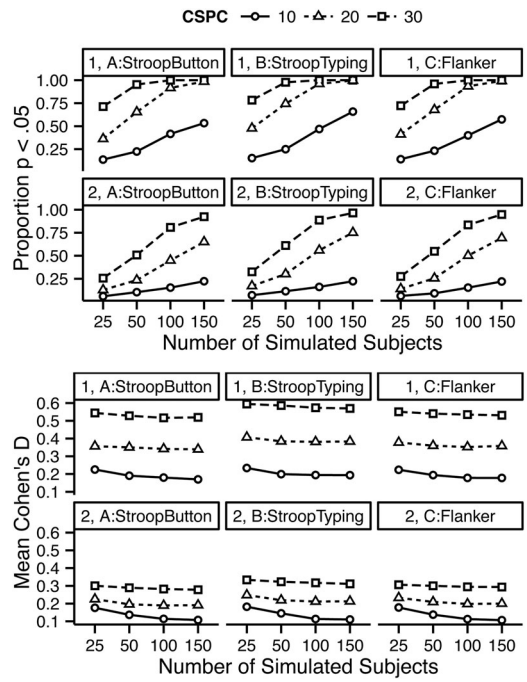

Figure 2. The graphs show the proportion of significant ($p < .05$) simulated experiments (1000 experiments per dot) and mean Cohen's d as a function of programmed values for the context-specific proportion congruent (CSPC) score and number of subjects. Within each graph, the top row (1) shows results for frequency unbiased items across Stroop and flanker tasks, and the bottom row (2) shows results for frequency biased item types based on the design of Experiment 2 from Crump and Milliken (2009).

demonstrate the utility of conducting simulations for reproducibility as a part of the exercise of constructing experimental designs. Specifically, by taking the variability of base reaction time distributions into account we produced more conservative power estimates that are more appropriate for the present data. As well, the simulation-based power estimates corresponded closely with our analysis of observed effect sizes. Finally, the simulation approach is highly flexible and can be used to estimate effect size, power, and sample size for a variety of designs that use a varying number of trials per condition.

General discussion

The focus of the present study was to assess the reproducibility of the CSPC effects reported by Crump and Milliken (2009). The study was motivated by recent failed attempts to replicate and reproduce CSPC effects for frequency unbiased items (Hutcheon & Spieler, 2016). These efforts have theoretical implications because the presence of CSPC effects for frequency unbiased items rules out classes of explanations that invoke frequency-sensitive learning processes. The important take-home message is that we find positive evidence of CSPC effects for frequency unbiased items in both Stroop and flanker tasks. As a result, our conclusion is that the critical result reported by Crump and Milliken, a CSPC effect for frequency unbiased items, was not a Type I error, and it continues to stand as an effect that cannot be explained by sole reference to frequency-sensitive learning processes.

At the same time, some of the results reported here did not reproduce those of Crump and Milliken (2009). Specifically, as with Hutcheon and Spieler (2016), we did not find clear evidence of CSPC effects for frequency unbiased items in the Stroop version of the second design—the design that included both congruent and incongruent frequency biased items in both location contexts. However, we did find CSPC effects for frequency unbiased items in the flanker version of this design. We also failed to find clear statistical support for the finding that CSPC effects for frequency unbiased items develop across the experiment, although some of the means were consistent with that trajectory. As part of our efforts, we conducted Monte Carlo simulations to estimate how often the present designs would be expected to replicate and found that the present designs would require much larger sample sizes than those used by Crump and Milliken to achieve greater than 80% replication success. That analysis also showed that the second design has low power to detect CSPC effects for frequency biased items.

Our reproducibility efforts were conducted online using web-browser technology to deploy tasks and Amazon's Mechanical Turk to recruit human subjects. This method probably introduces additional variance to our measures than would be expected from laboratory studies because subjects complete tasks in their own environment. The fact that we were able to successfully reproduce CSPC effects using these methods shows that these effects can be detected in less than optimal measurement conditions and further validates online tools for conducting attention and performance experiments (Crump et al., 2013).

The more general issue of reproducibility within psychological science is a timely topic, and one large-scale attempt to reproduce a wide berth of results across journals has shown generally low rates of reproducibility (Open Science Collaboration, 2015). In general, findings should replicate as a function of the power a design has to detect an effect of interest, and single replication efforts that fail can indicate Type II rather than Type I errors. Our study of reproducibility of the CSPC effect, together with that of Hutcheon and Spieler (2016), demonstrates how multiple replication efforts can be important for determining reproducibility.

To return to the more specific conceptual issues of the present work, we point out that measures of CSPC effects for frequency unbiased items using the Crump and Milliken (2009) designs are not the only available demonstrations that rule out frequency-sensitive learning accounts of CSPC effects. As already noted, King, Korb, et al. (2012) demonstrated CSPC effects using items that were unique across trials in a flanker task study. Two other studies have also produced CSPC effects (Cañadas, Rodríguez-Bailón, Milliken, & Lupiáñez, 2013) or CSPC-like effects (Crump, 2016) in flanker tasks with frequency unbiased items. Similar CSPC-like effects for frequency unbiased items have been shown in task-switching (Crump & Logan, 2010; Leboe, Wong, Crump, & Stobbe, 2008) and in masked-priming tasks (Heinemann, Kunde, & Kiesel, 2009). The fact that context-specific control effects have been observed for frequency unbiased items across this range of tasks bolsters confidence in the existence of these effects and undermines the view that frequency-sensitive learning processes on their own are sufficient to explain these contextual control phenomena.

Disclosure statement

No potential conflict of interest was reported by the authors.

Supplemental material

Supplemental material is available via the "Supplemental" tab on the article's online page (http://dx.doi.org/10.1080/17470218.2016.1206130).

The raw data for all experiments can be accessed at https://github.com/CrumpLab/CrumpBrosowskyMilliken_QJEP

References

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using qualtrics. *Behavior Research Methods*, *47*, 918–929.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

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., & 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*, 1–16, Article no. 367.

Cañadas, E., Rodríguez-Bailón, R., Milliken, B., & Lupiáñez, J. (2013). Social categories as a context for the allocation of attentional control. *Journal of Experimental Psychology: General*, *142*, 934–943.

Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*, 177–184.

Corballis, P. M., & Gratton, G. (2003). Independent control of processing strategies for different locations in the visual field. *Biological Psychology*, *64*, 191–209.

Crump, M. J. C. (2016). Learning to selectively attend from context-specific attentional histories: A demonstration and some constraints. *Canadian Journal of Experimental Psychology*, *70*, 59–77.

Crump, M. J. C., Gong, Z., & Milliken, B. (2006). The context-specific proportion congruent stroop effect: Location as a contextual cue. *Psychonomic Bulletin & Review*, *13*, 316–321.

Crump, M. J. C., & Logan, G. D. (2010). Contextual control over task-set retrieval. *Attention, Perception & Psychophysics*, *72*, 2047–2053.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), Article no. e57410.

Crump, M. J. C., & Milliken, B. (2009). The flexibility of context-specific control: Evidence for context-driven generalization of item-specific control settings. *The Quarterly Journal of Experimental Psychology*, *62*, 1523–1532.

Crump, M. J. C., Vaquero, J. M. M., & Milliken, B. (2008). Context-specific learning and control: The roles of awareness, task relevance, and relative salience. *Consciousness and Cognition*, *17*, 22–36.

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*, 374–380.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Heinemann, A., Kunde, W., & Kiesel, A. (2009). Context-specific prime-congruency effects: On the role of conscious stimulus representations for cognitive control. *Consciousness and Cognition*, *18*, 966–976.

Hutcheon, T., & Spieler, D. (2016). Context-driven control does not generalize to frequency unbiased stimuli. *Quarterly Journal of Experimental Psychology*. Epub online. doi:10.1080/17470218.2016.1182193

King, J. A., Donkin, C., Korb, F. M., & Egner, T. (2012). Model-based analysis of context-specific cognitive control. *Frontiers in Psychology*, *3*, 1–13, Article no. 358.

King, J. A., Korb, F. M., & Egner, T. (2012). Priming of control: Implicit contextual cuing of top-down attentional set. *The Journal of Neuroscience*, *32*, 8192–8200.

Leboe, J. P., Wong, J., Crump, M. J. C., & Stobbe, K. (2008). Probe-specific proportion task repetition effects on switching costs. *Perception & Psychophysics*, *70*, 935–945.

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

Logan, G. D., & Zbrodoff, N. J. (1998). Stroop-type interference: Congruity effects in colour naming with typewritten responses. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 978–992.

Massidda, D. (2013). *retimes: Reaction time analysis*. Retrieved from http://CRAN.R-project.org/package=retimes

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 1–8, aac4716.

Reimers, S., & Maylor, E. A. (2005). Task switching across the life span: effects of age on general and specific switch costs. *Developmental Psychology*, *41*, 661–671.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, *47*, 309–327.

Schmidt, J. R., & Besner, D. (2008). The stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 514–523.

Schmidt, J. R., Crump, M. J. C., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition*, *16*, 421–435.

Schmidt, J. R., Lemercier, C., & De Houwer, J. (2014). Context-specific temporal learning with non-conflict stimuli: Proof-of-principle for a learning account of context-specific proportion congruent effects. *Frontiers in Psychology*, *5*, 1–10, Article no. 1241.

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for collecting response

latencies in online studies of cognition. *PLoS ONE*, *8*, 1–12, e67769.

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon mechanical Turk and Adobe Flash. *Behavior Research Methods*, *1*, 95–111.

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

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*, 631–650.

Vecera, S. P., Cosman, J. D., Vatterott, D. B., & Roper, Z. J. (2014). The control of visual attention: toward a unified account. In R. H. Brian (Ed.), *Psychology of learning and motivation* (Vol. 60, pp. 303–347). Burlington: Academic Press.