AUTHOR ACCEPTED VERSION

The final version is available at **Advances in Methods and Practices in Psychological Science**

https://doi.org/10.1177/2515245920903079

Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003)

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

The final version of this paper is Advances in Methods and Practices in Psychological Science

https://doi.org/10.1177/2515245920903079

Author Note

Correspondence concerning this article should be addressed to **Lincoln J Colling** E-mail: lincoln@colling.net.nz

Abstract

The attentional spatial-numerical association of response codes (Att-SNARC) effect (Fischer, Castel, Dodd, & Pratt, 2003)—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 support strong claims about the link between number and space (e.g., a mental number line). We attempted to replicate Experiment 2 of Fischer et al. by collecting data from 1105 participants at 17 labs. Across all 1105 participants and four interstimulus-interval conditions, the proportion of times the effect we observed was positive (i.e., directionally consistent with the original effect) was .50. Further, the effects we observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. Given this, we conclude that we failed to replicate the effect reported by Fischer et al. In addition, our analysis of several participant-level moderators (finger-counting habits, reading and writing direction, handedness, and mathematics fluency and mathematics anxiety) revealed no substantial moderating effects. Our results indicate that the Att-SNARC effect cannot be used as evidence to support strong claims about the link between number and space.

Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003)

Introduction

A foundational issue in cognitive science is the question of how people represent concepts. Classical approaches to cognitive science, exemplified by Fodor's (1975) *language-of-thought hypothesis* 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 maintain that representations (a) are intimately linked to sensorimotor processing (see, e.g., Wilson, 2002, for an overview), (b) are analog rather than symbolic, and (c) represent by resembling their targets in some sense (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 this issue has been numerical cognition. In fact, Fischer and Brugger (2011) referred to numerical cognition as the "prime example of embodied cognition." In particular, they pointed to tasks examining spatial-numerical associations to make their case.

Researchers have long reasoned that numbers might be represented in a spatially organized 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 parity-judgment experiments conducted by Dehaene et al. (1993). In their experiments, Dehaene et al. (1993) asked participants to judge whether a number was odd or even and reported that responses to large numbers were faster when participants pressed a right-hand key rather than a left-hand key, whereas the opposite was true for small numbers. They labeled this number-magnitude-by-response-side interaction the spatial-numerical association of response codes (SNARC) effect.

In these experiments, there was no standard with which to compare the presented number. Consequently, whether a particular number was responded to more quickly with the left hand or the right hand was not determined by the absolute magnitude of the number, but rather by the relative magnitude

of the number within a stimulus set. Thus, the number 5 was responded to more quickly with the left hand when it appeared in a set of numbers ranging from 4 to 9 but more quickly with the right hand when it appeared in a set of numbers ranging from 0 to 5 (e.g., Dehaene et al., 1993; Fias et al., 1996).

Dehaene et al. (1993) reported that the effect was dependent on neither the handedness of participants nor the hand used to make the response, but instead depended on the side of space of the response: When participants' hands were crossed, responses to small numbers were quicker with the right hand than with the left (however, see, Wood et al., 2006). Nonetheless, Dehaene et al. (1993) did report that the effect was dependent on participants' reading and writing direction. Specifically, although they reported finding the effect in experiments with French participants, who had experience reading and writing from left to right, they also reported failing to find the effect in an experiment with Iranian participants, who had experience reading and writing from right to left (see Shaki et al. (2009) and Zebian (2005)). Together, the results from the nine experiments reported in Dehaene et al. (1993) were taken to support the idea of a mental number line and the association of numbers of increasing magnitude with the left-to-right axis of external space.

Although the SNARC effect appears to be robust (see Wood et al. (2008) and Toomarian and Hubbard (2018) for recent reviews), the great range of findings has resulted in debate about mechanism. One such debate concerns 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 a 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) noted, 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 behavioral (Keus & Schwarz, 2005) and psychophysiological (Keus et al., 2005) evidence to argue in favor 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 et al., 2006).

In addition, response-polarity-related accounts break the link between a number, space, and the SNARC effect. For example, Proctor and Cho (2006) argued 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. Thus, perceptual or conceptual overlap between the stimulus and response dimensions is not required for the SNARC effect to occur. In short, Gevers et al. (2006) model and Proctor and Cho (2006) account 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) reported that SNARC-like effects can be produced when left-right responses are replaced with unimanual close-far responses; small numbers are associated with close responses, and large numbers are associated with far responses. Further, Landy et al. (2008) reported that verbal “yes” and “no” responses on a parity-judgment task were facilitated by large numbers and small numbers, respectively.

Finally, still other researchers have argued in favor of a working memory account of the SNARC effect. For example, in an experiment reported by van Dijck and Fias (2011), participants performed a fruit/vegetable classification task after having been encouraged to store the stimuli as an ordered set in working memory. Specifically, a sequence of fruit and vegetable names was displayed in the center of the computer screen, and participants were tested on the order of the items. Then, in a subsequent classification task, responses to items that had appeared early in the sequence were faster if made with the left hand rather than the right hand, and responses to items that had appeared later in the sequence were faster if made with the right hand rather than the left hand. The authors argued that this working

memory account can also explain why SNARC-like effects emerge for other kinds of ordinal sequences, such as months of the year (Gevers et al., 2003) or days of the week (Gevers et al., 2004), as well as why spatial-numerical associations can be moderated by giving participants instructions to associate numbers with positions on a clockface (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 et al., 1998).

Given that several competing accounts of the SNARC effect exist and that many of these 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 representations 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 Posner's (1980) attentional cuing task developed by Fischer et al. (2003). In Fischer et al.'s experiment, participants were asked to press a single response button whenever a lateralized target, a white circle, appeared, regardless of whether it appeared on the left or the right. The target was always preceded by either a small number (1 or 2) or a large number (8 or 9), which was unrelated to the subsequent location of the target. Because the response was not lateralized, response-related effects were not possible. Results from this paradigm were consistent with the SNARC effect, as participants were quicker to detect left-side targets when they were preceded by small numbers and quicker to detect right-side targets when they were preceded by large numbers, at least when the numbers and targets were separated by an interstimulus interval (ISI) between 250 and 1000 ms. This finding—named the attentional SNARC (Att-SNARC) effect—suggests that viewing a number can cue spatial attention either to the left or to the right depending on the magnitude of the number.

Because the Att-SNARC effect is strong evidence in favor of an early, response-independent locus for the mechanism underlying 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, Fischer et al.'s original finding has been extremely influential (e.g., cited

746 times according to Google Scholar as of May 15, 2020). However, subsequent attempts to replicate the effect have produced a wide array of results.

Galfano et al. (2006) reported a so-called statistically significant effect for left-side targets when the data were aggregated over ISI conditions of 500 and 800 ms and a one-tailed test was employed, estimate = 6 ms, $t(25) = 1.75$, $p = .046$ (reported as $p = .04$). They also reported a statistically significant effect for right-side targets when the data were aggregated over these two ISI conditions and a one-tailed test was employed, but the claimed statistical significance reflected a reporting error, estimate = 5 ms, $t(25) = 1.59$, $p = .062$ (reported as $p = .04$). Although it is possible to obtain a point estimate for each of the ISI conditions with the data aggregated over the left- and right-side targets (500-ms ISI: 8 ms; 800-ms ISI: 4 ms), the corresponding variances and test statistics for these estimates were not reported and cannot be obtained from what was reported.

Ristic et al. (2006) reported a statistically significant effect when the data were aggregated over six ISI conditions ranging from 350 to 800 ms and over the left- and right-side targets, estimate = 3.79 ms (unreported; obtained via digitization of the figure), $F(1, 17) = 5.48$, $p = .032$ (reported as $p < .05$). Although it is possible, via digitization of the figure, to obtain a point estimate for each of the six ISI conditions with the data aggregated over the left- and right-side targets (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), the corresponding variances and test statistics for these estimates were not reported and cannot be obtained from what was reported.

Dodd et al. (2008) reported a statistically significant effect when the data were aggregated over three ISI conditions ranging from 250 to 750 ms and over the left- and right-side targets, but the claimed statistical significance reflected a reporting error, estimate = 5.5 ms (unreported), $F(1, 29) = 4.05$. $p = .054$ (reported as $p < .05$). They also reported statistically significant effects for the 500-ms ISI condition for left-side targets, estimate = 16 ms, $t(29) = 2.48$, $p = .010$ (reported as $p < .05$), and for right-side targets, estimate = 6 ms, $t(29) = 2.34$, $p = .013$ (reported as $p < .05$). Although it is possible to obtain a point estimate for each of the three ISI conditions with the data aggregated over the left- and

right-side targets (250-ms ISI: 6 ms; 500-ms ISI: 11 ms; 750-ms ISI: –0.5 ms), the variances and test statistics for these estimates were not reported and cannot be obtained from what was reported.

Salillas et al. (2008) reported a so-called statistically nonsignificant effect for a 450-ms ISI condition when the data were aggregated over the left- and right-side targets, estimate = 7.5 ms, $F(1, 11) = 1.3$, $p = .28$ (reported as "ns"). Additionally, Ranzini et al. (2009) reported a statistically nonsignificant effect when the data were aggregated over three ISI conditions ranging from 300 to 500 ms and over the left- and right-side targets, estimate = 3 ms (unreported; obtained via digitization of the figure), $F(1, 14) = 4.1$, $p = .06$. Point estimates and variances and test statistics for such estimates for the three ISI conditions with the data aggregated over the left- and right-side targets were not reported and cannot be obtained from what was reported.

More recently, van Dijck et al. (2014) reported a statistically nonsignificant effect when the data were aggregated over four ISI conditions ranging from 250 to 1000 ms and over the left- and right-side targets, estimate = 1 ms (unreported; obtained via digitization of the figure), reported $F(1, 42) < 1.05$, reported $p > .37$. Point estimates and variances and test statistics for such estimates for the four ISI conditions with the data aggregated over the left- and right-side targets were not reported and cannot be obtained from what was reported. In a second experiment, van Dijck et al. (2014) also reported a statistically nonsignificant effect when the data were aggregated over three ISI conditions ranging from 100 to 700 ms and over the left- and right-side targets, estimate = –2.5 ms (unreported; obtained via digitization of the figure), $F(1, 28) = 2.94$, $p = .097$ (no estimates were reported). Point estimates and variances and test statistics for such estimates for the three ISI conditions with the data aggregated over the left- and right-side targets were not reported and cannot be obtained from what was reported.

Zanolie and Pecher (2014) reported a statistically nonsignificant effect when the data were aggregated over four ISI conditions ranging from 250 to 1000 ms and over the left- and right-side targets, estimate = 0.5 ms (unreported; obtained via digitization of the figure), $F(1, 19) = 0.03$, $p = .863$. Although it is possible to obtain a point estimate for each of the four ISI conditions with the data aggregated over the left- and right-side targets (250-ms ISI: –1 ms; 500-ms ISI: 2 ms; 750-ms ISI: 5 ms;

1000-ms ISI: –4 ms), the variances and test statistics for these estimates were not reported and cannot be obtained from what was reported. In a second experiment, Zanolie and Pecher (2014) also reported a statistically nonsignificant effect when the data were aggregated over the same four ISI conditions and over the left- and right-side targets, estimate = –1.5 ms (unreported; obtained via digitization of the figure), $F(1, 23) = 0.17$, $p = .686$. Although it is possible to obtain a point estimate for each of the four ISI conditions with the data aggregated over the left- and right-side targets (250-ms ISI: –2 ms; 500-ms ISI: 5 ms; 750-ms ISI: –3 ms; 1000-ms ISI: –6 ms), the variances and test statistics for these estimates were not reported and cannot be obtained from what was reported.

Finally, Fattorini et al. (2015) reported a statistically nonsignificant effect when the data were aggregated over 500-ms and 700-ms ISI conditions and over the left- and right-side targets, estimate = 2 ms (unreported; obtained via digitization of the figure), $F(1, 59) = 1.69$, $p = .20$. Point estimates and variances and test statistics for such estimates for the two ISI conditions with the data aggregated over the left- and right-side targets were not reported and cannot be obtained from what was reported. In a second experiment, Fattorini et al. (2015) also reported a statistically nonsignificant effect when the data were aggregated over four ISI conditions ranging from 250 to 1000 ms and over the left- and right-side targets, estimate = –1.75 ms (unreported; obtained via digitization of the figure), $F(1, 31) = 1.5$, $p = .22$. Although it is possible to obtain a point estimate for each of the four ISI conditions with the data aggregated over the left- and right-side targets (250-ms ISI: –2 ms; 500-ms ISI: –1 ms; 750-ms ISI: –2 ms; 1000-ms ISI: –2 ms), the variances and test statistics for these estimates were not reported and cannot be obtained from what was reported.

A natural approach to assessing these various attempts to replicate the Att-SNARC effect would involve synthesizing the evidence across all published studies of the effect via meta-analysis. This would allow for, among other things, the estimation of 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 approach is complicated because (a) the statistical significance (or nonsignificance) of a study's results typically affects whether or not the study is published, which results in a set of published

studies that is not representative, and (b) meta-analytic results are biased when the set of studies analyzed is not representative (McShane et al., 2016; Ioannidis, 2008). Instead, the Registered Replication Report (RRR) format pursued in the present study provides an ideal means of assessing the Att-SNARC effect because in an RRR, results from all participating labs are included in the meta-analysis regardless of their statistical significance or nonsignificance. Further, preregistration of the primary hypotheses and statistical analyses mitigates some potential biases.

An additional benefit of the RRR format is that it allows for the investigation of potential moderators not previously considered, which might shed new light on mechanism and perhaps also the wide array of results observed in the various attempts to replicate the Att-SNARC effect. Consequently, in addition to replicating the experimental protocol of Fischer et al., we investigated several variables that could potentially moderate the Att-SNARC effect: finger-counting habits, reading and writing direction, handedness, and mathematics ability and mathematics anxiety (see Fischer (Fischer, 2006; Fischer, 2008), Fischer and Knops (2014), Georges et al. (2016), and Shaki et al. (2009) for details and conjectures).

Before proceeding, we note that alternative accounts of the effect reported by Fischer et al. have been suggested. These include, for example, accounts based on working memory (van Dijck et al., 2014). We also note that manipulations that make explicit associations between number and space have been able to produce Att-SNARC-like effects (e.g., Fattorini et al., 2015, experiment 3). Because these alternative accounts and additional manipulations have theoretical implications for the Att-SNARC effect that differ that originally proposed and our focus is on the latter, we do not consider them here.

Disclosures

Preregistration

This study was preregistered. All relevant documentation is available on the Open Science Framework (OSF) at https://osf.io/he5za/

Data, materials, and online resources

The data and materials are available on OSF at https://osf.io/he5za/. Links to the lab-specific pages of all participating labs are available on OSF at https://osf.io/7zyxj. Data and scripts to re-create the manuscript are available on a companion website at http://git.colling.net.nz/attentional_snarc/. An archived version of the companion website is available at https://doi.org/10.5281/zenodo.3738555.

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

All participating labs obtained ethical approval in accordance with their local requirements, and the research was carried out in according with the Declaration of Helsinki.

Methods

Sample size

Each participating lab was required to provide a target sample size no smaller than 60 participants and a stopping rule (see the lab-specific pages for details). We chose 60 participants as the minimum because, as required for RRRs, it provides high power conditional on a hypothetical assumed effect size of 0.92 for a one-tailed test at $\alpha = .05$, conditional on an effect size of 0.4 on the standardized Cohen's *d* scale, about the midpoint of previously published estimates. This value corresponds to a raw effect size of 6 ms assuming a between-participants standard deviation of 15 ms, again about the midpoint of previously published estimates.

Because of time constraints, not all labs were able to reach the minimum target of 60 participants (see Table 1 for the sample size achieved by each lab). However, given the sample sizes actually achieved, and again conditional on an effect size of 0.4 on the standardized Cohen's *d* scale, a statistically significant effect would be expected in 93% of the labs (i.e., about 16). Thus, if 0.4 is a

reasonable estimate of the effect size and there are no substantial moderators of the effect, statistically significant effects would be expected not only at the meta-analytic level but also at the level of the individual lab.

Materials

The participating labs all had (a) a testing station, such as a room or a cubicle, where participants could undertake the experiment without distraction; (b) a computer for presenting stimuli and recording responses; (c) a chin rest or similar device to ensure that participants remained a set distance from the computer monitor; and (d) a tape measure used to calibrate distance from the screen. Five labs also optionally made use of an eye tracker to record participants' eye movements during the attentional-cuing task (see the lab-specific pages for details).

An instruction booklet detailing how to perform the setup and calibration procedure and the finger-counting assessment was provided to the labs. These materials were initially written in English, but each lab conducted the experiment in the predominant language of its locale. Thus, the experiment was also conducted in German, Dutch, Czech, Spanish, Italian, and Chinese. All materials were translated from English into these other languages 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 et al., 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 et al., 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 et al., 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 et al., 2017), *reticulate* (Version 1.10; Allaire et al., 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 Experiment 2 had fewer ISI conditions and because the results were statistically significant in a greater proportion of the conditions. Before starting data collection, each lab performed a monitor calibration procedure using a supplied calibration script. This procedure involved measuring the viewing distance from the computer monitor and the size of standard stimuli presented on the screen (see https://osf.io/2m4ad/ for details). After participants provided informed consent, they were seated in front of the monitor with their chin placed in 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 was identical to that of Fischer et al. and did not include timing modifications for the eye tracker (see the Eye-Tracking Protocol subsection for details), is shown in Figure 1. The initial display on each trial consisted of a centrally located white fixation point (0.2° diameter) flanked by two white outline boxes (1° $\times$ 1°), all on a black background. The centers of the boxes were located 5° from the center of the fixation point. This initial display was shown for 500 ms. Next, a digit (1, 2, 8, or 9; height of 0.75°) replaced the fixation point for a fixed duration of 300 ms. After the digit was removed, the fixation point reappeared. Finally, a circular white target (0.7° diameter) appeared in either the left- or the right-side box after a variable duration (250 ms, 500 ms, 750 ms, or 1000 ms) on target trials, and no target appeared on catch trials.

Target trials ended after a response was made or 1000 ms after the onset of the target, whichever came first. Catch trials ended 1000 ms after the digit was removed. Trials advanced automatically,

separated by an intertrial interval of 1000 ms.

Participants responded to the appearance of the target by pressing the space bar with their preferred hand. When a participant responded before the target appeared or responded on a catch trial, the trial ended, and the following warning appeared: "Too quick! Please wait until the target appears in a box before pressing SPACE" [English version]. When a participant failed to respond on a target trial, the following warning was presented: "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; the trials in each block were evenly divided across the four ISI conditions, four digits, and two target locations, and the order of presentation was random.

Eye-tracking protocol

Code implementing an eye-tracking protocol using an EyeLink 1000 (SR Research, Ottawa, Ontario, Canada) eye tracker was provided to all labs and is available at Github (https://github.com/ljcolling/FischerRRR-eyetracking). Of the five labs that optionally made use of an eye tracker, one used a different eye tracker; this lab has provided information regarding deviations from the standard protocol on its lab-specific page. The standard nine-point grid was used for calibration and validation at the start of each block and when required during a block. The start of a trial was triggered after the detection of 500 ms of stable fixation within a 2° box centered 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 centered on the fixation point were recorded. Any deviations toward the lateral boxes that exceeded 2° resulted in the trial being marked as contaminated. These trials were excluded from primary analyses; however, they were analyzed separately in an attempt to determine any possible effect of eye movements on the results.

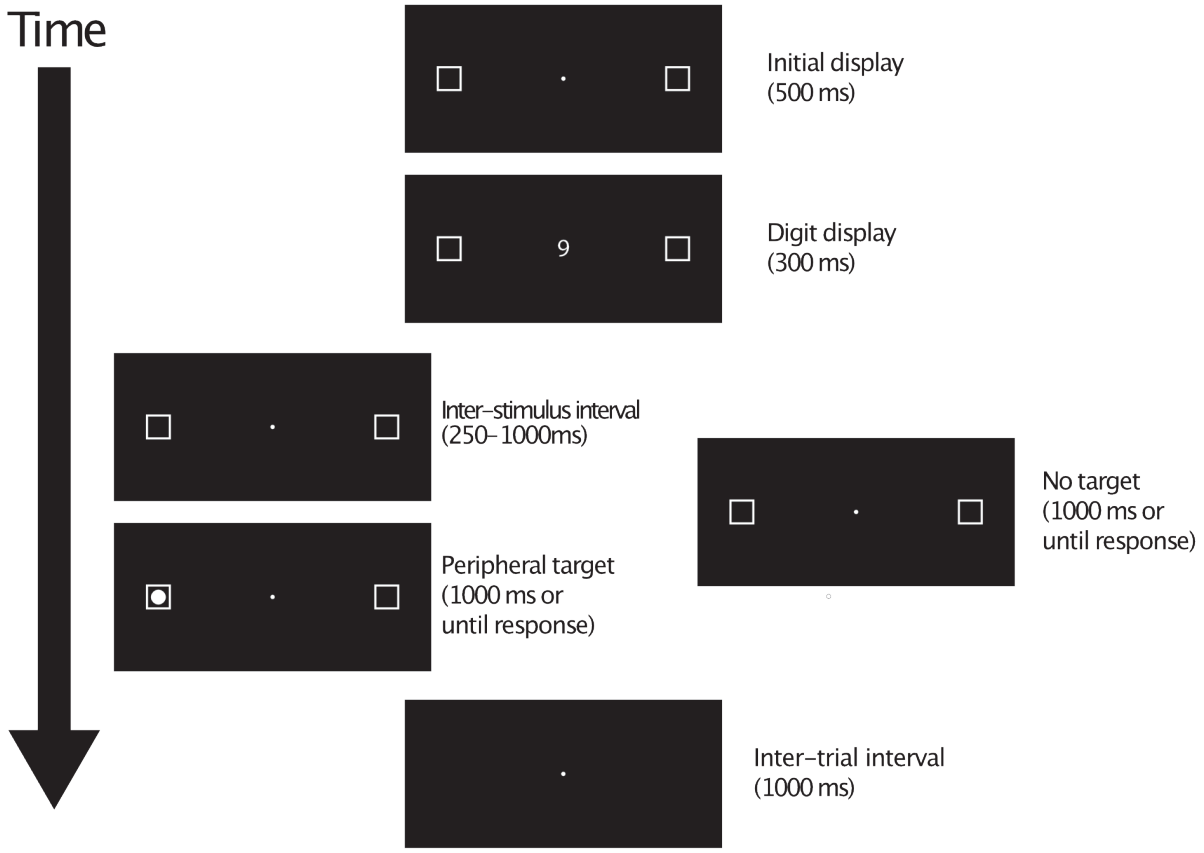


Figure 1. Trial structure for target trials and catch trials. The initial display on each trial consisted of a centrally located white fixation point flanked by two white outline boxes, all on a black background. Next, a digit replaced the fixation point. After the digit was removed, the fixation point reappeared. Finally, a circular white target appeared in either the left- or the right-side box after a variable duration on target trials, and no target appeared on catch trials. Target trials ended after a response was made or 1000 ms after the onset of the target, whichever came first. Catch trials ended 1000 ms after the digit was removed. Trials advanced automatically, separated by an intertrial interval of 1000 ms.

Finger counting

To assess finger-counting fluency, we used a task derived from that developed by Lucidi and Thevenot (2014). Participants were asked to read aloud four sentences while counting the number of syllables in each. Because reading aloud prevents verbalizing counting, most participants needed 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. Although most participants used their fingers for the task, some participants adopted a different strategy. Participants 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 the lab-specific pages for details).

The results from the finger-counting task were used to place participants into five groups: consistent left-starters, consistent right-starters, inconsistent left-starters, inconsistent right-starters, and others. This classification was determined not only by participants' hand choices, but also by how consistently they engaged in finger counting. The consistent left-starters and consistent right-starters included those participants who counted using a hand on all four occasions and started on the same hand on at least three of them. The inconsistent left-starters and inconsistent right-starters included participants who counted using a hand on two or three occasions and started on the same hand on at least two of them. The *other* group included all remaining participants (e.g., those who did not count on their fingers, those who counted on their fingers only once, and those who counted an equal number of times with each hand).

Reading/writing direction

To assess reading and writing direction, we used a simple question asking participants if they had experience with languages that are written exclusively from left to right (e.g., English and German), with languages that are not written exclusively from left to right (e.g., Hebrew), or with languages of both types (see https://osf.io/dqnkq/ for details). For the Chinese version of this question, participants were asked if they had experience with languages that are usually written horizontally, with languages that are usually written vertically, or with languages of both types (see https://osf.io/r3fhx/ for details).

Responses to this question were used to place participants into two groups: exclusively left-to-right readers-writers and not exclusively left-to-right readers-writers. Participants who selected the first option were placed in the left-to-right readers-writers group, and all the remaining participants were placed in the not-exclusively left-to-right readers-writers group.

Handedness

To assess handedness, we used Nicholls, Thomas, Loetscher, and Grimshaw's (2013) 10-item questionnaire. 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/r3fhx/ 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 of the Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock & Johnson, 1989). It contains 25 multiple-choice mathematics questions requiring addition, subtraction, multiplication, and division. Participants had 30 s to select the response on each trial; the timing was 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 25 questions were split into five sets of 5 questions each. Two errors on a single set or errors on consecutive sets terminated the test. The final score was the total number of correct answers.

Mathematics anxiety

To assess mathematics anxiety, we used the Abbreviated Math Anxiety Scale (AMAS; Hopko et al., 2003). The AMAS contains nine questions that ask participants to rate (on a scale from 1 to 5) how anxious they would feel during particular events, including thinking of an upcoming mathematics test, taking a mathematics examination, and listening to a mathematics lecture. In labs conducting the experiment in a language other than English, the AMAS was translated. The final score was the sum of the individual ratings; possible scores ranged from 9 (low anxiety) to 45 (high anxiety).

Exit questionnaire

An exit questionnaire that asked participants to describe the purpose of the experiment was used to determine whether they had guessed its purpose. Participants who guessed correctly, as judged by the experimenter, were excluded from primary analyses; however, their data were analyzed separately to determine whether guessing the experiment's purpose moderated the Att-SNARC effect.

Exclusion criteria

Participants who committed errors on more than 5% of the catch trials, who correctly guessed the purpose of the experiment, or who did not undertake all tasks were excluded from the analysis.

Analysis

The dependent variables of interest were the congruency effects in the four ISI conditions (i.e., 250 ms, 500 ms, 750 ms, and 1000 ms). The congruency effect was defined as the average difference in response time between congruent and incongruent trials; congruent trials were defined as trials with left-side targets preceded by low digits (1 or 2) and trials with right-side targets preceded by high digits (8 or 9), and incongruent trials were defined as trials with left-side targets preceded by high digits and trials with right-side targets preceded by low digits. A positive value for the congruency effect indicates that participants were faster responding on congruent trials than on incongruent trials, and a negative value indicates the reverse.

We analyzed 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 can take account of the dependence between multiple dependent variables (here, the congruency effect in each of the four ISI conditions). Second, rather than assuming a simple two-level structure, with participants nested within labs, they can take account of more complex nesting structures (here, participants nested within moderator groups, such as consistent left-starters, consistent right-starters, etc., and moderator groups nested within labs). In short, the standard approach necessitates treating several variance components as zero, and thereby makes unwarranted

independence assumptions.

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

In analyzing moderators, it is ideal to consider them all jointly within a single model. Unfortunately, data sparsity precluded this. When the moderators were considered jointly, many combinations of them resulted in either zero or very few participants per moderator group in each lab. Indeed, this was also the case for some moderators when considered alone (i.e., reading and writing direction and handedness; see Tables S4 and S6, respectively, in the Supplemental Material). Consequently, we consider each moderator separately.

For models featuring no moderators (Model 1) or discrete moderators (finger counting, reading and writing direction, and handedness; Models 2–4, respectively), for simplicity we analyzed the data at the moderator-group level, as per McShane and Böckenholt (2018), using data from moderator groups not precluded for reasons of data sparsity. For the model featuring continuous moderators (mathematics fluency and mathematics anxiety; Model 5), this was not possible, so we analyzed the data at the participant level using an analogous specification (see the Model 5 subsection for details) and using data from all participants. Our motivation for considering these moderators follows.

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

Model 2: Finger counting. Recent work suggests that spatial-numerical compatibility effects in general (Fischer, 2008)—including attentional-cuing effects in response to numbers (Fischer & Knops, 2014)—might be moderated by finger-counting behavior. Specifically, this work suggests that these

effects are stronger among people 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 possibility, and consequently this model 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 categorized as consistent left-starters or consistent right-starters. We restricted the analysis to these two groups principally because if finger-counting behavior has an effect, we would expect it to be most prominent in participants whose finger-counting habits are clear and unambiguous.

Model 3: Reading/writing direction. Recent work suggests that the congruency effect might be weaker or possibly even reversed among people who have experience with languages that are not read and written exclusively from left to right (Fischer, 2008; Shaki et al., 2009). The purpose of Model 3 was to assess this possibility, and consequently this model took account of the reading-and-writing-direction moderator. Specifically, participants were placed into two groups according to their responses on the reading-writing questionnaire: those who read and 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 this model took account of the handedness moderator. Specifically, participants were classified as left-handed or right-handed according to their responses on 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 possibility, and consequently this model jointly took account of both mathematics fluency and mathematics anxiety, as measured by the math test and AMAS, respectively. Specifically, we employed a multilevel model with fixed effects included for the full set of ISI Condition $\times$ Math Test $\times$ AMAS interactions, and random effects included for each participant, for each ISI condition for each lab (with equal variance and zero correlation), and for the math test, the AMAS, and the Math Test $\times$ AMAS interaction for each lab

(independently).

Secondary analyses. The purpose of our secondary analyses was to assess whether insight into the purpose of the experiment or eye movements moderated the congruency effect. Specifically, Model 1 was estimated separately on data from participants who correctly guessed the purpose of the experiment and also separately on data from eye-movement-contaminated trials of participants with contaminated trials in every combination of ISI and congruency condition.

Results

Replication operationalisation

According to the common definition of replication employed in practice, a subsequent experiment has successfully replicated a prior experiment if the results from the two experiments either (a) failed to attain statistical significance or (b) were directionally consistent and attained statistical significance. This definition has been applied analogously in large-scale replication projects such as the present one by comparing the statistical significance (or nonsignificance) of the results from a meta-analysis of the replication studies with the statistical significance (or nonsignificance) of the results from the original study.

However, the null-hypothesis significance-testing paradigm upon which this operationalization of replication is based has been the subject of no small amount of criticism over the decades (Rozenboom, 1960; Meehl, 1978; Cohen, 1994; Gelman et al., 2003; McShane & Gal, 2016; McShane & Gal, 2017), and recent calls to abandon it abound (Amrhein, Trafimow, et al., 2019; McShane, Gal, et al., 2019; Wasserstein et al., 2019; Amrhein, Greenland, et al., 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 light of the scientific hypothesis advanced by Fischer et al. (2003).

Exclusions

In total, seventeen labs contributed data from 1267 participants; 162 were excluded as per our exclusion criteria, which left a total of 1105. See Table 1 for details of the total number of participants recruited by each lab, the number included in the analysis, and the number excluded for each reason; the technical-error category includes those participants who were excluded for having incomplete data because of, for example, equipment failure or experimenter error.

Five labs used an eye tracker for at least some of their participants. Table S11 in the Supplemental Material shows the number of participants in each of these labs tested with an eye tracker, the number of participants whose data were analyzed in our secondary analysis of trials contaminated by eye movement, and the number of such contaminated trials in each combination of ISI condition and congruency condition.

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

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 in the 250-ms ISI condition, 0.53 in the 500-ms ISI condition, 0.48 in the 750-ms ISI condition, and 0.50 in the 1000-ms ISI condition. Further, the number of ISI conditions with a positive congruency effect was zero for 0.06 of the participants, one for 0.26 of the participants, two for 0.36 of the participants, three for 0.26 of the participants, and four for 0.06 of the participants. All of these results are compatible

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

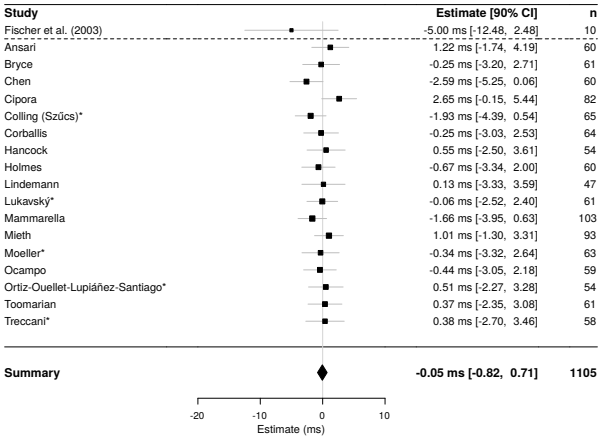
with the relevant binomial distribution with probability parameter .5 (i.e., the distribution of the number of heads on tosses of a fair coin).

Primary analyses

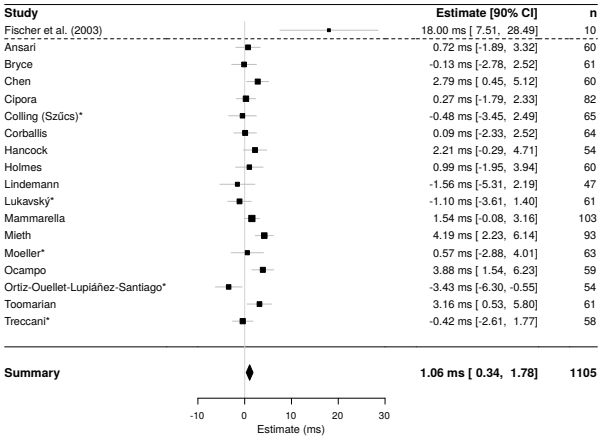
Model 1: No moderators. The effects we observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. Specifically, Fischer et al. estimated an effect of -5.00 ms at the 250 ms ISI condition, 18.00 ms at the 500 ms ISI condition, 23.00 ms at the 750 ms ISI condition, and 11.00 ms at the 1000 ms ISI condition. In contrast, Model 1 estimated effects of -0.05 ms, 1.06 ms, 0.19 ms, and 0.18 ms, respectively, in the four ISI conditions.

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

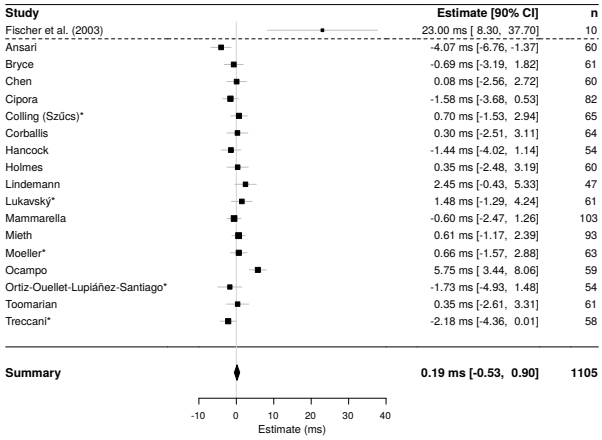
The effects we observed were highly consistent across ISI conditions. They were also highly consistent across labs, perhaps surprisingly given a recent report—contrary to both substantive and statistical expectations—of a nontrivial degree of heterogeneity across labs in large-scale replication projects like the present study (McShane, Tackett, et al., 2019). Specifically, we estimated heterogeneity across labs at 1.02 ms—nonzero but practically unimportant for many purposes (see Table S1 in the Supplemental Material for details). This result suggests that lab-level moderators are unlikely to have driven our results.



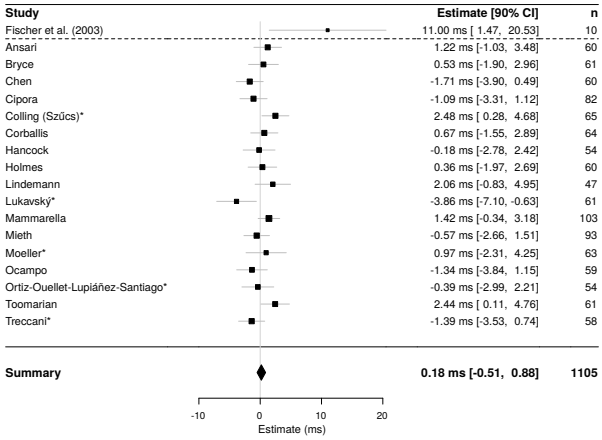
(a) 250 ms ISI Condition



(b) 500 ms ISI Condition



(c) 750 ms ISI Condition



(d) 1000 ms ISI Condition

Figure 2. Summary of results from Experiment 2 of Fischer, Castel, Dodd, and Pratt (2003), each lab in the present study, and Model 1. Each panel presents the estimate for a given interstimulus-interval (ISI) condition: (a) 250 ms, (b) 500 ms, (c) 750 ms, and (d) 1000 ms. The squares give the effect observed in each lab in each ISI condition; the size of each square is inversely proportional to the sample size. The horizontal lines give the 90% confidence interval (CI) in each lab in each ISI condition, and the diamond gives the Model 1 estimate and 90% CI. Labs are identified by the last name of their first authors (as listed in the appendix); labs that used an eye tracker are marked with an asterisk. The effects observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. (2003). They were also highly consistent both across ISI conditions and across labs; the latter result suggests that lab-level moderators are unlikely to have driven our results.

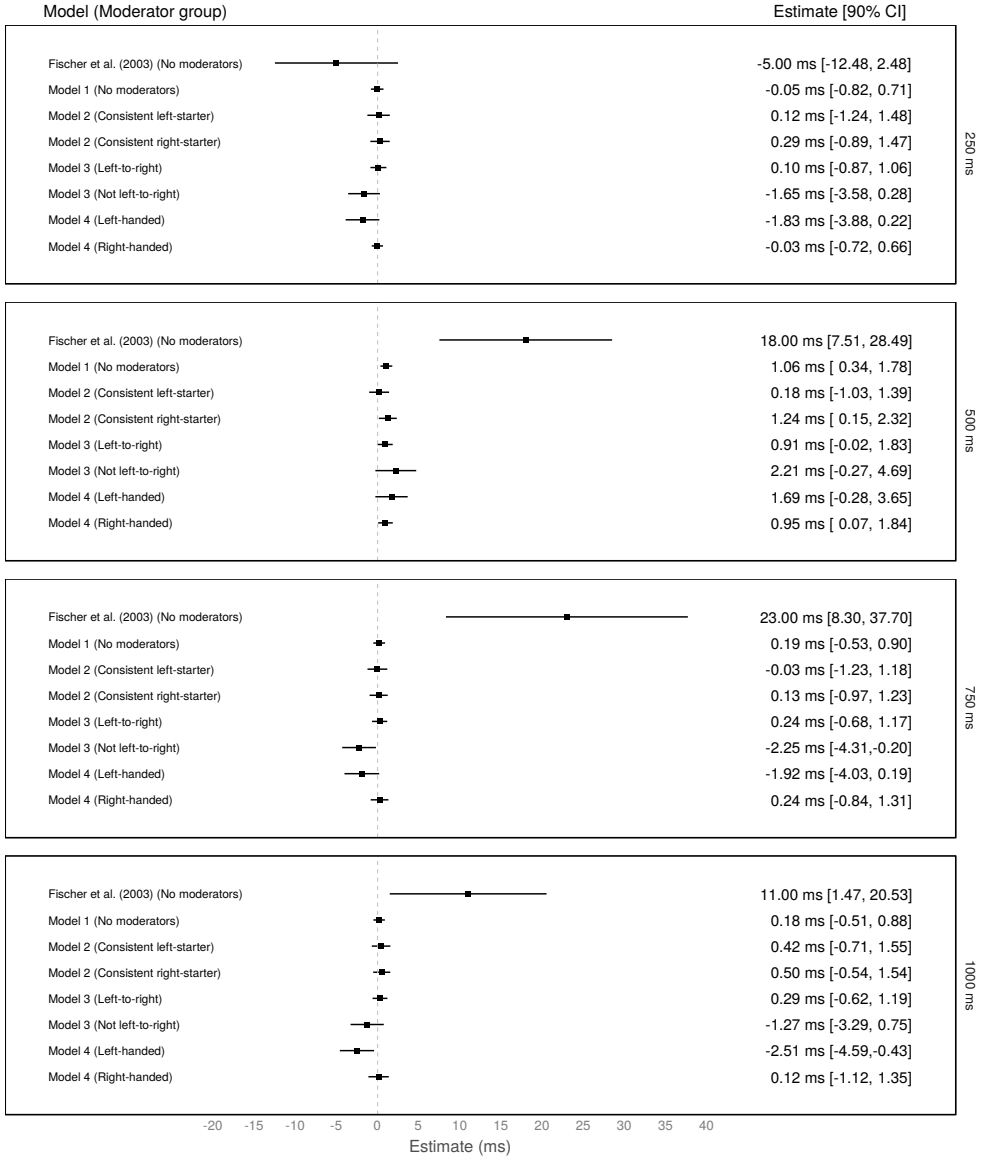


Figure 3. Summary of results from Experiment 2 of Fischer, Castel, Dodd, and Pratt (2003) and Models 1 through 4. Each panel presents the estimates for a given interstimulus-interval condition: from top to bottom, 250 ms, 500 ms, 750 ms, and 1000 ms. The squares give the point estimates, and the horizontal lines give 90% confidence intervals (CIs). The effects observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. (2003). They were also highly consistent across ISI conditions.

Model 2: Finger counting. Model 2 was estimated on data from 343 consistent left-starters from 17 labs and 482 consistent right-starters from 17 labs. We summarize the results from Experiment 2 of Fischer et al. (2003) along with results from Models 1 through 4 in Figure 3. Although previous work suggests a stronger congruency effect among left-starters and a weaker or possibly even reversed effect among right-starters, Figure 3 shows that finger counting had no substantial impact on the results. Specifically, the figure shows a minuscule congruency effect for each finger-counting group in each ISI condition and minuscule differences between congruency effects for the two finger-counting groups in each ISI condition (see Tables S2 and S3 in the Supplemental Material for details).

Model 3: Reading/writing direction. Model 3 was estimated on data from 1014 exclusively left-to-right readers-writers from 17 labs and 76 not exclusively left-to-right readers-writers from 8 labs. Although previous work suggests a weaker or possibly even reversed congruency effect among participants who have experience with languages that are not read and written exclusively from left to right. Figure 3 shows that reading and writing direction had no substantial impact on the results. Specifically, the figure shows a minuscule effect for each reading-and-writing-direction group in each ISI condition and minuscule differences between the congruency effects for the two reading-and-writing direction groups in each ISI condition (see Tables S4 and S5 in the Supplemental Material for details).

Model 4: Handedness. Model 4 was estimated on data from 69 left-handed participants from 9 labs and 1007 right-handed participants from 17 labs. Figure 3 shows that handedness had no substantial impact on the results. Specifically, the figure shows a minuscule effect for each handedness group in each ISI condition and minuscule differences between the congruency effects for the two handedness groups in each ISI condition (see Tables S6 and S7 in the Supplemental Material for details).

Model 5: Mathematics fluency and mathematics anxiety. Model 5 was estimated on data from 1105 participants from 17 labs. Although previous work suggests that mathematics fluency and mathematics anxiety might moderate congruency effects, we observed no substantial moderating effects (see Table S8 in the Supplemental Material for details).

Secondary analyses

Model 1 was estimated separately on data from 41 participants (from four labs) who correctly guessed the purpose of the experiment and also separately on data from 10468 eye-movement-contaminated trials of 132 participants (from five labs) with contaminated trials in every combination of ISI and congruency condition. These analyses yielded no results of substantive interest (see the Supplemental Material for details).

Discussion

The Att-SNARC effect 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 support strong claims about the link between number and space (e.g., a mental number line).

We attempted to replicate Experiment 2 of Fischer et al. (2003) by collecting data from 1105 participants at 17 labs. Across these 1105 participants and four ISI conditions, the proportion of times the congruency effect we observed was positive was .50. Further, the effects we observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. Given this, we conclude that we failed to replicate the effect reported by Fischer et al.

The effects we observed were highly consistent both across ISI conditions and across labs; the latter result suggests that lab-level moderators are unlikely to have driven our results. In addition, our analyses of several participant-level moderators (finger-counting habits, reading and writing direction, handedness, and mathematics fluency and mathematics anxiety) revealed no substantial moderating effects.

We conclude with two important points. First, on the basis of the common definition of replication employed in practice, one might object that we did in fact successfully replicate Fischer et al. (2003), at least in the 500-ms ISI condition. In response, we argue that this objection illustrates one major flaw of that definition: Our result in 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 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., and that potential new paradigms could reveal an effect. Nonetheless, even if such paradigms are forthcoming, we maintain on the basis of our results that the paradigm of Fischer et al. provides no evidence of such a mechanism.

Second, we note several limitations of the present study. First and foremost, although 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 earlier, our results do not refute such accounts. Specifically, although one might, on the basis of our results, prefer accounts of the SNARC effect that do not imply a mental number line, the evidence for and against different claims about the SNARC effect must be viewed in its entirety. The Att-SNARC effect provides only one such piece of evidence—albeit a particularly strong and valuable one.

In addition, a set of limitations relates to our sample of participants. Our sample was recruited primarily from North America, Europe, and Australasia. Consequently, participants who read and wrote exclusively from left to right are overrepresented in our data. As reading and writing direction has been shown to strongly moderate spatial-numerical associations, it would have been preferable to have more participants with experience with languages that are not read and written exclusively from left to right. In addition, data sparsity precluded considering all moderators jointly in a single model, and thus we considered 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 employed finger counting inconsistently, and they were therefore 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

L. J. Colling and D. Szűcs proposed the study. L. J. Colling programmed the experiments. L. J. Colling and B. B. McShane developed the analysis plan and conducted the analyses. L. J. Colling wrote an initial manuscript. L. J . Colling and B. B. McShane wrote revised and final manuscripts. All authors critically reviewed the manuscript by providing comments, feedback, and edits at all stages of writing, and all authors approved the final manuscript. All authors were involved in data collection. Authors from the contributing labs provided translated materials where required (see the appendix).

References

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

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

Amrhein, V., Greenland, S., & McShane, B. B. (2019). Retire statistical significance. *Nature*, *7748*(567), 305–307. https://doi.org/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. https://doi.org/10.1080/00031305.2018.1543137

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

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r* [R package version 1.5]. R package version 1.5. 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. https://doi.org/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. https://doi.org/10.18637/jss.v067.i01

Bates, D., & Maechler, M. (2018). *Matrix: Sparse and dense matrix classes and methods* [R package version 1.2-14]. R package version 1.2-14. 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]. R package version 3.6.2. https://CRAN.R-project.org/package=R.matlab

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/135467996387552

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

Fischer, M. H. (2008). Finger counting habits modulate spatial-numerical associations. *Cortex*, *44*(4), 386–392. https://doi.org/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. https://doi.org/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. https://doi.org/10.1038/nn1066

Fischer, M. H., & Knops, A. (2014). Attentional cueing in numerical cognition. *Frontiers in Psychology*, *5*(325), 426. https://doi.org/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. https://doi.org/10.3758/BF03194011

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis*. Chapman; 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. https://doi.org/10.1177/1745691614551642

Georges, C., Hoffmann, D., & Schiltz, C. (2016). How Math Anxiety Relates to Number–Space Associations. *Frontiers in Psychology*, *7*(33), 143. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1007/s10539-017-9562-6

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

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

Hester, J. (2018). *Glue: Interpreted string literals* [R package version 1.3.0]. R package version 1.3.0. 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. https://doi.org/10.1177/1073191103252351

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/0022-006X.46.4.806

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

Müller, K., & Wickham, H. (2018). *Tibble: Simple data frames* [R package version 1.4.2]. R package version 1.4.2. 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. https://doi.org/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. https://doi.org/10.1016/j.cortex.2013.02.002

Ooms, J. (2018). *Magick: Advanced graphics and image-processing in r* [R package version 1.9]. R package version 1.9. 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). https://doi.org/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]. R package version 3.1-137. https://CRAN.R-project.org/package=nlme

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/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. https://doi.org/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. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/BF03194010

Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/PBR.16.2.328

Tibber, M. S., Manasseh, G. S. L., Clarke, R. C., Gagin, G., Swanbeck, S. N., Butterworth, B., Lotto, R. B., & Dakin, S. C. (2013). Sensitivity to numerosity is not a unique visuospatial psychophysical predictor of mathematical ability. *Vision Research*, *89*, 1–9. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/00031305.2019.1583913

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

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

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

Wickham, H. (2018b). *Stringr: Simple, consistent wrappers for common string operations* [R package version 1.3.1]. R package version 1.3.1. 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]. R package version 0.7.6. 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]. R package version 0.8.1. 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]. R package version 1.1.1. 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. https://doi.org/10.1007/s11229-017-1578-6

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636. https://doi.org/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. https://doi.org/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; Hall/CRC.

Zanolie, K., & Pecher, D. (2014). Number-induced shifts in spatial attention: a replication study. *Frontiers in Psychology*, *5*(e85048), 667. https://doi.org/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. https://doi.org/10.1163/1568537054068660

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

Supplementary Results

Primary analyses

Model 1: No Moderators. Model 1 was estimated on 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 S1.

Model 2: Finger counting. Model 2 was estimated on data from 343 consistent left-starters from seventeen labs and 482 consistent right-starters from seventeen labs (see Supplementary Table S2 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 S3.

Model 3: Reading/writing direction. Model 3 was estimated on 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 S4 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 S5.

Model 4: Handedness. Model 4 was estimated on data from 69 left-handed participants from nine labs and 1007 right-handed participants from seventeen labs (see Supplementary Table S6 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 S7.

Model 5: Mathematics fluency and mathematics anxiety. Model 5 was estimated on 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 effects for each ISI condition for each lab and (ii) the math 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 S8.

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 analyzed separately to determine whether insight into the purpose of the experiment moderated the effect. Specifically, Model 1 was estimated on data from the 41 participants from four labs who correctly guessed the purpose of the experiment (see Supplementary Table S9 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 S10.

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 analyzed separately to determine whether eye movements moderated the effect. Specifically, Model 1 was estimated on data from 10468 eye movement contaminated trials of 132 participants from five labs with contaminated trials in every combination of ISI and congruency congruency condition (see Supplementary Table S11 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 S12

Table S1

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 S2

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

Lab	Consistent Left-starter	Inconsistent Left-starter	No Group	Inconsistent Right-starter	Consistent 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 S3

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	Consistent Right-starter	0.29	0.72	0.40
250 ms	Consistent Left-starter	0.12	0.83	0.14
500 ms	Consistent Right-starter	1.24	0.66	1.88
500 ms	Consistent Left-starter	0.18	0.74	0.24
750 ms	Consistent Right-starter	0.13	0.67	0.19
750 ms	Consistent Left-starter	-0.03	0.73	-0.04
1000 ms	Consistent Right-starter	0.50	0.63	0.79
1000 ms	Consistent 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 S4

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 S5

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 S6

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 S7

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 S8

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 S9

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 S10

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 S11

Number of participants tested with an eye-tracker, number of participants analyzed 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 combination of ISI and 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 S12

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