Colling LJ, Szűcs D, De Marco D, et.al., A multilab registered replication of the attentional SNARC effect, *Advances in Methods and Practices in Psychological Science* (Forthcoming). Copyright © The Authors 2020. Reprinted by permission of SAGE Publications.

A multilab registered replication of the attentional SNARC effect

Lincoln J Colling[1], **Dénes Szűcs**[1], Damiano De Marco[1, 12], Krzysztof Cipora[2], Rolf Ulrich[2], Hans-Christoph Nuerk[2], Mojtaba Soltanlou[2], Donna Bryce[2], Sau-Chin Chen[3], Philipp Alexander Schroeder[4], Dion T Henare[5], Christine K Chrystall[5], Paul M Corballis[5], Daniel Ansari[6], Celia Goffin[6], H Moriah Sokolowski[6], Peter JB Hancock[7], Ailsa E Millen[7], Stephen RH Langton[7], Kevin J Holmes[8], Mark S Saviano[8], Tia A Tummino[8], Oliver Lindemann[9], Rolf A Zwaan[9], Jiří Lukavský[10], Adéla Becková[11], Marek A Vranka[11], Simone Cutini[12], Irene Cristina Mammarella[12], Claudio Mulatti[12], Raoul Bell[13], Axel Buchner[13], Laura Mieth[13], Jan Philipp Röer[14, 2], Elise Klein[15], Stefan Huber[15], Korbinian Moeller[15,2], Brenda Ocampo[16], Juan Lupiáñez[17], Javier Ortiz-Tudela[17], Juanma De la fuente[17], Julio Santiago[17], Marc Ouellet[17], Edward M Hubbard[18], Elizabeth Y Toomarian[18], Remo Job[19], Barbara Treccani[19], & Blakeley B McShane[20]

[1] Department of Psychology, University of Cambridge

[2] Department of Psychology, University of Tübingen

[3] Department of Human Development and Psychology, Tzu-Chi University

[4] Department of Psychiatry and Psychotherapy, University of Tübingen

[5] School of Psychology, University of Auckland

[6] Department of Psychology & Brain and Mind Institute, The University of Western Ontario

[7] Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, UK

[8] Department of Psychology, Colorado College

[9] Department of Psychology, Education & Child Studies, Erasmus University Rotterdam, Netherlands

[10] Institute of Psychology of the Czech Academy of Sciences
[11] Department of Psychology, Faculty of Arts, Charles University
[12] Department of Developmental Psychology, University of Padova
[13] Department of Experimental Psychology, Heinrich Heine University Düsseldorf
[14] Department of Psychology and Psychotherapy, Witten/Herdecke University
[15] Leibniz-Institut für Wissensmedien, Tübingen
[16] School of Psychology, The University of Queensland
[17] Research Center for Mind, Brain, and Behavior, University of Granada
[18] Department of Educational Psychology, University of Wisconsin-Madison
[19] Department of Psychology and Cognitive Science, University of Trento
[20] Kellogg School of Management, Northwestern University

Author Note

Correspondence concerning this article should be addressed to **Dénes Szűcs**, Downing Street, CB2 3EB, Cambridge, UK. E-mail: ds377@cam.ac.uk

Abstract

The attentional Spatial-Numerical Association of Response Codes (att-SNARC) effect (Fischer et al., 2003; Nature Neuroscience)—the finding that participants are quicker to detect left-side targets when the targets are preceded by small numbers and quicker to detect right-side targets when they are preceded by large numbers—has been used as evidence for *embodied* number representations and to allow for strong claims about the link between number and space (e.g., a mental number line). We attempted to replicate Study 2 of Fischer et al. (2003) by collecting data from 1105 participants across seventeen labs. Across all 1105 participants and four ISI conditions, the proportion of times the direction of the observed effect was consistent with the original effect was 0.50. Further, the effects we observed both within and across labs were minuscule and incompatible with those observed in Fischer et al. (2003). Given this, we conclude that we have *failed* to replicate the effect reported by Fischer et al. (2003). In addition, our analysis of several participant-level moderators (finger counting preferences, reading/writing direction experience, handedness, and mathematics fluency and mathematics anxiety) revealed no substantial moderating effects. Our results demonstrate that the att-SNARC effect cannot be used as evidence to support the strong claims about the link between number and space discussed above

A multilab registered replication of the attentional SNARC effect

Introduction

A foundational issue in cognitive science is the question of how we *represent* concepts. Classical approaches to cognitive science, exemplified by Fodor's (1975) "language of thought" and Newell and Simon's (1976) "physical symbol systems" hypothesis, view representations as abstract or amodal and as distinct from sensorimotor processing. In contrast to these traditional views, a range of other views that go under labels such as "embodied", "situated", or "grounded" cognition argue that representations (i) are intimately linked to sensorimotor processing (see, e.g., Wilson, 2002, for an overview); (ii) are analogue rather than symbolic; and (iii) represent by in some sense resembling their targets (e.g., see Gładziejewski & Miłkowski, 2017; Williams & Colling, 2018).

One area of research that has provided a wealth of empirical findings valuable for debates about the nature of concept representation has been numerical cognition. In fact, Fischer and Brugger (2011) have referred to numerical cognition as the "prime example of embodied cognition". In particular, Fischer and Brugger (2011) point to tasks examining spatial-numerical associations to make their case.

Researchers have long reasoned that numbers might be represented in a spatially organised manner (Galton, 1880), for example, as a *mental number line* (e.g., Restle, 1970). Key support for this notion comes from a series of nine experiments conducted by Dehaene, Bossini, and Giraux (1993). In these experiments, Dehaene et al. (1993) asked participants to judge whether the parity of a number was odd or even, finding that responses to large numbers were faster when pressing a right-hand key relative to a left-hand key while the opposite was true for small numbers. They labelled this number magnitude by response side interaction the Spatial-Numerical Association of Response Codes (SNARC) effect.

In these parity judgement experiments, there was no standard with which to compare the presented number. Consequently, whether a particular number was responded to quicker with the left hand or the right hand was not determined by the absolute magnitude of the number, but by the relative magnitude of the number within a stimulus set. Thus, the number five was responded to more quickly with the left hand when appearing in a set of numbers ranging from four to nine but more quickly with the right hand when appearing in a set of numbers ranging from zero to five (e.g., Dehaene et al., 1993; Fias, Brysbaert, Geypens, & d'Ydewalle, 1996).

Dehaene et al. (1993) reported that the effects were neither dependent on the handedness of participants nor the hand used to make the response. Instead, they tracked the side of space of the response, with responses to small numbers being quicker with the right hand when the participants' hands were crossed (see, however, Wood, Nuerk, & Willmes, 2006). Nonetheless, Dehaene et al. (1993) did report that the effect was dependent on reading/writing direction. Specifically, while they initially found the effect in French participants who had experience reading/writing from left to right, the did not replicate it in a follow-up experiment with Iranian participants who had experience reading/writing from right to left (see Shaki, Fischer, and Petrusic (2009) and Zebian (2005)). Together, the results of the nine experiments reported in Dehaene et al. (1993) were taken to support the idea of a mental number line with numbers of increasing magnitude associated with the left-to-right axis of external space.

While SNARC effects appear to be robust (see Wood, Willmes, Nuerk, and Fischer (2008) and Toomarian and Hubbard (2018) for recent reviews), the great range of findings has resulted in some debate about the underlying mechanism(s) that produce them. One such debate is concerned with whether the SNARC effect is produced by early, *response-independent* mechanisms or whether processes at the stage of *response selection* are responsible. According to theories that place the origin of the SNARC effect at an early stage, the mere observation of the number should be sufficient to activate the spatial code because the spatial code is intimately connected to the numerical representation. Consequently, these theories make the strongest claims about the link between number and space. Theories that place the origin of the SNARC effect at the response selection stage, however, make weaker claims about the connection between number and space. As Pecher and Boot (2011) note, if the response selection stage gives rise to the SNARC effect, then no underlying spatial-numerical representation need be assumed.

Most recent work has tended to support the notion that the response selection stage is the locus of the SNARC effect. In particular, Keus and colleagues have used both behavioural (Keus & Schwarz, 2005) and psychophysiological (Keus, Jenks, & Schwarz, 2005) evidence to argue in favour of a later, response-related origin of the SNARC effect. Further support comes from a computational model that relies on task-dependent conceptual coding of the number at a stage distinct from the numerical representation itself (Gevers, Verguts, Reynvoet, Vaessens, & Fias, 2006).

Additional accounts that break the link between number, space, and the SNARC effect are so-called response polarity-related accounts. Specifically, Proctor and Cho (2006) argue that on binary classification tasks, items in the task set are coded as being positive or negative in polarity. Response selection can then be facilitated when there is a structural overlap between the polarity of the item (the number in the case of the SNARC effect) and the response. As with the model from Gevers et al. (2006), the account of Proctor and Cho (2006) does not require any perceptual or conceptual overlap between the stimulus and the response dimensions for the SNARC effect to occur. That is, these accounts do not rely on the notion of a mental number line or sensorimotor-linked representations. A range of empirical findings support these types of accounts. For example, Santens and Gevers (2008) found that SNARC-like effects can be produced when left-right responses are replaced with unimanual close-far responses, with small numbers associated with close responses and large numbers associated with far responses. Further, Landy, Jones, and Hummel (2008) found that verbal “Yes” and “No” responses on a parity judgement task were facilitated by large numbers and small numbers respectively.

Finally, still other researchers have argued in favour of a working memory account of the SNARC effect. For example, in the task reported by van Dijck and Fias (2011), participants performed a fruit/vegetable classification after having been encouraged to store the stimuli as an ordered set in working memory. This was done by presenting participants with a sequence of fruit and vegetable names (displayed in the centre of the screen) before the classification task and then testing them on the order of the items. A spatial response-compatibility effect emerged with participants responding faster to items early in the sequence with their left hand and items later in the sequence with the right hand. van Dijck and Fias (2011) argue that this working memory account can also explain why SNARC-like effect emerge for other kinds of ordinal sequences such as months of the year (Gevers, Reynvoet, & Fias, 2003) or days of the week (Gevers, Reynvoet, & Fias, 2004) as well as why spatial-numerical associations can be moderated by giving participants instructions to associate numbers with positions on a clock-face (1–5 on the right and 6–10 on the left) rather than on a ruler (1–5 on the left and 6–10 on the right; Bächtold, Baumüller, & Brugger, 1998)

Given that several competing accounts of the SNARC effect exist and that many of the accounts do not require a mental number line, one may doubt whether spatial-numerical associations provide evidence for anything like "embodied" number representations or number representation that are intimately linked with space. However, there is evidence that does support an early, response-independent locus for the SNARC effect and thus does provide support for the notion of a mental number-line and spatially-linked number representation—the modified version of the Posner (1980) attentional cueing task developed by Fischer, Castel, Dodd, and Pratt (2003). In this study, participants were asked to detect the appearance of lateralised targets. The target, a white circle, was preceded by either a small number (1 or 2) or a large number (8 or 9). Importantly, the digit did not predict the subsequent location of the target, that is, it was not task-relevant. Instead, the task was merely to press a single response button when the target appeared regardless of whether it appeared on the left or the right. Importantly, not requiring a spatially lateralised response negates the possibility of any response-related effects. The finding from this paradigm was consistent with the SNARC effect, as participants were quicker to detect left-side targets when the targets were preceded by small numbers and quicker to detect right-side targets when they were preceded by large numbers, at least for digits and targets that were separated by an inter-stimulus interval (ISI) between 250 and 1000 ms. This finding—named the attentional SNARC (att-SNARC) effect—suggests that viewing numbers alone was able to cue spatial attention either to the left or the right depending on the magnitude of the number.

Because the att-SNARC effect argues strongly in favour of an early, response-independent locus for the cause of the SNARC effect, the att-SNARC effect plays a crucially important role in adjudicating debates about the origin of the SNARC effect and the nature of number representations. As a result, the original finding has been extremely influential (e.g., cited 704 times according to Google Scholar as of 12 September 2019). However, subsequent attempts to replicate the effect have produced mixed results.

Galfano, Rusconi, and Umiltà (2006) report a statistically significant effect for right-side targets when the data was collapsed across two ISI conditions of 500 and 800 ms using a one-tailed test [Estimate = 6.5 ms; $t(25) = 1.75$; $p = .046$ (reported as $p = .04$)]. They also report a statistically significant effect for left-side targets collapsed across the two ISI conditions, but the claimed statistical significance reflected a reporting error [Estimate = 5.5 ms; $t(25) = 1.59$; $p = .062$ (reported as $p = .04$)]. Finally, they report an overall estimate (collapsed across the left and right target locations) of 8 ms for the 500 ms ISI condition and 4 ms for the ISI 800 ms condition, but the reporting is such that the corresponding variances or test statistics for these estimates cannot be obtained.

In addition, Dodd, Van der Stigchel, Leghari, Fung, and Kingstone (2008) report a statistically significant effect when the data was collapsed across three ISI conditions between 250 and 750 ms and across both left and right target locations, but again the claimed statistical significance reflected a reporting error [Estimate = 5.5 ms; $F(1,29) = 4.05$; $p = .054$ (reported as $p < .05$)]. At the level of individual inter-stimulus intervals, they report statistically significant effects at 500 ms for right-side targets [Estimate = 6 ms; $t(29) = 2.34$; $p = .013$] and left-side targets [Estimate = 16 ms; $t(29) = 2.48$; $p = .010$]. Finally, they report estimates of 6 ms for the 250 ms ISI condition, 11 ms for the 500 ms ISI condition, and -0.5 ms for the 750 ms ISI condition (collapsed across left and right target locations), but the reporting is such that the corresponding variances or test statistics for these estimates cannot be obtained.

Ristic, Wright, and Kingstone (2006) also report a statistically significant effect when the data was collapsed across six ISI conditions between 350 and 800 ms and across right and left side targets [Estimate = 3.79 ms; $F(1,17) = 5.48$; $p = .032$]. Although it is possible to obtain point estimates for each of the six inter-stimulus intervals [350 ms ISI = 11.24 ms; 400 ms ISI = 2.81 ms; 500 ms ISI = -1.44 ms; 600 ms ISI = 6.17 ms; 700 ms ISI = 6.05 ms; 800 ms ISI = -2.17 ms] (collapsed across left and right target locations), the reporting is such that the corresponding variances or test statistics for these estimates cannot be obtained.

Several other failed replications have also been reported. Zanolie and Pecher (2014) report two experiments that failed to find a statistically significant effect when collapsed across four ISIs between 250 and 750 ms and across left and right side targets [Experiment 1: No estimates reported; $F(1,19) = 0.03$, $p = .863$; Experiment 2: No estimates reported; $F(1,23) = 0.13$, $p = .772$]. Ranzini, Dehaene, Piazza, and Hubbard (2009) also failed to find a statistically significant effect when collapsed across three ISIs between 300 and 500 ms and across left and right side targets [Estimate = 3 ms; $F(1,14) = 4.1$, $p = .06$]. Salillas, El Yagoubi, and Semenza (2008) failed to find a statically signifiant effect at a 400 ms ISI when collapsed across left and right side targets [Estimate = 2 ms; $F(1,11) = 1.3$, $p = .28$]. More recently, van Dijck, Abrahamse, Acar, Ketels, and Fias (2014) failed to find an effect collapsed across four ISIs between 250 and 1000 ms and left and right side targets [Experiment 1: No estimates reported] and three ISIs between 100 and 700 ms [Experiment 2: No estimates reported; $F(1,28) = 2.94$, $p = .097$]. While Fattorini, Pinto, Rotondaro, and Doricchi (2015) failed to find an effect collapsed across two ISI of 500 and 750 ms and across left and right side targets [Experiment 1: No estimates reported; $F(1,59) = 1.69$, $p = .20$] and four ISIs between 250 and 1000 ms [Experiment 2: No estimates reported; $F(1,31) = 1.5$, $p = .22$]. The final two studies by van Dijck et al. (2014) and Fattorini et al. (2015) are particularly notable for their large sample size.

It should be noted that alternative accounts of the effect reported by Fischer et al. (2003) have been suggested. These alternative accounts include, for example, accounts based on working memory (van Dijck et al., 2014). Similarly, manipulations that make explicit associations between number and space have also been able to produce att-SNARC-like effect (e.g., Fattorini et al., 2015, Experiment 3). However, because of these modifications, the findings of these studies have different theoretical implications to the att-SNARC and, therefore, they will not be discussed here. Instead, the focus of the present work will be on the att-SNARC as originally proposed.

In sum, prior studies have demonstrated—at best—only qualified and partial success at replicating Fischer et al. (2003). That said, one might argue that failure to replicate Fischer et al. (2003), reflects more the definition of replication employed—namely one based on statistical significance—than any true failure to replicate the scientific hypothesis as opposed to the statistical hypothesis examined by Fischer et al. (2003). As we discuss in greater depth below, we are sympathetic to this view and prefer alternative operationalisations of replication.

One component of such a better approach to assessing replication might involve synthesising the evidence across all published studies of the effect via meta-analysis in order to estimate, for example, an overall average effect size, the heterogeneity in effect sizes across studies, and the effects of potential moderators at the study-level or otherwise. However, this is complicated because (i) the statistical significance of a study's results typically impacts whether or not the study is published therefore resulting a set of published studies that is not representative and (ii) meta-analytic results are biased when the set of studies analysed is not representative (McShane, Böckenholt, & Hansen, 2016; Ioannidis, 2008).

Given this, the Registered Replication Report (RRR) format that we pursue here provides an ideal means of assessing the att-SNARC effect because results from all participating labs are included in the meta-analysis regardless of the results. Further, pre-registration of the primary hypotheses and statistical analyses further mitigates many potential biases.

An additional benefit of an RRR is that it allows for the investigation of potential moderator variables not previously considered thereby shedding light on mechanism and perhaps also the current mixed record of replication success. Consequently, in addition to replicating the experimental protocol of Fischer et al. (2003), we investigate several variables that could potentially moderate the att-SNARC effect including finger counting habits, reading/writing direction, handedness, mathematics ability, and mathematics anxiety (see Fischer (Fischer, 2006; Fischer, 2008), Fischer and Knops (2014), Georges, Hoffmann, and Schiltz (2016), and Shaki et al. (2009) for details and conjectures).

Methods

Design

Sample size

Each participating lab was required to provide a target sample size and stopping rule on a lab-specific OSF page (https://osf.io/7zyxj/), with labs agreeing to a minimum target sample size of sixty participants. We chose sixty as the minimum because it provides more than adequate power (0.92 using a one-tailed test at $\alpha = 0.05$) assuming an effect size on the standardised Cohen's *d* scale of 0.4, about the midpoint of previously published estimates. This corresponds to a raw effect size of 6 ms assuming a between-participant standard deviation of 15 ms, again about the midpoint of previously published estimates.

Due to time constraints, not all labs were able to reach the minimum target of sixty (see Table 1 for sample sizes achieved by each lab). However, again assuming an effect size of 0.4, we would expect to see a statistically significant effect in 93% of the labs (i.e., about sixteen) given the sample sizes actually achieved. Given this, if 0.4 is a reasonable estimate of the effect size and there are no substantial moderators of the effect, we would expect statistically significant effects not only at the meta-analytic level but also at the level of the individual lab.

Materials

The participating labs all had: (i) a testing station, such as a room or a cubicle, where participants could undertake the experiment without distraction; (ii) a computer for presenting stimuli and recording responses; (iii) a chin rest or similar device to ensure that the participant remained a set distance from the computer monitor; and (iv) a tape measure for use in the screen calibration process. Five labs also optionally made use of an eye-tracker to record participants' eye movements during the replication task; see the lab-specific OSF pages for details.

Additionally, an instruction booklet detailing how to perform the setup and calibration procedure and the finger counting assessment was provided. The materials were initially written in English. The experiment was also conducted in German, Dutch, Czech, Spanish, Italian, and Chinese, reflecting the predominant language in the locale of the individual labs; for these labs, the English language instructions were translated into the new language and then independently back-translated into English to ensure accuracy.

All materials including translations are available on OSF (see https://osf.io/7zyxj/). To perform analyses, we used R (Version 3.5.1; R Core Team, 2018) and the R packages *bindrcpp* (Version 0.2.2; Müller, 2018), *checkmate* (Version 1.8.5; Lang, 2017), *dplyr* (Version 0.7.6; Wickham, François, Henry, & Müller, 2018), *forcats* (Version 0.3.0; Wickham, 2018a), *forestplot* (Version 1.7.2; Gordon & Lumley, 2017), *ggplot2* (Version 3.0.0; Wickham, 2016), *glue* (Version 1.3.0; Hester, 2018), *kableExtra* (Version 0.9.0; Zhu, 2018), *knitr* (Version 1.20; Xie, 2015), *lme4* (Version 1.1.18.1; Bates, Mächler, Bolker, & Walker, 2015), *magick* (Version 1.9; Ooms, 2018), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.14; Bates & Maechler, 2018), *nlme* (Version 3.1.137; Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *purrr* (Version 0.2.5; Henry & Wickham, 2018), *pwr* (Version 1.2.2; Champely, 2018), *R.matlab* (Version 3.6.2; Bengtsson, 2018), *readr* (Version 1.1.1; Wickham, Hester, & Francois, 2017), *reticulate* (Version 1.10; Allaire, Ushey, & Tang, 2018), *stringr* (Version 1.3.1; Wickham, 2018b), *tibble* (Version 1.4.2; Müller & Wickham, 2018), *tidyr* (Version 0.8.1; Wickham & Henry, 2018), and *tidyverse* (Version 1.2.1; Wickham, 2017).

Procedure

We employed an experimental paradigm based on Experiment 2 of Fischer et al. (2003). We chose Experiment 2 over Experiment 1 because it had fewer ISI conditions and because the results were statistically significant in a greater proportion of conditions. Before starting data collection, each lab performed a monitor calibration procedure using a supplied calibration script which involved measuring the viewing distance and the size of standard stimuli presented on the screen; see OSF for details. After participants provided informed consent, they were seated in front of a computer monitor with their heads placed into a chin rest that was located a fixed distance from the monitor (set during the calibration procedure) and then data collection commenced.

The standard trial structure, which is identical to that of Fischer et al. (2003) and which does not contain timing modifications for the eye-tracker (see below for details), is shown in Figure 1. The initial display of each trial contained a centrally presented white fixation point on a black background (0.2° diameter), and two white boxes (1° × 1°) presented on either side of the fixation point. The centres of the boxes were located 5° from the centre of the fixation point. This initial display was shown for 500 ms. Following the initial display, a digit (1, 2, 8, or 9; 0.75° height) was presented at a fixed duration of 300 ms. After the digit was removed, the fixation point reappeared for a variable duration (250 ms, 500 ms, 750 ms, or 1000 ms). This was followed by a circular white target (0.7° diameter) appearing in either the left- or right-side box on target trials or no target appearing on catch trials.

Target trials ended after a response was made or 1000 ms after target onset, whichever came first. Catch trials ended 1000 ms after the digit was removed. Trials automatically advanced and were separated by an inter-trial interval of 1000 ms.

Participants responded by pressing the spacebar with the preferred hand. Participants who responded before the target appeared or who responded on a catch trial were presented with the warning "Too quick! Please wait until the target appears in a box before pressing SPACE" [English text] and the trial ended. Participants who failed to respond on a target trial were presented with the warning "Too slow! Please press SPACE as soon as the target appears". Participants who erred on more than 5% of trials were excluded from analyses.

Participants performed a total of 800 trials (640 target trials and 160 catch trials), split into five blocks of 160 trials each with 128 target trials and 32 catch trials per block; each block contained an equal number of trials for each ISI, digit, and target location, and these were presented in a random order.

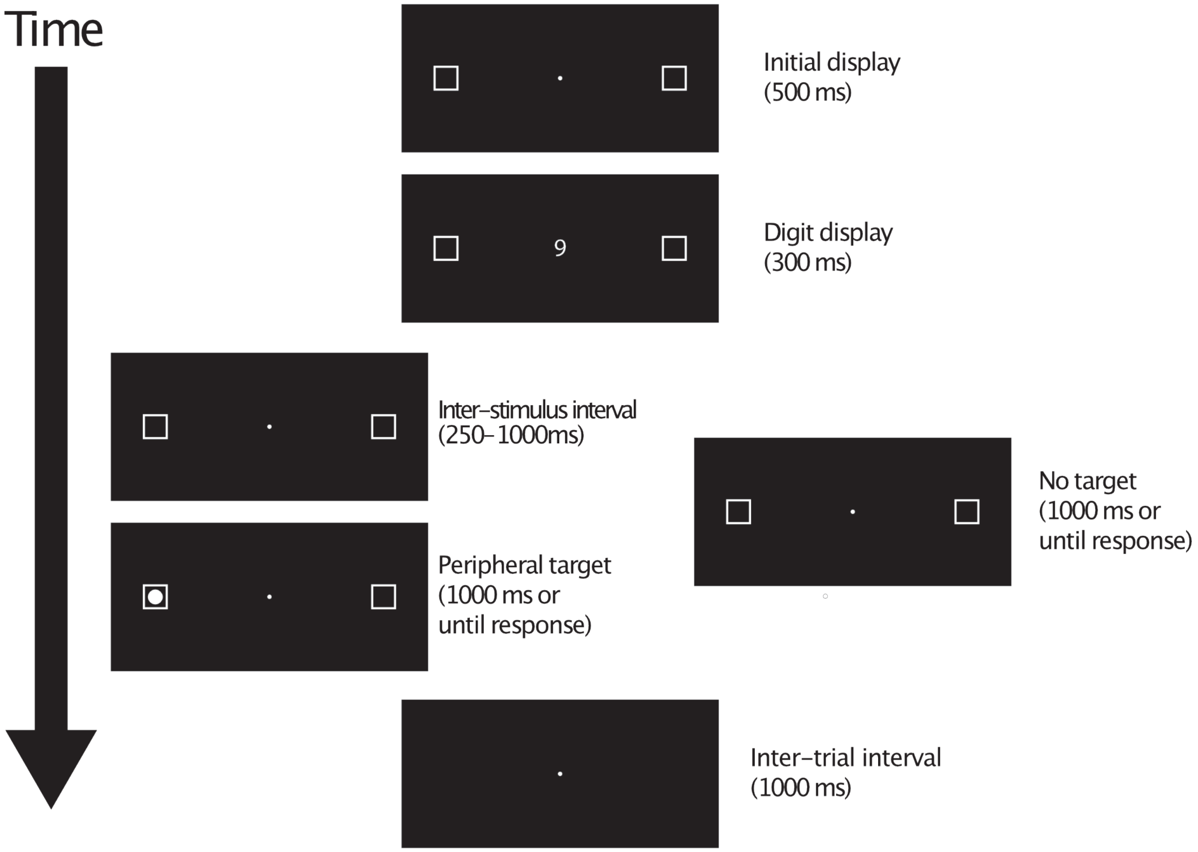


Figure 1. Outline of the trial structure for target trials and catch trials.

Eye-tracking protocol

Code implementing an eye-tracking protocol using an EyeLink 1000 eye-tracker was provided to all labs (and is available at https://github.com/ljcolling/FischerRRR-eyetracking). For labs using an eye-tracker other than an EyeLink 1000, deviations from the standard protocol are listed on the lab-specific OSF page. The standard nine-point grid was used for calibration and validation at the start of each block or when required during a block. The start of trials was triggered after the detection of 500 ms of stable fixation within a 2° box centred on the fixation point. If the system could not detect a stable fixation within a 2000 ms time window, the calibration process was repeated. After the digit was presented, and before the target appeared, the gaze position was monitored and any deviations outside a

1° box centred on the fixation point were recorded. Any deviations towards the lateral boxes that exceeded 2° resulted in the trial being marked as contaminated. These trials were excluded from primary analyses; however, they were analysed separately to attempt to determine any possible effect of eye movements on the results.

Finger counting

The finger counting assessment was derived from the task developed by Lucidi and Thevenot (2014). Participants were asked to read aloud four sentences while counting the number of syllables in each. As reading aloud prevents prevents participants from verbalising counting, most participants would need to resort to finger counting while sounding out the syllables. For each sentence, the experimenter recorded the first finger and first hand the participant used. While most participants used their fingers for the task, some participants did not use their fingers and instead adopted a different strategy. Participant who failed to engage in finger counting after two sentences were prompted to do so. Details of the prompting were recorded in lab logs. See OSF for details.

The results of the finger counting task were used to place participants into one of five groups: left-starters, right-starters, left-prefer, right-prefer, and no group. The finger counting group was determined not only by participants' hand preferences but also by how consistently they engaged in finger counting. The left- (right-)starter group was defined as those who counted using a hand on all four occasions and used the left (right) hand on at least three of them. The left (right)-prefer group was defined as those who counted using a hand on two or three occasions and used the left (right) hand on at least two of them. The no group group was defined as all other participants (for example, those who did not count on their fingers, those who only counted on their fingers once, and those who counted an equal number of times with each hand).

Reading/writing direction

Reading and writing direction was determined with a simple three option questionnaire asking if participants had experience with languages that are written exclusively from left to right (e.g., English and German), not exclusively left to right (e.g., Hebrew), or both types (see https://osf.io/he5za/ for details). This was used to cluster participants into two groups: exclusively left-to-right readers/writers and not exclusively left-to-right readers/writers.

Handedness

To assess handedness, we used the 10-item questionnaire from Nicholls, Thomas, Loetscher, and Grimshaw (2013). In labs conducting the experiment in a language other than English, the questionnaire was translated and some questions were replaced with more culturally appropriate versions when required (see https://osf.io/he5za/ for details).

Mathematics assessment

To assess mathematics fluency, we used the short mathematics assessment employed by Tibber et al. (2013). This test is adapted from the Mathematics Calculation Subtest (WJ-RCalc) of the Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock & Johnson, 1989). It contains twenty-five multiple choice mathematics questions requiring addition, subtraction, multiplication, and division. Participants had thirty seconds to select the response on each trial, with the timing controlled by the computer software. A countdown timer was stationed in the top left of the screen to inform participants of the time remaining. The twenty-five questions were split into five levels of five questions. Two errors on a single level or errors on consecutive levels terminated the test. The final score was the total number of correct answers.

Mathematics anxiety

Mathematics anxiety was assessed using the Abbreviated Math Anxiety Scale (AMAS; Hopko, Mahadevan, Bare, & Hunt, 2003). The AMAS contains nine questions that ask participants to rate (on a one to five scale) how anxious they would feel during particular events including thinking of an upcoming mathematics test, sitting a mathematics examination, and listening to a mathematics lecture. In labs conducting the experiment in a language other than English, the AMAS was translated (see https://osf.io/dhnf8/ for details). The final score was the sum of the individual ratings, with scores ranging from nine (low anxiety) to forty-five (high anxiety).

Exit questionnaire

An exit questionnaire that asked participants to describe the purpose of the experiment was used to determine whether participants could guess the purpose of the experiment. Participants who correctly guessed the purpose of the experiment, as judged by the experimenter, were excluded from primary analyses; however, they were analysed separately to determine whether this moderated the effect.

Exclusion criteria

Participants whose reaction time data contained more than 5% catch trial errors, who correctly guessed the purpose of the experiment or who who did not undertake all additional assessments were excluded from the analysis as per our pre-registration plan (see https://osf.io/6a2ny/).

Analysis

The dependent variables of interest were the congruency effect at each of the four ISI conditions (i.e., 250 ms, 500 ms, 750 ms, and 1000 ms). This is defined as the average difference in response time between congruent and incongruent targets, with congruent targets being defined as left targets preceded by low digits and right targets preceded by high digits and incongruent targets being defined as left targets preceded by high digits and right targets preceded by low digits. A positive value for the congruency effect indicates that participants were faster at responding to congruent targets relative to incongruent targets, and a negative value indicates the reverse.

We analysed our data via multilevel multivariate meta-analytic models (McShane & Böckenholt, 2018). Such models have at least two advantages over the standard random effects meta-analytic model. First, they better account for the dependence between our multiple dependent variables (i.e., the congruency effect at each of the four ISI conditions). Second, rather than assuming a simple two-level structure, with participants nested within labs, they can account for more complex nesting structures such as participants nested within with moderator groups (e.g., left-starters, right-starters, etc.) and moderator groups nested within within labs. In short, the standard approach necessitates treating several variance components as zero, thereby making unwarranted independence assumptions.

For each analysis, we consider several simplifications to the equal allocation multilevel multivariate compound symmetry specification detailed in McShane and Böckenholt (2018); we also consider an equal variance version of the single correlation equal allocation multilevel multivariate compound symmetry specification that, using the notation of that paper, sets the $\sigma_{d,d}$ equal for all dependent variables d (i.e., the congruency effect at each of the four ISI conditions). We chose among the six specifications via the Akaike Information Criterion (AIC; Akaike, 1974).

In analysing the effect of moderators, it would be ideal to consider them jointly within a single model. However, this would require a sufficient number of participants in each moderator group. Specifically, a minimum number of five participants is necessary to compute a 4 $\times$ 4 covariance matrix of full rank (i.e., corresponding to the congruency effect at each of the four ISI conditions) as required. Therefore, the decision on whether to consider all the moderators jointly within a single model or separately in different models was left until the sample sizes were known.

Unfortunately, data sparsity prevented us from considering all the moderators jointly in a single model: when considered jointly, many combinations of moderators (e.g., finger counting, reading/writing direction, handedness) result in either zero or very few participants per moderator group; indeed, this is also the case for some moderators (i.e., reading/writing direction and handedness) when considered alone as can be seen in Supplementary Tables A4 and A6 respectively. Consequently, we consider each moderator separately analysing only moderator groups with a minimum of five participants. All analyses were pre-registered (see https://osf.io/6a2ny/) and carried out in accordance with this plan.

For models featuring no moderators (Model 1) or discrete moderators (finger counting, reading/writing direction, and handedness; Models 2–4 respectively), we analysed the data at the moderator group level as per McShane and Böckenholt (2018). For the model featuring continuous moderators (mathematics fluency and mathematics anxiety; Model 5), we analysed the data at the participant level using an analogous specification (see below for details). Our motivation for considering these moderators and predictions follow as applicable.

Model 1: No Moderators. Fischer et al. (2003) suggests a positive congruency effect. The purpose of Model 1 was to assess this by replicating the analysis performed by Fischer et al. (2003); consequently, it did not account for any moderators.

Model 2: Finger counting. Recent work suggests that spatial-numerical compatibility effects in general (Fischer, 2008)—including attentional cueing effects in response to numbers (Fischer & Knops, 2014)—might be moderated by finger counting behaviour, specifically being stronger among those who start finger counting on the left hand and weaker or possibly even reversed among those who start finger counting on the right hand. The purpose of Model 2 was to assess this and consequently it took account of the finger counting moderator.

This model used only data from participants who consistently engaged in finger counting and consistently started on the same hand, that is, participants categorised as left-starters or right-starters. We restricted the analysis to these two groups principally because, if the finger counting moderator is to have an effect, then we would expect it to be most prominent in those whose finger counting is clear and unambiguous.

Model 3: Reading/writing direction. Recent work suggests that the congruency effect might be weaker or possibly even reversed among those who have experience with languages that are not read/written exclusively from left to right (Fischer, 2008; Shaki et al., 2009). The purpose of Model 3 was to assess this and consequently it took account of the reading/writing direction moderator. Specifically, participants were placed into two groups based on the reading/writing questionnaire: those who read/wrote exclusively left to right and those who did not.

Model 4: Handedness. The purpose of Model 4 was to assess whether handedness moderates the congruency effect and consequently it took account of the handedness moderator. Specifically, participants were classified as left-handed or right-handed according to the handedness questionnaire.

Model 5: Mathematics fluency and mathematics anxiety. Recent work suggests that numerical abilities (Fischer, 2006) and mathematics anxiety (Georges et al., 2016) may influence the strength of spatial-numerical associations. The purpose of Model 5 was to assess this and consequently it jointly took account of both mathematics fluency and mathematics anxiety as measured by the maths test and AMAS respectively.

Specifically, we fit a multilevel model to the participant-level congruency effects at each of the four ISI conditions; fixed effects were included for the full set of ISI Condition × Maths test × AMAS interactions and random effects were included for (i) each participant, (ii) each Lab × ISI Condition (with equal variance and zero correlation), and (iii) independently each Lab × Maths test, Lab × AMAS, and Lab × Maths test × AMAS.

Secondary analyses. The purpose of our secondary analyses was to assess whether insight into the purpose of the experiment or eye movements moderate the congruency effect. Specifically, Model 1 was refit separately to data from participants who correctly guessed the purpose of the experiment and to data from eye movement contaminated trials from participants with contaminated trials at each ISI × congruency condition.

Results

Replication operationalisation

The common definition of replication employed in practice is that a subsequent study is considered to have successfully replicated a prior study if either both failed to attain statistical significance or both attained statistical significance and were directionally consistent. This definition has been applied analogously in large-scale replication projects like the present one by comparing the results of a meta-analysis of the replication studies to the original study in terms of statistical significance.

However, the null hypothesis significance testing paradigm upon which this operationalisation of replication is based has been the subject of no small amount of criticism over the decades (see, for example, Rozenboom, 1960; Meehl, 1978; Cohen, 1994; Gelman, Carlin, Stern, & Rubin, 2003;

McShane & Gal, 2016; McShane & Gal, 2017) and recent calls to abandon it abound (Amrhein, Trafimow, & Greenland, 2019; McShane, Gal, Gelman, Robert, & Tackett, 2019; Wasserstein, Schirm, & Lazar, 2019; Amrhein, Greenland, & McShane, 2019). Further, recent work discussing alternative statistical paradigms specifically in the context of replication (Colling & Szűcs, 2018) has called for a better understanding of how statistical inference relates to scientific inference. A key point is that any assessment of whether a theory is supported by data depends on whether the magnitude of the observed effect is consistent with the theory (Gelman & Carlin, 2014). Consequently, in assessing replication, we distinguish between *statistical hypotheses* and *scientific hypotheses* and focus on that latter. Specifically, in discussing our results, we do so in light of the scientific hypothesis advanced by Fischer et al. (2003).

Exclusions

In total, seventeen labs contributed data from a total of 1267 participants; 162 were excluded as per our pre-registered criteria leaving a total of 1105. See Table 1 for details of the number of participants collected by each lab, the number analysed, and the number excluded based on each criterion; the technical error category includes those participants that were excluded for having incomplete data due to, for example, equipment failure, experimenter error, or other technical errors.

Five labs used an eye-tracker for at least some of their participants. See Table A11 for details of the number of participants tested with an eye-tracker, number of participants analysed in our secondary analysis of eye movement contaminated trials, and number of eye movement contaminated trials at each ISI × congruency condition for each lab.

Preliminary analyses

Across all 1105 participants and four ISI conditions, the congruency effect we observed had a mean of 0.24 ms and a standard deviation of 12.48 ms. In addition, across all 1105 participants, it had a mean of -0.07 ms and a standard deviation of 13.45 ms at the 250 ms ISI condition, a mean of 0.94 ms and a standard deviation of 12.42 ms at the 500 ms ISI condition, a mean of -0.02 ms and a standard deviation of 12.12 ms at the 750 ms ISI condition, and a mean of 0.10 ms and a standard deviation of

Table 1

Total number of participants, number analysed, number excluded for reasons of technical error, number excluded for more than 5% catch trial errors, and number excluded for guessing the purpose of the experiment for each lab.

Lab	Total Participants	Analysed Participants	Technical Error	Catch Trial Error	Guessed Purpose
Ansari	68	60	2	6	0
Bryce	68	61	0	3	4
Chen	62	60	1	1	0
Cipora	93	82	1	3	7
Colling (Szűcs)	72	65	4	3	0
Corballis	68	64	2	2	0
Hancock	66	54	5	6	1
Holmes	77	60	3	8	6
Lindemann	50	47	0	1	2
Lukavský	62	61	1	0	0
Mammarella	126	103	15	1	7
Mieth	124	93	2	8	21
Moeller	77	63	13	1	0
Ocampo	60	59	0	0	1
Ortiz-Ouellet-Lupiáñez-Santiago	60	54	3	2	1
Toomarian	74	61	4	7	2
Treccani	60	58	0	1	1

11.84 ms at the 1000 ms ISI condition. Further, the correlation between conditions had a mean of 0.00 (and a mean of 0.03 in magnitude) across the six possible pairs of conditions.

In terms of sign, across all 1105 participants and four ISI conditions, the proportion of times the congruency effect we observed was positive was 0.50. In addition, across all 1105 participants, this proportion was 0.49 at the 250 ms ISI condition, 0.53 at the 500 ms ISI condition, 0.48 at the 750 ms ISI condition, and 0.50 at the 1000 ms ISI condition. Further, the proportion of times the number of positive congruency effects per participant was equal to zero, one, two, three, and four was respectively 0.06, 0.26, 0.36, 0.26, and 0.06. All of these results are compatible with the relevant binomial distribution with probability parameter one-half (i.e., the distribution of the number of heads on tosses of a fair coin).

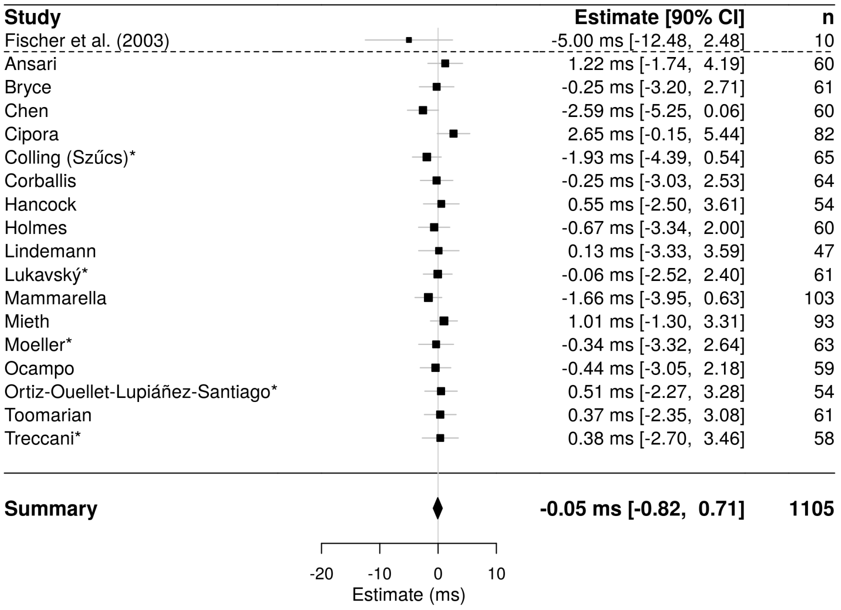
Primary analyses

Model 1: No moderators. The purpose of Model 1 was to replicate the analysis performed by Fischer et al. (2003), and thus it did not account for any moderators. Model 1 was fit to data from 1105 participants from seventeen labs. We summarise the results from Study 2 of Fischer et al. (2003) along with results from each lab and from Model 1 in Figure 2.

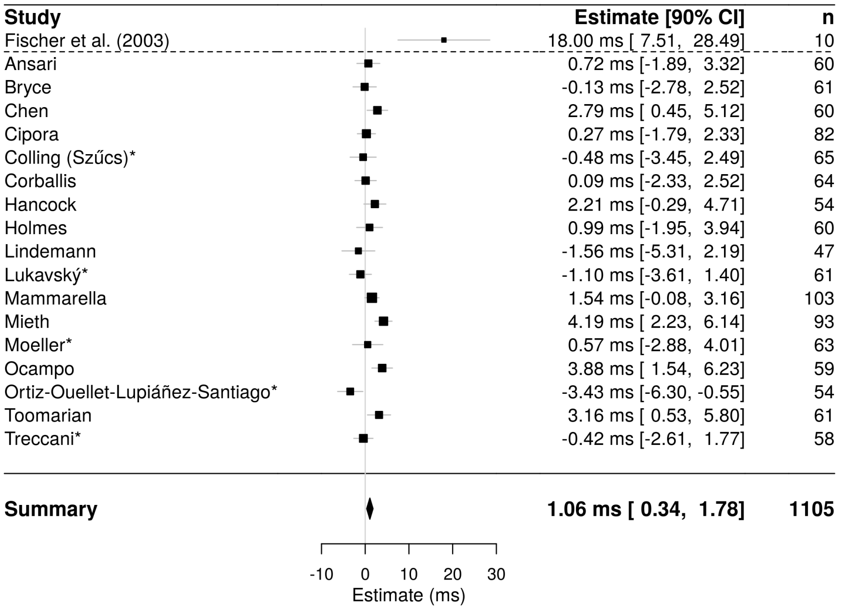
The effects we observed both within and across labs were minuscule and incompatible with those observed in Fischer et al. (2003). Specifically, Fischer et al. (2003) estimated an effect of -5.00 ms at the 250 ms ISI condition, 18.00 ms at the 500 ms ISI condition, 23.00 ms at the 750 ms ISI condition, and 11.00 ms at the 1000 ms ISI condition. In contrast, Model 1 estimates an effect of -0.05 ms, 1.06 ms, 0.19 ms, and 0.18 ms at each of the four respective ISI conditions.

Given this in tandem with the results of our preliminary analyses, we conclude that we have *failed* to replicate the effect reported by Fischer et al. (2003).

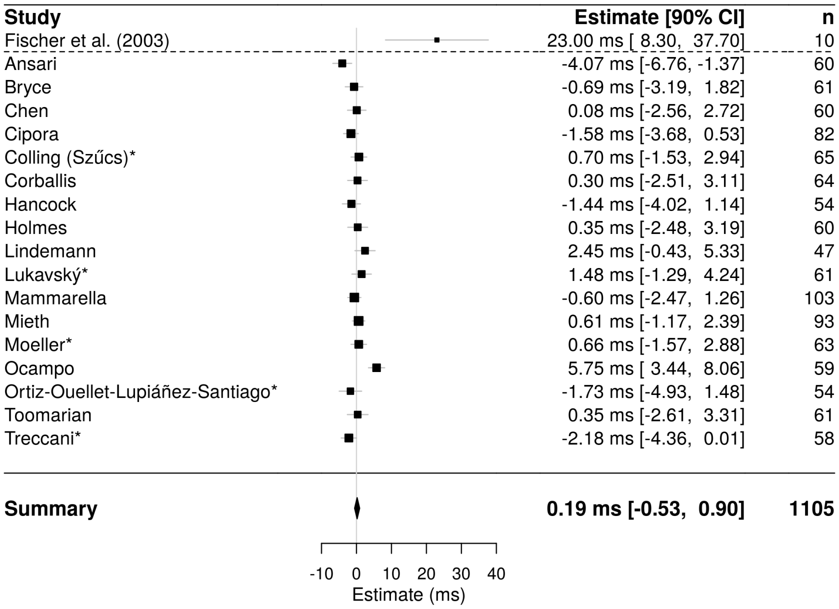
Another major finding was that the effects we observed were highly consistent not only across ISI conditions but also—perhaps more surprisingly—across labs. Recent work has found that, contrary to both substantive and statistical expectations, large-scale replications projects like the present one tend to show a nontrivial degree of heterogeneity across labs (McShane, Tackett, Böckenholt, & Gelman, 2019). In contrast, we estimate heterogeneity across labs at 1.02 ms and thus practically unimportant for most purposes. This suggests that, at least across the labs involved in the present project, there are unlikely to be lab-level moderators driving our results. See Table 1 and Supplementary Table A1 for additional details.



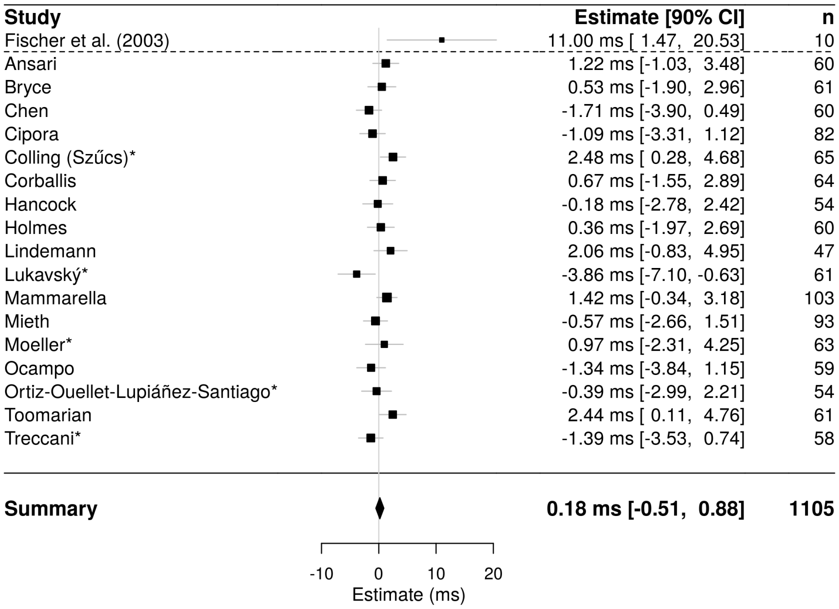
(a) 250 ms ISI Condition



(b) 500 ms ISI Condition



(c) 750 ms ISI Condition



(d) 1000 ms ISI Condition

Figure 2. Summary of Results from Study 2 of Fischer et al. (2003), Each Lab, and Model 1. The effects we observed both within and across labs were miniscule—around 1 ms—and incompatible with those of around 20 ms observed in Fischer et al. (2003). They were also highly consistent not only across ISI conditions but also—perhaps more surprisingly—across labs with the latter suggesting there are unlikely to be lab-level moderators driving our results Labs using an eye-tracker are marked with an asterisk.

Model 2: Finger counting. Model 2 was fit to data from 343 left-starter participants from seventeen labs and 482 right-starter participants from seventeen labs. We summarize the results from Model 2 along with the results from Study 2 of Fischer et al. (2003) as well as Model 1, Model 3, and Model 4 in Figure 3. While the evidence presented above suggests a stronger congruency effect among left-starters and a weaker or possibly even reversed effect among right-starters, as can be seen in Figure 3, finger counting had no substantial impact on the results: we observed minuscule effects for each ISI condition and finger counting group and minuscule differences between the two finger counting groups at each ISI condition. See Supplementary Table A2 and Supplementary Table A3 for additional details.

Model 3: Reading/writing direction. Model 3 was fit to data from 1014 exclusively left-to-right readers/writers from seventeen labs and 76 not exclusively left-to-right readers/writers from eight labs. While the evidence presented above suggests a weaker or possibly even reversed congruency effect among those who have experience with languages that are not read/written exclusively from left to right, as can be seen in Figure 3, reading/writing direction had no substantial impact on the results: we observed minuscule effects for each ISI condition and reading/writing direction group and minuscule differences between the two reading/writing direction groups at each ISI condition. See Supplementary Table A4 and Supplementary Table A5 for additional details.

Model 4: Handedness. Model 4 was fit to data from 69 left-handed participants from nine labs and 1007 right-handed participants seventeen labs. As can be seen in Figure 3, handedness had no substantial impact on the results: we observed minuscule effects for each ISI condition and handedness group and minuscule differences between the two handedness groups at each ISI condition. See Supplementary Table A6 and Supplementary Table A7 for additional details.

Model 5: Mathematics fluency and mathematics anxiety. Model 5 was fit to data from 1105 participants from seventeen labs. While the evidence presented above suggests mathematics fluency and mathematics anxiety might moderate congruency effects, we observed no substantial moderating effects. See Table 1 and Supplementary Table A8 for additional details.

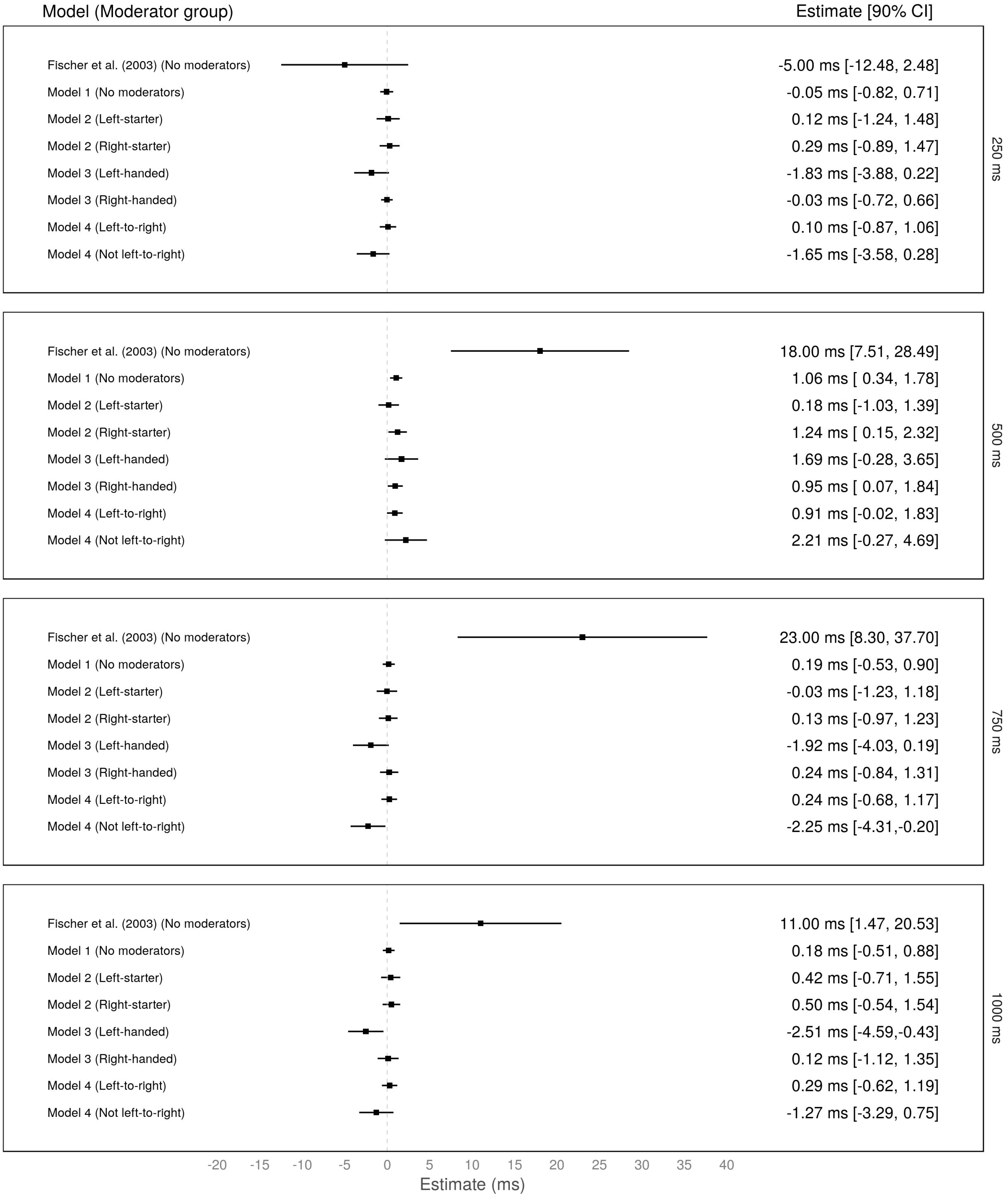


Figure 3. Summary of Results from Study 2 of Fischer et al. (2003) and Models 1–4. The effects we observed were minuscule and incompatible with those observed in Fischer et al. (2003). They were also highly consistent across ISI conditions.

Secondary analyses

Model 1 was refit separately to data from 41 participants from four labs who correctly guessed the purpose of the experiment and to data from 10468 eye movement contaminated trials from 132 participants from five labs with contaminated trials at each ISI × congruency condition. These analyses yielded nothing of substantive interest. See Supplementary Materials for details.

Discussion

The att-SNARC effect (Fischer et al., 2003) has been used to argue for an early, response-independent, and automatic origin of the SNARC effect. If the SNARC effect is produced by early mechanisms, this would provide good evidence for “embodied” number representations and allow for strong claims about the link between number and space (e.g., a mental number line).

We attempted to replicate Study 2 of Fischer et al. (2003) by collecting data from 1105 participants across seventeen labs. Across all 1105 participants and four ISI conditions, the proportion of times the congruency effect we observed was positive was 0.50. Further, the effects we observed both within and across labs were miniscule and incompatible with those observed in Fischer et al. (2003). Given this, we conclude that we have *failed* to replicate the effect reported by Fischer et al. (2003).

The effects we observed were also highly consistent not only across ISI conditions but also—perhaps more surprisingly—across labs. The latter suggests there are unlikely to be lab-level moderators driving our results. In addition, our analysis of several participant-level moderators (finger counting preferences, reading/writing direction experience, handedness, and mathematics fluency and mathematics anxiety) revealed no substantial moderating effects.

We conclude with two important points. First, one might, on the basis of the common definition of replication employed in practice, object that we have successfully replicated Fischer et al. (2003), at least at the 500 ms ISI condition. In response, we argue this illustrates one major flaw of that definition: our result at the 500 ms ISI condition is manifestly incompatible with the analogous result of Fischer et al. (2003). In addition, we view a difference of about 1 ms, even if “real”, as too small for any neurally or psychologically plausible mechanism—particularly one constrained to operate only within a narrow time window of 500 ms after the digit display stimulus. That said, we recognize that some such mechanism could be subject to an arbitrarily large attenuation factor in any particular experimental paradigm such as that of Fischer et al. (2003), and that potential new paradigms could reveal an effect. Nonetheless, even were such paradigms forthcoming, we maintain based on these results that Fischer et al. (2003) provides no evidence of such a mechanism.

Second, we point to several limitations of the present study. First and foremost, while our results demonstrate that the att-SNARC effect cannot be used as evidence to support the strong claims about the link between number and space discussed above, this does not refute such accounts. Specifically, while one might, on the basis of our results, prefer accounts of the SNARC effect that do not imply a mental number line, the entirety of the evidence for and against different claims about the SNARC effect must be viewed as a whole. The att-SNARC effect provides only one such piece of evidence—albeit a particularly strong and valuable one.

The second set of limitations relates to our sample of subjects. Our sample was primarily collected from North America, Europe, and Australasia. Consequently, participants who read/wrote exclusively left to right are overrepresented in our data. As reading/writing direction has been shown to strongly moderate spatial-numerical associations, it would have been preferable to have more participants in this subgroup. In addition, data sparsity prevented us from considering all the moderators jointly in a single model and thus we were required to consider each moderator separately.

Finally, the finger counting assessment we employed did not contain an explicit instruction to engage in finger counting. As a result, some participants inconsistently employed finger counting, resulting in them being excluded from the Model 2 analysis.

Acknowledgements

LJC and DS are funded by James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition (grant number 220020370; received by DS). We acknowledge the help of the original authors, in particular Martin Fischer and Jay Pratt. We also note this project would not have been possible without editor Alex Holcombe's patient and thoughtful help at every step of the process.

Author contributions

LJC and DS proposed the study. LJC programmed the experiments. LJC and BBM conducted the analyses. LJC wrote an initial manuscript. LJC and BBM wrote revised and final manuscripts. All authors critically reviewed the final manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Allaire, J., Ushey, K., & Tang, Y. (2018). *Reticulate: Interface to 'python'*. R package version 1.10. Retrieved from https://CRAN.R-project.org/package=reticulate

Amrhein, V., Greenland, S., & McShane, B. B. (2019). Retire statistical significance. *Nature*, *7748*(567), 305–307. doi:10.1038/d41586-019-00857-9

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistican*, *73*(sup1), 262–270. doi:10.1080/00031305.2018.1543137

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. R package version 0.1.0.9842. Retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r*. R package version 1.5. Retrieved from https://CRAN.R-project.org/package=magrittr

Bächtold, D., Baumüller, M., & Brugger, P. (1998). Stimulus-response compatibility in representational space. *Neuropsychologia*, *36*(8), 731–735. doi:10.1016/S0028-3932(98)00002-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bates, D., & Maechler, M. (2018). *Matrix: Sparse and dense matrix classes and methods*. R package version 1.2-14. Retrieved from https://CRAN.R-project.org/package=Matrix

Bengtsson, H. (2018). *R.matlab: Read and write mat files and call matlab from within r*. R package version 3.6.2. Retrieved from https://CRAN.R-project.org/package=R.matlab

Champely, S. (2018). *Pwr: Basic functions for power analysis*. R package version 1.2-2. Retrieved from https://CRAN.R-project.org/package=pwr

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997

Colling, L. J., & Szűcs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*, 1–27. doi:10.1007/s13164-018-0421-4

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. doi:10.1037//0096-3445.122.3.371

Dodd, M. D., Van der Stigchel, S., Leghari, M. A., Fung, G., & Kingstone, A. (2008). Attentional SNARC: There’s something special about numbers (let us count the ways). *Cognition*, *108*(3), 810–818. doi:10.1016/j.cognition.2008.04.006

Fattorini, E., Pinto, M., Rotondaro, F., & Doricchi, F. (2015). Perceiving numbers does not cause automatic shifts of spatial attention. *Cortex*, *73*, 298–316. doi:10.1016/j.cortex.2015.09.007

Fias, W., Brysbaert, M., Geypens, F., & d’Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition*, *2*(1), 95–110. doi:10.1080/135467996387552

Fischer, M. H. (2006). The Future for Snarc Could Be Stark… *Cortex*, *42*(8), 1066–1068. doi:10.1016/S0010-9452(08)70218-1

Fischer, M. H. (2008). Finger counting habits modulate spatial-numerical associations. *Cortex*, *44*(4), 386–392. doi:10.1016/j.cortex.2007.08.004

Fischer, M. H., & Brugger, P. (2011). When digits help digits: Spatial-numerical associations point to finger counting as prime example of embodied cognition. *Frontiers in Psychology*, *2*, 260. doi:10.3389/fpsyg.2011.00260

Fischer, M. H., Castel, A. D., Dodd, M. D., & Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. *Nature Neuroscience*, *6*(6), 555–556. doi:10.1038/nn1066

Fischer, M. H., & Knops, A. (2014). Attentional cueing in numerical cognition. *Frontiers in Psychology*, *5*(325), 426. doi:10.3389/fpsyg.2014.01381

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard Unvieristy Press.

Galfano, G., Rusconi, E., & Umiltà, C. (2006). Number magnitude orients attention, but not against one’s will. *Psychonomic Bulletin & Review*, *13*(5), 869–874. doi:10.3758/BF03194011

Galton, F. (1880). Visualised numerals. *Nature*, *21*, 252–256. doi:10.1038/021252a0

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis*. Chapman and Hall/CRC: Boca Raton, FL.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives in Psychological Science*, *9*(6), 641–651. doi:10.1177/1745691614551642

Georges, C., Hoffmann, D., & Schiltz, C. (2016). How Math Anxiety Relates to Number–Space Associations. *Frontiers in Psychology*, *7*(33), 143. doi:10.3389/fpsyg.2016.01401

Gevers, W., Reynvoet, B., & Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. *Cognition*, *87*(3), B87–B95. doi:10.1016/S0010-0277(02)00234-2

Gevers, W., Reynvoet, B., & Fias, W. (2004). The Mental Representation of Ordinal Sequences is Spatially Organised: Evidence from Days of the Week. *Cortex*, *40*(1), 171–172. doi:10.1016/S0010-9452(08)70938-9

Gevers, W., Verguts, T., Reynvoet, B., Vaessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 32–44. doi:10.1037/0096-1523.32.1.32

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & philosophy*, *32*(3), 337–355. doi:10.1007/s10539-017-9562-6

Gordon, M., & Lumley, T. (2017). *Forestplot: Advanced forest plot using ’grid’ graphics*. R package version 1.7.2. Retrieved from https://CRAN.R-project.org/package=forestplot

Henry, L., & Wickham, H. (2018). *Purrr: Functional programming tools*. R package version 0.2.5. Retrieved from https://CRAN.R-project.org/package=purrr

Hester, J. (2018). *Glue: Interpreted string literals*. R package version 1.3.0. Retrieved from https://CRAN.R-project.org/package=glue

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability. *Assessment*, *10*(2), 178–182. doi:10.1177/1073191103252351

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. doi:10.1097/EDE.0b013e31818131e7

Keus, I. M., Jenks, K. M., & Schwarz, W. (2005). Psychophysiological evidence that the SNARC effect has its functional locus in a response selection stage. *Cognitive Brain Research*, *24*(1), 48–56. doi:10.1016/j.cogbrainres.2004.12.005

Keus, I. M., & Schwarz, W. (2005). Searching for the functional locus of the SNARC effect: Evidence for a response-related origin. *Memory & Cognition*, *33*(4), 681–695. doi:10.3758/BF03195335

Landy, D. H., Jones, E. I., & Hummel, J. E. (2008). Why spatial-numerical associations aren't evidence for a mental number line. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 257–362). Austin, TX.

Lang, M. (2017). checkmate: Fast argument checks for defensive r programming. *The R Journal*, *9*(1), 437–445. doi:10.32614/RJ-2017-028

Lucidi, A., & Thevenot, C. (2014). Do not count on me to imagine how I act: behavior contradicts questionnaire responses in the assessment of finger counting habits. *Behavior Research Methods*, *46*(4), 1079–1087. doi:10.3758/s13428-014-0447-1

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1), 255–271. doi:10.1007/s11336-017-9571-z

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. doi:10.1177/1745691616662243

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*(6), 1707–1718. doi:10.1287/mnsc.2015.2212

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*(519), 885–895. doi:10.1080/01621459.2017.1289846

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician*, *73*(sup1), 235–245. doi:10.1080/00031305.2018.1527253

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large scale replication projects in contemporary psychological research. *The American Statistican*, *73*(sup1), 99–105. doi:10.1080/00031305.2018.1505655

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. doi:10.1037/0022-006X.46.4.806

Müller, K. (2018). *Bindrcpp: An ’rcpp’ interface to active bindings*. R package version 0.2.2. Retrieved from https://CRAN.R-project.org/package=bindrcpp

Müller, K., & Wickham, H. (2018). *Tibble: Simple data frames*. R package version 1.4.2. Retrieved from https://CRAN.R-project.org/package=tibble

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi:10.1145/360018.360022

Nicholls, M. E. R., Thomas, N. A., Loetscher, T., & Grimshaw, G. M. (2013). The Flinders Handedness survey (FLANDERS): A brief measure of skilled hand preference. *Cortex*, *49*(10), 2914–2926. doi:10.1016/j.cortex.2013.02.002

Ooms, J. (2018). *Magick: Advanced graphics and image-processing in r*. R package version 1.9. Retrieved from https://CRAN.R-project.org/package=magick

Pecher, D., & Boot, I. (2011). Numbers in space: Differences between concrete and abstract situtations. *Frontiers in Psychology*, *2*(121). doi:10.3389/fpsyg.2011.00121

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-137. Retrieved from https://CRAN.R-project.org/package=nlme

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. doi:10.1080/00335558008248231

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. doi:10.1037/0033-2909.132.3.416

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Ranzini, M., Dehaene, S., Piazza, M., & Hubbard, E. M. (2009). Neural mechanisms of attentional shifts due to irrelevant spatial and numerical cues. *Neuropsychologia*, *47*(12), 2615–2624. doi:10.1016/j.neuropsychologia.2009.05.011

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*(2, Pt.1), 274–278. doi:10.1037/h0028573

Ristic, J., Wright, A., & Kingstone, A. (2006). The number line effect reflects top-down control. *Psychonomic Bulletin & Review*, *13*(5), 862–868. doi:10.3758/BF03194010

Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. doi:10.1037/h0042040

Salillas, E., El Yagoubi, R., & Semenza, C. (2008). Sensory and cognitive processes of shifts of spatial attention induced by numbers: An erp study. *Cortex*, *44*(4), 406–413. doi:10.1016/j.cortex.2007.08.006

Santens, S., & Gevers, W. (2008). The SNARC effect does not imply a mental number line. *Cognition*, *108*(1), 263–270. doi:10.1016/j.cognition.2008.01.002

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*(2), 328–331. doi:10.3758/PBR.16.2.328

Tibber, M. S., Manasseh, G. S. L., Clarke, R. C., Gagin, G., Swanbeck, S. N., Butterworth, B., … Dakin, S. C. (2013). Sensitivity to numerosity is not a unique visuospatial psychophysical predictor of mathematical ability. *Vision Research*, *89*, 1–9. doi:10.1016/j.visres.2013.06.006

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of spatial-numerical associations: Evolutionary factors co-construct the mental number line. *Neuroscience and Biobehavioural Reviews*, *90*, 184–199. doi:10.1016/j.neubiorev.2018.04.010

van Dijck, J.-P., Abrahamse, E. L., Acar, F., Ketels, B., & Fias, W. (2014). A working memory account of the interaction between number and spatial attention. *Quarterly Journal of Experimental Psychology*, *67*(8), 1500–1513. doi:10.1080/17470218.2014.903984

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition*, *119*(1), 114–119. doi:10.1016/j.cognition.2010.12.013

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond p < 0.05. *The American Statistican*, *73*(sup1), 1–19. doi:10.1080/00031305.2019.1583913

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H. (2018a). *Forcats: Tools for working with categorical variables (factors)*. R package version 0.3.0. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2018b). *Stringr: Simple, consistent wrappers for common string operations*. R package version 1.3.1. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2018). *Dplyr: A grammar of data manipulation*. R package version 0.7.6. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2018). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. R package version 0.8.1. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2017). *Readr: Read rectangular text data*. R package version 1.1.1. Retrieved from https://CRAN.R-project.org/package=readr

Williams, D., & Colling, L. J. (2018). From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese*, *195*(5), 1941–1967. doi:10.1007/s11229-017-1578-6

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636. doi:10.3758/BF03196322

Wood, G., Nuerk, H.-C., & Willmes, K. (2006). Crossed Hands and the Snarc Effect: A failure to Replicate Dehaene, Bossini and Giraux (1993). *Cortex*, *8*, 1069–1079. doi:10.1016/S0010-9452(08)70219-3

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science*, *50*(4), 489–525.

Woodcock, R., & Johnson, M. (1989). *Woodock Johnson—Revised-tests of academic achievement*. Chicago: The Riverside Publishing Company.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd). Boca Raton, Florida: Chapman and Hall/CRC.

Zanolie, K., & Pecher, D. (2014). Number-induced shifts in spatial attention: a replication study. *Frontiers in Psychology*, *5*(e85048), 667. doi:10.3389/fpsyg.2014.00987

Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC Effect and the REVERSE SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture*, *5*(1–2), 165–190. doi:10.1163/1568537054068660

Zhu, H. (2018). *Kableextra: Construct complex table with ’kable’ and pipe syntax*. R package version 0.9.0. Retrieved from https://CRAN.R-project.org/package=kableExtra

Supplementary Results

Primary analyses

Model 1: No Moderators. Model 1 was fit to data from 1105 participants from seventeen labs (see Table 1 for details). Of the six equal allocation multilevel multivariate compound symmetry (EAMMCS) model specifications, the *Equal Variance, Zero Correlation* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A1.

Model 2: Finger counting. Model 2 was fit to data from 343 left-starter participants from seventeen labs and 482 right-starter participants from seventeen labs (see Supplementary Table A2 for details). Of the six EAMMCS model specifications, the *Equal Variance, Zero Correlation* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A3.

Model 3: Reading/writing direction. Model 3 was fit to data from 1014 exclusively left-to-right readers/writers from seventeen labs and 76 not exclusively left-to-right readers/writers from eight labs (see Supplementary Table A4 for details). Of the six EAMMCS model specifications, the *Equal Variance, Zero Correlation* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A5.

Model 4: Handedness. Model 4 was fit to data from 69 left-handed participants from nine labs and 1007 right-handed participants from seventeen labs (see Supplementary Table A6 for details). Of the six EAMMCS model specifications, the *Unequal Variance, Zero Correlation* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A7.

Model 5: Mathematics fluency and mathematics anxiety. Model 5 was fit to data from 1105 participants from seventeen labs (see Table 1). See the main text for model specification details, but note that (i) for consistency with Model 1 we employed the *Equal Variance, Zero Correlation* specification for the Lab × ISI Condition effects and (ii) the maths test and AMAS were centred and scaled by their respective means and standard deviations across the 1105 participants prior to estimation of the model. Fixed effect estimates, standard errors, and t-statistics and variance component estimates are shown in Supplementary Table A8.

Secondary analyses

Purpose of experiment. Data from several participants were not included in the primary analysis because they correctly guessed the purpose of the experiment (as assessed by the exit questionnaire). The data from these participants was analysed separately to determine whether insight into the purpose of the experiment moderated the effect. Specifically, Model 1 was refit to data from the 41 participants from four labs who correctly guessed the purpose of the experiment (see Supplementary Table A9 for details). Of the six model EAMMCS model specifications, the *Equal Variance, Zero Correlation* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A10.

Eye-movement contaminated trials. Data from individual trials that were contaminated with eye movements were also not included the primary analysis. The data from these trials was analysed separately to determine whether eye movements moderated the effect. Specifically, Model 1 was refit to data from 10468 eye movement contaminated trials from 132 participants from five labs with contaminated trials at each ISI × congruency condition (see Supplementary Table A11 for details). Of the six EAMMCS model specifications, the *Fixed Effects* specification was chosen by AIC. AIC; fixed effect estimates, standard errors, and z-statistics; and variance component estimates are shown in Supplementary Table A12

Table A1

Model 1 Estimates.

(a) *AIC*

Specification	AIC
Fixed Effects	264.12
Equal Variance, Zero Correlation	259.66
Equal Variance, Single Correlation	261.64
Unequal Variance, Zero Correlation	261.04
Unequal Variance, Single Correlation	260.87
No Constraints	270.83

(b) *Fixed Effect Estimates*

ISI Condition	Estimate	Std. Err.	z
250 ms	-0.05	0.47	-0.11
500 ms	1.06	0.44	2.43
750 ms	0.19	0.43	0.43
1000 ms	0.18	0.42	0.44

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale.*

ISI Condition	Estimate
250 ms	1.02
500 ms	1.02
750 ms	1.02
1000 ms	1.02

Table A2

Number of participants in each finger counting group for each of the seventeen labs.

Lab	Left-Starter	Left-Prefer	No Group	Right-Prefer	Right-Starter
Ansari	23	2	2	3	30
Bryce	13	8	2	17	21
Chen	22	0	2	0	36
Cipora	19	9	5	18	31
Colling (Szűcs)	21	3	11	3	27
Corballis	18	3	5	4	34
Hancock	22	6	0	3	23
Holmes	14	2	1	8	35
Lindemann	22	1	4	1	19
Lukavský	12	7	2	16	24
Mammarella	30	8	6	23	36
Mieth	32	10	10	16	25
Moeller	23	0	6	0	34
Ocampo	27	0	2	0	30
Ortiz-Ouellet-Lupiáñez-Santiago	10	8	4	22	10
Toomarian	19	0	0	0	42
Treccani	16	7	4	6	25

Table A3

Model 2 Estimates.

(a) *AIC*

Specification	AIC
Fixed Effects	665.97
Equal Variance, Zero Correlation	637.31
Equal Variance, Single Correlation	639.00
Unequal Variance, Zero Correlation	638.57
Unequal Variance, Single Correlation	640.13
No Constraints	646.51

(b) *Fixed Effect Estimates*

ISI Condition	Finger counting group	Estimate	Std. Err.	z
250 ms	Right-starter	0.29	0.72	0.40
250 ms	Left-starter	0.12	0.83	0.14
500 ms	Right-starter	1.24	0.66	1.88
500 ms	Left-starter	0.18	0.74	0.24
750 ms	Right-starter	0.13	0.67	0.19
750 ms	Left-starter	-0.03	0.73	-0.04
1000 ms	Right-starter	0.50	0.63	0.79
1000 ms	Left-starter	0.42	0.69	0.61

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale. 39% of the variance is estimated to be at the lab-level and 61% at the group-level.*

ISI Condition	Estimate
250 ms	1.74
500 ms	1.74
750 ms	1.74
1000 ms	1.74

Table A4

Number of participants in each of the reading/writing direction groups for each of the seventeen labs.

Lab	Exclusively Left-to-Right	Not exclusively Left-to-Right
Ansari	55	5
Bryce	59	2
Chen	39	21
Cipora	76	6
Colling (Szűcs)	55	10
Corballis	60	4
Hancock	53	1
Holmes	54	6
Lindemann	47	0
Lukavský	58	3
Mammarella	103	0
Mieth	79	14
Moeller	54	9
Ocampo	55	4
Ortiz-Ouellet-Lupiáñez-Santiago	54	0
Toomarian	56	5
Treccani	57	1

Table A5

Model 3 Estimates.

(a) *AIC*

Specification	AIC
Fixed Effects	495.58
Equal Variance, Zero Correlation	448.05
Equal Variance, Single Correlation	449.41
Unequal Variance, Zero Correlation	451.89
Unequal Variance, Single Correlation	453.44
No Constraints	457.83

(b) *Fixed Effect Estimates*

ISI Condition	Reading/Writing Direction	Estimate	Std. Err.	z
250 ms	Exclusively LTR	0.10	0.59	0.17
250 ms	Not exclusively LTR	-1.65	1.17	-1.41
500 ms	Exclusively LTR	0.91	0.56	1.62
500 ms	Not exclusively LTR	2.21	1.51	1.46
750 ms	Exclusively LTR	0.24	0.56	0.43
750 ms	Not exclusively LTR	-2.25	1.25	-1.80
1000 ms	Exclusively LTR	0.29	0.55	0.53
1000 ms	Not exclusively LTR	-1.27	1.23	-1.03

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale. 10% of the variance is estimated to be at the lab-level and 90% at the group-level.*

ISI Condition	Estimate
250 ms	1.71
500 ms	1.71
750 ms	1.71
1000 ms	1.71

Table A6

Number of participants in each handedness group for each of the seventeen labs.

Lab	Left-handed	Right-handed
Ansari	4	56
Bryce	4	57
Chen	5	55
Cipora	3	79
Colling (Szűcs)	7	58
Corballis	9	55
Hancock	6	48
Holmes	4	56
Lindemann	5	42
Lukavský	7	54
Mammarella	6	97
Mieth	14	79
Moeller	4	59
Ocampo	4	55
Ortiz-Ouellet-Lupiáñez-Santiago	3	51
Toomarian	10	51
Treccani	3	55

Table A7

Model 4 Estimates.

(a) *AIC*

Specification	AIC
Fixed Effects	598.41
Equal Variance, Zero Correlation	473.56
Equal Variance, Single Correlation	475.56
Unequal Variance, Zero Correlation	470.86
Unequal Variance, Single Correlation	472.48
No Constraints	480.12

(b) *Fixed Effect Estimates*

ISI Condition	Handedness Group	Estimate	Std. Err.	z
250 ms	Right-handed	-0.03	0.42	-0.07
250 ms	Left-handed	-1.83	1.25	-1.46
500 ms	Right-handed	0.95	0.54	1.76
500 ms	Left-handed	1.69	1.19	1.42
750 ms	Right-handed	0.24	0.65	0.37
750 ms	Left-handed	-1.92	1.28	-1.50
1000 ms	Right-handed	0.12	0.75	0.16
1000 ms	Left-handed	-2.51	1.27	-1.98

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale. 12% of the variance is estimated to be at the lab-level and 88% at the group-level.*

ISI Condition	Estimate
250 ms	0.01
500 ms	1.57
750 ms	2.19
1000 ms	2.71

Table A8

Model 5 Estimates.

(a) *Fixed Effect Estimates*

Effect	Estimate	Std. Err.	t
250 ms ISI	-0.03	0.44	-0.07
500 ms ISI	0.88	0.44	2.02
750 ms ISI	0.01	0.44	0.02
1000 ms ISI	0.21	0.44	0.48
250 ms ISI × Maths test	-0.15	0.42	-0.35
500 ms ISI × Maths test	-0.80	0.42	-1.90
750 ms ISI × Maths test	-0.24	0.42	-0.57
1000 ms ISI × Maths test	0.08	0.42	0.18
250 ms ISI × AMAS	-0.66	0.40	-1.66
500 ms ISI × AMAS	0.29	0.40	0.73
750 ms ISI × AMAS	-0.21	0.40	-0.54
1000 ms ISI × AMAS	-0.57	0.40	-1.44
250 ms ISI × Maths test × AMAS	-0.12	0.39	-0.30
500 ms ISI × Maths test × AMAS	-0.38	0.39	-0.98
750 ms ISI × Maths test × AMAS	-0.24	0.39	-0.63
1000 ms ISI × Maths test × AMAS	0.22	0.39	0.56

(b) *Variance Component Estimates. Estimates are presented on the standard deviation scale.*

ISI Condition	Estimate	Additional Effects	Estimate
250 ms	0.85	Participant	0.00
500 ms	0.85	Maths Test	0.61
750 ms	0.85	AMAS	0.33
1000 ms	0.85	Maths test × AMAS	0.50

Table A9

Number of participants who correctly guessed the purpose of the experiment for each lab.

Lab	n
Cipora	7
Holmes	6
Mammarella	7
Mieth	21

Table A10

Model 1 Estimates (only participants who correctly guessed the purpose of the experiment).

(a) *AIC*

Specification	AIC
Fixed Effects	80.21
Equal Variance, Zero Correlation	71.39
Equal Variance, Single Correlation	73.39
Unequal Variance, Zero Correlation	73.83
Unequal Variance, Single Correlation	75.83
No Constraints	85.42

(b) *Fixed Effect Estimates*

ISI Condition	Estimate	Std. Err.	z
250 ms	1.49	2.21	0.67
500 ms	0.36	2.32	0.16
750 ms	-0.68	2.17	-0.31
1000 ms	1.15	2.37	0.48

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale.*

ISI Condition	Estimate
250 ms	3.08
500 ms	3.08
750 ms	3.08
1000 ms	3.08

Table A11

Number of participants tested with an eye-tracker, number of participants analysed in our secondary analysis of eye movement contaminated trials, and number of eye movement contaminated trials in the analysis (total number of eye movement contaminated trials) at each ISI × congruency condition for each lab.

Lab	Participants	Analysed	Trial Type	250 ms	500 ms	750 ms	1000 ms
Colling (Szűcs)	52	18	Congruent	64 (88)	93 (133)	109 (173)	107 (162)
			Incongruent	71 (97)	95 (144)	103 (140)	95 (142)
Lukavský	61	29	Congruent	158 (182)	201 (240)	235 (278)	252 (292)
			Incongruent	146 (176)	202 (238)	231 (280)	233 (282)
Moeller	64	53	Congruent	593 (600)	723 (734)	774 (787)	851 (868)
			Incongruent	621 (635)	711 (729)	774 (802)	842 (858)
Ortiz-Ouellet-Lupiáñez-Santiago	28	18	Congruent	127 (135)	165 (177)	176 (186)	184 (197)
			Incongruent	130 (138)	147 (157)	167 (174)	160 (175)
Treccani	30	14	Congruent	89 (99)	113 (136)	129 (139)	133 (152)
			Incongruent	99 (109)	116 (126)	124 (144)	125 (141)

Table A12

Model 1 Estimates (only eye movement contaminated trials).

(a) *AIC*

Specification	AIC
Fixed Effects	120.28
Equal Variance, Zero Correlation	122.28
Equal Variance, Single Correlation	124.28
Unequal Variance, Zero Correlation	127.98
Unequal Variance, Single Correlation	129.75
No Constraints	139.65

(b) *Fixed Effect Estimates*

ISI Condition	Estimate	Std. Err.	z
250 ms	-5.35	6.27	-0.85
500 ms	-2.65	4.95	-0.54
750 ms	-5.52	3.98	-1.39
1000 ms	3.86	4.17	0.93

(c) *Variance Component Estimates. Estimates are presented on the standard deviation scale.*

ISI Condition	Estimate
250 ms	0
500 ms	0
750 ms	0
1000 ms	0