One and only SNARC?

Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Hans-Christoph Nuerk[1,4,5], Annika Tave Overlander[2], Krzysztof Cipora[3]*

[1]Department of Psychology, University of Tübingen, Germany

[2]Department of Psychology, University of Konstanz, Germany

[3]Centre for Mathematical Cognition, Loughborough University, United Kingdom

[4]LEAD Graduate School & Research Network, University of Tübingen, Germany

[5]German Center for Mental Health (DZPG)

*Corresponding author:

k.cipora@lboro.ac.uk

Abstract

Numbers are associated with space, but it is unclear how flexible these associations are. We investigated whether the SNARC effect (Spatial-Numerical Association of Response Codes; Dehaene et al., 1993; i.e., faster responses to small/large number magnitude with the left/right hand, respectively) is fully flexible (depending only on relative magnitude within a stimulus set), or not (depending on absolute magnitude as well). Evidence for relative-magnitude dependency came from studies observing that numbers 4 and 5 were associated with the right when presented in a 0 – 5 range but with the left in a 4 – 9 range (Dehaene et al., 1993; Fias et al., 1996). However, this conclusion was drawn from the absence of evidence for absolute-magnitude dependency in frequentist analyses in underpowered studies. Within this Registered Report, we conducted two~~In two highly powered Registered Report~~ online experiments running Bayesian analyses with optional recruitment stopping at moderate evidence for (BF_{10} above 3) or against (BF_{10} below 1/3) each hypothesis. Experiment 1 (N = 200) replicated relative-magnitude dependency using the same stimuli as Fias et al. and Dehaene et al. ~~(https://doi.org/10.17605/OSF.IO/AE2C8, IPA by PCI: 12/03/2023), we observed and here report such a dependency on absolute magnitude (in addition to the replication of effects of relative magnitude).~~ However, Experiment 2 (N = 300) additionally demonstrated absolute-magnitude dependency, while considering recent advances in SNARC research (mainly by improving the stimulus sets and using 1 – 5 excluding 3 and 4 – 8 excluding 6). We conclude that the frequently perpetuated notion of fully flexible spatial-numerical associations is wrong. Some fixed relation to absolute magnitude prevails, especially for some numbers. We suggest that these findings have important consequences for how spatial-numerical associations might support numerical ~~and arithmetic~~ processing.

Keywords: spatial-numerical associations, SNARC effect, mental number line, replication, flexibility

One and only SNARC?

Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude

Numbers are highly relevant in everyday life. Therefore, much research has been devoted to understanding how we process and represent them in our minds. Interestingly, various aspects of numerical information such as cardinality and ordinality are systematically associated with different aspects of space such as extensions or directions (Cipora et al., 2020; Cipora, Schroeder et al., 2018; Patro et al., 2014). This broad range of phenomena is referred to under the umbrella term Spatial-Numerical Associations (SNAs; Fischer & Shaki, 2014; Toomarian & Hubbard, 2018). Investigating these associations is fundamental for models of number representation and – considering the bigger picture – of human cognition.

The hallmark directional SNA is the Spatial-Numerical Association of Response Codes (SNARC) effect, which denotes that in left-to-right reading cultures, participants respond faster to small/large magnitude numbers on the left/right side, respectively (Dehaene et al., 1993). Interestingly, the SNARC effect can be observed in a parity judgment task, in which the magnitude of the numbers in not task-relevant. This effect has been replicated using different modalities, setups and tasks (see Cipora, Soltanlou, et al., 2019, for an online replication; Fias et al., 1996; Toomarian & Hubbard, 2018, for a recent review; Wood et al., 2008, for a meta-analysis). The SNARC effect is typically quantified using the repeated-measures regression originally proposed by Lorch and Myers (1990) and applied to the SNARC effect by Fias et al. (1996). In the first step mean differences in reaction times (RTs) between the right and left hand (dRTs) are regressed on numerical magnitude for each participant separately. A negative slope indicates an increasing right-hand advantage with increasing number magnitude (the more negative the so-called SNARC slope, the stronger the SNARC effect). Subsequently, to check for the SNARC effect at the group level, individual SNARC slopes are tested against zero with a one-sample *t*-test.

Interestingly, several studies have documented that the SNARC effect is not fixed but might be prone to several types of manipulation (Cipora, Patro, & Nuerk, 2018, for a taxonomy), for instance, changing the number range of the used stimuli, which has been classified as representational, intra-experimental manipulation. The spatial mental number representation seems to be adapted to fit the task at hand. In this work we focus on the extent to which the SNARC effect flexibly adjusts to the specific range of the numbers being used in the task set.

Relative-magnitude dependency of the SNARC effect

The seminal paper by Dehaene et al. (1993) has already demonstrated in Experiment 3 that the SNARC effect depends on the relative rather than the absolute magnitude of numbers. They found the SNARC effect in two different numerical intervals ranging from 0 to 5 and from 4 to 9. In the lower interval, responses to numbers 4 and 5 were faster with the right hand than with the left (typical response pattern for large numbers) and right-hand responses to these numbers were faster than right-hand responses to lower numbers. In contrast, in the higher interval, responses to these numbers were faster with the left hand than with the right (typical response pattern for small numbers) and left-hand responses to these numbers were faster than left-hand responses to higher numbers. This finding was replicated by Fias et al. (1996, Experiment 1). It suggests that the SNARC effect dynamically adapts to the current task set (i.e., numbers being used) and is determined by the relative magnitude of the number within the set rather than its absolute magnitude. We refer to this claim about the SNARC effect as relative-magnitude dependency (RMdependency).

RMdependency is considered as one of the crucial features of the SNARC effect and is taken for granted since these early findings. The results of Dehaene et al.'s (1993) and Fias et al.'s (1996) experiments are widely cited as an argument for the SNARC effect being dependent on the given number range (e.g., by Antoine & Gevers, 2016; Deng et al., 2016; Ginsburg et al., 2014; Ginsburg & Gevers, 2015; Schwarz & Keus, 2004; Pinhas et al., 2013). The

RMdependency of the SNARC effect has been demonstrated by several other studies even going beyond a basic setup comprising judgments on single digit numbers. For instance, Tlauka (2002) found a SNARC effect both when using the two numbers 1 and 100 and when using the two numbers 100 and 900. The number 100 was associated to the right/left when it was the larger/smaller of the two numbers, respectively. Ben Nathan et al. (2009) went even further, showing that the SNARC effect is not only RMdependent on the task level but built up on a trial-to-trial basis. They found the right- and left-key response speed advantages in magnitude judgment tasks to depend on the relative magnitude in comparison to the ever-changing reference number. What is more, evidence for RMdependency of SNARC-like effects goes beyond numerical stimuli. Wühr and Richter (2022) found a SNARC-like effect (association of physically smaller/larger stimuli with the left/right, respectively) to depend on relative rather than absolute stimulus size.

Importantly, RMdependency has also been used as a methodological tool to show that a spatial-numerical phenomenon is in fact the SNARC effect. For instance, Rugani et al. (2015), Di Giorgio et al. (2019), and Giurfa et al. (2022) showed RMdependency to claim that a certain effect they observed in newly hatched chickens, in newborn children, and in honeybees is of the same nature as the SNARC effect. To sum up, there is evidence for the RMdependency of the SNARC effect in various tasks and setups, and it has even been used to validate SNAs.

RMdependency in the light of number-representation models

RMdependency fits well with most theoretical accounts of number representation. The seminal work of Restle (1970) outlining the Mental Number Line (MNL) account, which has been proposed as the first explanation for the SNARC effect (Dehaene et al., 1993), postulates that the MNL is flexible and dynamically adapts to the task demands. In line with this, Pinhas et al. (2013) claim that the resolution of the MNL can be adjusted to the numerical context. The accounts of verbal-spatial coding (Gevers et al., 2010) and polarity correspondence (Proctor & Cho, 2006) are on the one hand in line with RMdependency, but on the other hand they do not

make clear statements about relative magnitude being the *only* decisive factor determining the SNARC effect. Crucially, both accounts assume that long-term number representations underlie the SNARC effect, which hardly justifies the SNARC effect's flexibility (Ginsburg & Gevers, 2015; van Dijck et al., 2015). The working memory account (Fias & van Dijck, 2016; van Dijck & Fias, 2011) originally claimed that the SNARC effect does not rely on long-term number representations, but is instead constructed during task execution, which speaks in favor of pure RMdependency. However, Ginsburg et al. (2014) and Koch et al. (2023) argue that short-term number representations do not always fully overrule long-term number representations. This idea has been incorporated in the hybrid account proposed by van Dijck et al. (2015) as well, and it allows the coexistence of RMdependency and dependency of the SNARC effect on absolute number magnitude (henceforth AMdependency). Furthermore, concurrent RMdependency and AMdependency would also be in line with the idea that multiple number representations and multiple spatial reference frames can be activated and operated simultaneously (Weis et al., 2018). To conclude, the assumption that absolute magnitude plays no role can hardly be derived from theoretical accounts of the SNARC effect.

Hints towards AMdependency of the SNARC effect

In addition to the prominent claims on the RMdependency of the SNARC effect, the literature also provides hints towards an AMdependency of the SNARC effect. It is important to note that AMdependency can, on the one hand, influence the strength of the SNARC effect (reflected by the SNARC slope), and on the other, the location of numbers on the MNL in absolute terms (reflected by the intercept of the regression line and by dRTs of critical numbers that are part of both number ranges). Crucially, the SNARC effect seemed to be stronger in the lower than in the higher number range in both initial studies demonstrating the RMdependency (-20.1 ms vs. -10.9 ms in Dehaene et al., 1993; and -10.18 ms vs. -7.19 ms in Fias et al., 1996), suggesting AMdependency as well. In Fias et al.'s (1996) results, the observed slope difference had approximately an effect size of Cohen's $d = 0.16$ (i.e., the slope difference of 2.99 divided

by the standard deviation for this slope difference of 18.34 ms, which has been calculated with SD = 15.1 ms and SD = 11.2 ms for the lower and higher number ranges, assuming a rather conservative correlation between them of r = 0.05, which corresponds to the correlation we have observed in our previous color judgment tasks, see Roth, Caffier, et al., 2024,, where we also found a stronger SNARC effect in the lower than in the higher half of the stimulus set ranging from 1 to 9). Moreover, the results pointed towards an overall shift of small/large numbers to the left/right on the MNL, respectively, since the smallest-number intercept (i.e., the predicted dRT for the smallest number magnitude of the range, which was 0/4 in the lower/higher range, respectively) was larger in the lower than in the higher range (37.52 ms vs. 14.03 ms in Dehaene et al., 1993; and 15.43 ms vs. 8.82 ms in Fias et al., 1996). However, the mean-number intercepts (i.e., the predicted dRT for the mean number magnitude of the range, which was 2.5/6.5 in the lower/higher range, respectively) did not differ much in Fias et al.'s results (-10.02 ms vs. -9.16 ms). In Dehaene et al.'s results, this intercept seemed to be smaller in the higher number range, but it cannot be calculated exactly based on the data reported in the paper.

Methodological limitations of the two initial studies demonstrating RMdependency

Even if we use the two original studies as a guidance for further investigations, their findings are not very reliable because of several important limitations regarding the design and the interpretation of the results. Both Dehaene et al. (1993) and Fias et al. (1996) found a significant two-way interaction of response side (left vs. right) and magnitude (small vs. medium vs. large). Apart from the repeated-measures regression approach, the SNARC effect can also be quantified as a two-way interaction of response side and magnitude (for methodological considerations, see Fias et al., 1996) or as linear contrast in an ANOVA (Tzelgov et al., 2013). However, the three-way interaction of response side and magnitude with interval (0 to 5 vs. 4 to 9) remained non-significant in both studies. In Fias et al.'s (1996) additional repeated-measures regression the resulting SNARC slopes differed significantly

from zero in both intervals in a one-sample *t*-test, and the difference in SNARC slopes between both intervals remained non-significant in a *t*-test for two dependent samples. Crucially, the strong conclusion of pure RMdependency that has been derived from these null results is dangerously close to mistaking absence of evidence for evidence of absence. Importantly, no Bayesian analysis was conducted to test whether the null results supported the null hypothesis (and it is not possible to run a post-hoc Bayesian analysis due to the lacking report of the exact *t*-statistic). What is more, neither Dehaene et al. (1993) nor Fias et al. (1996) tested whether the dRT pattern for the same number differed significantly between number ranges – even if the right-hand advantage (reflected by negative dRTs) for numbers 4 and 5 in the range from 0 to 5 and the left-hand advantage (reflected by positive dRTs) for these numbers in the range from 4 to 9 are often cited. Also, the smallest-number intercepts and the mean-number intercepts were not compared between ranges.

Moreover, the design was most likely underpowered for the relevant statistical comparisons in both studies (see below for calculations). On the one hand, this was due to the relatively low sample sizes (n = 12 in Dehaene et al., 1993; and n = 24 in Fias et al., 1996). On the other, only 15 repetitions per experimental cell (i.e., per number magnitude and response-key assignment) were used. Later methodological studies proposed to use at least 20 repetitions and 20 participants to detect the SNARC effect, and even more repetitions and participants to detect differences in the size of the SNARC effect (Cipora & Wood, 2017). Following the *effect-size sensitivity approach* (Giner-Sorolla et al., 2020), we have run power calculations to determine SNARC slope differences between the two number ranges that are detectable in a *t*-test for two dependent samples at different adequate power levels (adapting Monte-Carlo simulations by Wickelmaier, 2022). For the sample size used by Fias et al. (1996) and with the standard deviations they observed, our calculations revealed that at power levels of .80, .90, and .95, only SNARC slope differences between the two number ranges of minimum 11.0 ms (d = 0.60), 12.7 ms (d = 0.69) and 14.1 ms (d = 0.77) could have been detected, respectively.

Note that we ran these calculations within the frequentist framework, which corresponds to the data analysis by Fias et al. (for calculations in both the frequentist and the Bayesian framework, see "PCI Registered Report Materials" at https://doi.org/10.17605/OSF.IO/Z43PM, created using the R packages *rmarkdown* by Allaire et al., 2022; *knitr* by Xie, 2022; and *BayesFactor* by Morey et al., 2015). However, such differences in SNARC slopes are very unlikely, even in case of AMdependency, because they would be larger than the typically observed SNARC slopes themselves. Because of the lack of related information in Dehaene et al.'s (1993) paper, we were not able to run such power calculations for their results; but because their sample was even smaller, they could have detected only even larger differences.

Moreover, the stimuli used in both studies (0, 1, 2, 3, 4, 5 and 4, 5, 6, 7, 8, 9) lead to two problems. First, the average number magnitude in both number ranges is larger for odd than for even numbers (3 vs. 2 in the lower and 7 vs. 6 in the higher number range). This can lead to a confound with the MARC (Linguistic Markedness of Response Codes) effect that denotes a left/right-hand advantage when responding to odd/even numbers, respectively (Nuerk et al., 2004). Such a confound may decrease the SNARC effect (Tzelgov et al., 2013; Zohar-Shai et al., 2017). The association of small/large numbers to the left/right side, respectively, should be weaker if small/large numbers are more often even/odd, respectively. More recent studies have addressed this issue by using stimuli sets in which number magnitude and contrast-coded parity are orthogonal (e.g., Cipora, Soltanlou, et al., 2019). Typically, it is done by using the number set 1, 2, 3, 4, 6, 7, 8, 9, which importantly also excludes 0 (see below).

Second, using the number 0 is problematic due to its special status shown in several studies: Reading time for 0 is significantly longer than for any other single digit number and is not predicted by factors determining reading time of other single digit numbers (Brysbaert, 1995). Nuerk et al. (2004) and Nieder (2016) provide further empirical evidence that 0 may not be represented on the MNL along with other numbers (but see Pinhas & Tzelgov, 2012, for another conclusion). Additionally, quite often participants have problems understanding the

parity status of 0 (Levenson et al., 2007). Using 0 also turned out problematic in SNARC studies: The RTs and dRTs for the number 0 do not strongly correlate with the RTs and dRTs of other even numbers (Nuerk et al., 2004). Later studies on the SNARC effect have excluded 0 from the stimuli set (e.g., Cipora, van Dijck, et al., 2019; Cleland & Bull, 2018; Deng et al., 2016; Gevers et al., 2010, Gökaydin et al., 2018). Ultimately, both the parity status and the presence of 0 might have confounded the results of the previous studies (see Table 1). Therefore, in addition to the replication that we conducted as close as possible to the original studies by Dehaene et al. (1993) and Fias et al. (1996), we also ran a conceptual replication using suitable stimulus sets to disentangle these potential confounds and tackle all the above-mentioned limitations.

The SNARC effect operating on two reference frames at once

As we laid out so far, there is a general tendency to interpret the SNARC effect as entirely flexible based on the findings of RMdependency and on the inference-statistical null effects concerning AMdependency (in underpowered studies). However, the SNARC effect could be operating concurrently in both relative and absolute terms. Indeed, one of us has proposed that the SNARC effect operates on multiple number lines (Weis et al., 2018). However, that paper is not about whether the SNARC effect operates on multiple number lines in terms of RMdependency and AMdependency, but instead it used two-digit numbers as stimuli to see whether separate number lines are activated for decade and unit numbers. The operations on different number ranges are for decade and unit digits of one two-digit number (i.e., the same number, but different digits of its decomposition). Thus, the paper by Weis et al. provides the principal account that the SNARC effect could operate on multiple reference frames at once. The current study goes beyond their findings because it seeks to demonstrate that both RMdependent and AMdependent spatial mappings are concurrently present in the same digit.

The current study

In this study, we aim to answer the question whether the SNARC effect depends only on relative magnitude or whether absolute magnitude plays a role as well. Crucially, in contrast to previous literature about the flexibility of the SNARC effect, we differentiate between two concepts that can be affected by RMdependency and AMdependency:

(i) On the one hand, the number mapping on the MNL (e.g., dRT for number 4) may be different depending on the experimental setup. In our setup, it can be RMdependent (i.e., depending on the position on the used range, e.g., position 5 for range 0 – 5, or 1 for range 4 – 9), AMdependent (i.e., depending on the magnitude, e.g., 4), or both at the same time.

(ii) On the other hand, the strength of the SNARC effect relies on the relative increase of right-hand advantage per increase in magnitude (i.e., the steepness of the SNARC slopes, e.g., -5 ms per number or -10 ms per number) and these slopes can differ between ranges.

For a more detailed and complex elaboration of six possible scenarios combining different parameters of (i) and (ii), see Figures S1 and S2 in the Supplementary Material (see "PCI Registered Report Materials" at https://doi.org/10.17605/OSF.IO/Z43PM) .

To answer the research question, we first replicated Experiment 3 by Dehaene et al. (1993), which has also been replicated in Experiment 1 by Fias et al. (1996), using the original number ranges from 0 to 5 and from 4 to 9. Second, we conducted a conceptual replication to address confounds due to the unequal distribution of odd and even numbers and due to the presence of 0 in both stimuli sets, using the number ranges 1 to 5 (excluding 3) and 4 to 8 (excluding 6). The middle number of the range is also excluded in most SNARC studies using the typical set from 1 to 9. This way, the critical numbers that appear in both ranges were the same in both experiments, namely 4 and 5. Table 1 (see Method section) gives an overview of the used number ranges and of confounds between number parity and number magnitude in Experiment 1 that were avoided in Experiment 2.

In both of our replication experiments, a high statistical power was obtained by testing much larger samples than Dehaene et al. (1993) and Fias et al. (1996) and by increasing the

number of repetitions per experimental cell from 15 to 25. To be able to quantify evidence both for differences between number ranges and lack thereof, we applied the Bayesian instead of frequentist approach in statistical analysis. For the interpretation of different values for the Bayes Factors, we followed the recommendations by Dienes (2021): A BF_{10} greater than 3 or 10 was treated as moderate or strong evidence for the alternative hypothesis, while a BF_{10} smaller than 1/3 or 1/10 was treated as moderate or strong evidence for the null hypothesis, respectively.

Online experiments offer the possibility to collect data from large samples and therefore reach high statistical power (Reips, 2000, 2002). The SNARC effect has been successfully replicated in online settings (e.g., Cipora, Soltanlou, et al., 2019; Gökaydin et al., 2018; Koch et al., 2023). The measurement in the online setup showed a similar reliability and magnitude compared to the SNARC effect that is typically observed in lab studies. Further, it seems to be valid as regards the correlations of the SNARC effect with mean RTs and standard deviations of RTs, which are similar compared to lab studies.

In this study, we expected to replicate the findings by Dehaene et al. (1993) and by Fias et al. (1996) as concerns RMdependency. However, we also expected to find evidence towards AMdependency of the number mapping on the MNL and of the strength of the SNARC effect. Previous studies have indicated tendencies that cannot be explained by RMdependency alone. Thus, we hypothesized:

1. A SNARC effect in both (a) the lower and (b) the higher number ranges in each experiment. (a) The SNARC effect in the lower range served as a manipulation check and was considered as a prerequisite for testing Hypotheses 2 and 3 in each experiment. Both (a) and (b) aimed at replicating the results by Dehaene et al. (1993) and Fias et al. (1996).
2. Both (a) RMdependency and (b) AMdependency of the number mapping on the MNL, such that small/large numbers in relative and absolute terms are shifted towards the

left/right, respectively. (a) RMdependency is reflected by dRTs for the same critical numbers (i.e., 4 and 5) differing between ranges, showing that the MNL adapts flexibly and relative to the range. (b) AMdependency is reflected by dRTs for these critical numbers being equal between ranges, and by dRTs for the smallest number (Experiment 1: 0 in the 0 – 5 range vs. 4 in the 4 – 9 range; Experiment 2: 1 in the 1 – 5 range [excluding 3] vs. 4 in the 4 – 8 range [excluding 6]) differing between ranges. AMdependency means that small/large numbers are shifted to the left/right on the MNL, although they are exactly on the same position within their range, but differ in terms of absolute magnitude.

3. AMdependency of the strength of the SNARC effect, such that it is stronger in the lower than in the higher ranges. This is reflected by steeper (i.e., more negative) SNARC slopes in the lower than in the higher ranges, which was descriptively observed in the two seminal studies by Dehaene et al. (1993) and Fias et al. (1996).

Method

This study has been approved by the ethics committee of the University of Tübingen's Department of Psychology. The Stage-1 Registered Report received in-principle acceptance by the Peer Community In after peer review on December 3rd 2023 (https://doi.org/10.17605/OSF.IO/AE2C8).

Sample size considerations

For this study, we defined Cohen's d = 0.15 as the minimal effect size of interest, because the most crucial aim of the present study was to find out whether AMdependency of the strength of the SNARC effect exists or not (Hypothesis 3). By choosing this minimal effect size of interest, we were able to find evidence for or against the SNARC slope differences between number ranges that had been descriptively reported in the original studies that we wished to replicate. Due to the lacking report of standard deviations, it was not possible to

calculate Cohen's *d* for the slope difference of 9.2 ms found by Dehaene et al. (1993), but the slope difference of 2.99 ms with its standard deviation of 18.34 ms found by Fias et al. (1996) corresponds to an effect size of $d = 0.16$. Note that in the two original studies, the symmetric confidence intervals for these estimates must also include at least the double slope difference and effect size due to their non-significance. Hence, in case of AMdependency of the strength of the SNARC effect, the true effect size might in fact be larger than $d = 0.15$. This sample size estimation was also valid for testing Hypotheses 1 and 2, which required one-sample *t*-tests. The reason was that an effect smaller than $d = 0.15$ would not be meaningful for the SNARC effect in the lower (Hypothesis 1a) or higher (Hypothesis 1b) number range, or for RMdependency (Hypothesis 2a) and AMdependency (Hypothesis 2b) of the number mapping on the MNL either. Similarly, the chosen maximal sample size was large enough to find at least moderate evidence in case these hypotheses are false.

To ensure a probability of .90 for finding at least moderate evidence for a true underlying effect (i.e., BF_{10} above 3, according to Dienes, 2021) with a minimally relevant effect size of Cohen's $d = 0.15$ in one-sample or paired *t*-tests, the sample needed to consist of 800 participants in each experiment (for power calculations, see "PCI Registered Report Materials" at https://doi.org/10.17605/OSF.IO/Z43PM). The sample size of 800 participants was required for a proportion of at least .90 Bayesian *t*-tests to yield a BF_{10} above 3, when 5000 samples of SNARC slope differences randomly drawn from a normal distribution around the minimally relevant effect size of $d = 0.15$ were simulated (for a similar approach, see Kelter, 2021). Following the same procedure, we found that the sample needed to consist of 180 participants to ensure a probability of .90 for finding at least moderate evidence against a truly absent effect (i.e., BF_{10} below 1/3 for $d = 0$, according to Dienes, 2021). Note that the sample size of 180 was smaller than the initial sample size of 200 that was collected in the "Sequential Bayes Factor with maximal n" (SBF+maxN) approach as described by Schönbrodt & Wagenmakers (2018; see explanation below). For these calculations, we used $SD = 15.1$ ms and $SD = 11.2$ ms for the

lower and higher number ranges, as reported by Fias et al. (1996), although the standard deviations in our previous color judgment experiments were only *SD* = 4.2 ms and *SD* = 3.9 ms (Roth, Caffier, et al., 2024). Hence, our calculations were rather conservative, and the probability to find evidence for a true underlying effect thus was most probably even higher. While in the frequentist framework, low error type II rates (and high statistical power) need to be achieved, in the Bayesian framework, a low probability of misleading evidence for the null hypothesis in case of a true underlying effect and a high probability of finding evidence for a true underlying effect need to be ensured. To achieve the same probability for error type II and misleading evidence, Bayesian *t*-tests (using the default *r*-scale of 0.707 as uninformed prior in the Cauchy distribution) require larger samples as compared to frequentist *t*-tests (Kelter, 2021).

Importantly, as we run Bayesian instead of frequentist analyses, we made use of the SBF+maxN approach and defined an optional stopping threshold to make our data collection more efficient. Namely, we used moderate evidence in favor of all hypotheses ($BF_{10} > 3$) or against them ($BF_{10} < 1/3$) as thresholds. More precisely, for each experiment, we first recruited 200 participants (i.e., complete individual datasets) and computed the BF_{10} for the SNARC effect in lower (Hypothesis 1a) and higher (Hypothesis 1b) number ranges, for the shift of critical small/large numbers in both relative (Hypothesis 2a) and absolute (Hypothesis 2b) terms towards the left/right, respectively, and for the SNARC slope difference between ranges (Hypothesis 3). As long as the BF_{10} did not reach any of the two thresholds for all hypotheses, we collected another 20 complete individual datasets and recalculated the BF_{10}. If no threshold had been reached with our maximal sample size of 800 participants (that is required for obtaining at least moderate evidence for a true underlying minimally relevant effect with a probability of at least .90, as explained above), we would have stopped the sequential recruiting of participants in any case.

Participants

For each experiment, adults were recruited via the recruiting platform Prolific. To comply with our ethics proposal, they had to be at least 18 years old, and because of possible age differences in RTs, we set the maximum age to 40 years. As the experiments were conducted in English, participation was only possible for native English speakers (as per Prolific's screening based on self-reports). Participation took approximately 20 minutes and was compensated with £5 (partial payment for partial participation).

Design and experimental task

In the parity judgment task with binary response-key setup, participants had to indicate as fast and as accurately as possible whether the number presented on the screen was odd or even. The parity judgment task is widely used in numerical cognition and the standard task to investigate the SNARC effect (see Toomarian & Hubbard, 2018, for a review, and Wood et al., 2008, for a meta-analysis). We assigned participants randomly to one of our two experiments. In Experiment 1 (close replication of Dehaene et al., 1993, and Fias et al., 1996), the numbers from 0 to 5 were used in the lower number range and the numbers from 4 to 9 in the higher number range. In Experiment 2 (conceptual replication), the numbers from 1 to 5 (excluding 3) were used in the lower number range and the numbers from 4 to 8 (excluding 6) in the higher number range, eliminating confounds between number parity and number magnitude (see Table 1) and special influences of number 0.

In both experiments, we used 25 repetitions per number magnitude in each number range (lower vs. higher) and each response-key assignment (MARC congruent, i.e., left-hand responses to odd and right-hand responses to even numbers, vs. MARC incongruent, i.e., right-hand responses to odd and left-hand responses to even numbers). This lead to a total of 600 trials for Experiment 1 and 400 trials for Experiment 2. In each experiment, the trials were equally divided into four blocks (one per combination of number range and response-key assignment), with a break of minimum 30 seconds between them. Participants were randomly assigned to one of four block orders of congruency and number range (see Figure 1). The order

of stimulus presentation within blocks was fully randomized. Each trial started with a square (extended ASCII 254 with the font size 72px) as an eye fixation point (300 ms). Then the number (Open Sans font, size 72px) was presented until a response was given. A blank screen (500 ms) concluded the trial. Stimuli as well as fixation squares were presented in black (0, 0, 0 in RGB notation), while the background remained gray (150, 150, 150 in RGB notation). The time course of an exemplary trial is illustrated in Figure 2. Each block was preceded by a short practice session, in which each number was presented twice (i.e., 12 practice trials before each block in Experiment 1 and eight practice trials before each block in Experiment 2). Accuracy feedback appeared during practice sessions only.

Table 1

Stimulus sets and their characteristics

Experiment 1 (close replication: number ranges used by Dehaene et al., 1993, and Fias et al., 1996)				Experiment 2 (conceptual replication)			
Lower range		Higher range		Lower range		Higher range	
Absolute magnitude predictor	Contrast-coded parity predictor	Absolute magnitude predictor	Contrast-coded parity predictor	Absolute magnitude predictor	Contrast-coded parity predictor	Absolute magnitude predictor	Contrast-coded parity predictor
0	+0.5	**4**	+0.5	1	–0.5	**4**	+0.5
1	–0.5	**5**	–0.5	2	+0.5	**5**	–0.5
2	+0.5	6	+0.5	**4**	+0.5	7	–0.5
3	–0.5	7	–0.5	**5**	–0.5	8	+0.5
4	+0.5	8	+0.5				
5	–0.5	9	–0.5				
Mean number magnitude depending on number parity:							
M_{even} = 2 M_{odd} = 3		M_{even} = 6 M_{odd} = 7		M_{even} = 3 M_{odd} = 3		M_{even} = 6 M_{odd} = 6	
Correlation between number magnitude and number parity:							
r = –.293				r = 0			

Note. This table lists all stimuli used in the two experiments. It shows the confound between number parity and number magnitude in both number ranges of Experiment 1 and illustrates how we avoided it in both number ranges of Experiment 2, such that number parity and number magnitude were uncorrelated (i.e., they were orthogonal to each other as predictors in regression models). Number parity was contrast-coded with –0.5 for odd and +0.5 for even numbers when measuring the MARC effect. The number 0 was included in Experiment 1, but not used it in the conceptual replication in Experiment 2 because of its special features and irregular mental representation (as outlined in the Introduction). The numbers 4 and 5, which are written in bold in the table, were present in each of the number ranges.

Figure 1

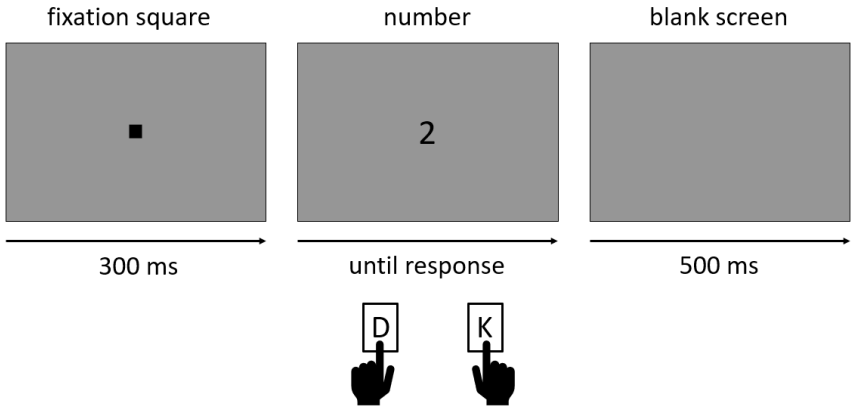
Counterbalancing block orders in Experiments 1 and 2

	Block order 1	Block order 2	Block order 3	Block order 4
Block 1	lower range MARC incongruent	lower range MARC congruent	higher range MARC incongruent	higher range MARC congruent
Block 2	lower range MARC congruent	lower range MARC incongruent	higher range MARC congruent	higher range MARC incongruent
Block 3	higher range MARC incongruent	higher range MARC congruent	lower range MARC incongruent	lower range MARC congruent
Block 4	higher range MARC congruent	higher range MARC incongruent	lower range MARC congruent	lower range MARC incongruent

Note. This figure shows the four block orders resulting from the combination of range (lower range vs. higher range) and response-key assignment (MARC congruent, i.e., odd-left and even-right, vs. MARC incongruent, i.e., even-left and odd-right). Each block was preceded by two repetitions per number as practice trials (12 trials for Experiment 1 and eight trials in Experiment 2), consisted of 25 repetitions per number as experimental trials (150 trials for Experiment 1 and 100 trials in Experiment 2) and was followed by a break.

Figure 2

Time course of an exemplary trial



Procedure

The experiments were set up with WEXTOR (https://wextor.eu; Reips & Neuhaus, 2002) in its HTML and JavaScript framework and adapted (see demo version for Experiment 1 at https://luk.uni-konstanz.de/numcog_3/?demo&e1 and for Experiment 2 at https://luk.uni-konstanz.de/numcog_3/?demo&e2). Our previous experiments (Roth, Caffier, et al., 2024) have demonstrated that this software is suitable for detecting the SNARC effect in an online setup. At the very beginning of the experiment, a seriousness check (e.g., Reips, 2009; Aust et al., 2013) was applied and participants were asked whether they wanted to participate seriously. Participants were asked to take part only if they gave their informed consent, if they were using a desktop computer or laptop, and if they were between 18 and 40 years old. Then, participants were asked to provide basic demographic data such as age, gender (*man*, *woman*, *other*), first native language (English and potentially others), handedness (*right-handed*, *left-handed*, *ambidextrous*), and finger-counting habits (starting hand: *left hand*, *right hand*, *does not know or no preference*; and stability: *always*, *usually*, *does not know or no preference*). For each of the above-mentioned questions, we also provided the option "I prefer not to answer" to respect some participants' unwillingness to share information with us and to not force them to choose any option that might not reflect the truth (Jenadeleh et al., 2023; Stieger et al., 2007). Note that in earlier studies, only very few participants chose this option in any of the above-mentioned questions. Next, if not already the case for the default response keys D and K, participants could

choose response keys for the experimental task which were to be located in the same row and about one hand width apart from each other on their keyboard. Then, instructions were displayed before the first block of the experimental task started with its practice trials. For instance, the instructions were as follows for the block with the lower number range in Experiment 1 (only numbers and response-to-key assignments are replaced for the higher number range or for Experiment 2): “In our experiment, your task is to distinguish the parity of numbers, that is, to decide whether a number is even or odd. For this, please place the index finger of your left hand on the [D] key and the index finger of your right hand on the [K] key on your keyboard. In each run, a black square will appear in the center of the screen. Please look at this square. It will soon be replaced by either an even or an odd number. If the number is even (0, 2, 4), press [D]. If the number is odd (1, 3, 5), press [K]. Please answer as quickly and as accurately as possible.”

After completion of the whole experimental task, participants were asked to self-rate their math skills compared to people of their age on a visual analogue scale from *very bad* to *very good* operationalized as 0 to 400 (Funke & Reips, 2012). Next, data quality was assessed by asking participants how they would describe their environment during participation (*silent*, *very quiet*, *fairly quiet*, *fairly noisy*, *very noisy*, or *extremely noisy)*, whether there were any major distractions during participation (*none*, *one*, or *multiple*), and whether there were any difficulties during participation (*yes* or *no*, text field for comments). Moreover, we asked participants whether they had used their left index finger for the left response key and their right index finger for the right response key throughout the experiment (*yes*, *partly*, or *no*). Participants were provided a completion code for Prolific and contact information of our research team. To prevent search engine bots (e.g., Googlebot) from submitting data on our experiment, we equipped the experiment materials with a standardized "noindex, nofollow" meta tag, which prompts search engine bots not to index the experiment pages and also not to visit subsequent pages (see Reips, 2007, p. 379). Further, we restricted participation to devices

over 600 pixel screen width (i.e., no smartphones). In addition, to exclude multiple submissions by the same participants, submissions from the same IP addresses were not permitted.

Data preprocessing

We used the same analysis pipeline as in another of our studies, except for not applying any color vision check (Roth, Caffier, et al., 2024). This pipeline is similar to that used by Cipora, van Dijck, et al. (2019) in an extensive re-analysis of 18 datasets and permits to reliably detect the SNARC effect. The preprocessing steps were applied in the exact order as they are listed in the following. Specifically, only datasets of participants who indicated to be between 18 and 40 years old and to seriously participate were included. Individual datasets were excluded if participants described their environment as very/extremely noisy, if they reported multiple major distractions, or if they reported that they were not using their left/right index finger for the left/right response key, respectively. As outlined by Reips (2002) and Birnbaum (2004), experimenters are recommended to control for potential multiple submissions from the same participants by checking their User-Agents (OS and browser information) and IP addresses[1]. Regarding the data from all remaining participants, practice trials and incorrectly answered trials were not analyzed. Only trials with RTs of minimum 200 ms were included in the analysis, because parity judgments faster than 200 ms are very unlikely and faster responses can therefore be treated as anticipations. Moreover, only trials with RTs of maximum 1500 ms were included, because healthy educated adults should be capable to judge the parity status of single-digit numbers in less than 1500 ms, so that slower responses are unlikely to reflect only the mental process underlying parity judgment but instead might be caused by distractions. In a next step, further outliers were removed in an iterative trimming procedure for each

[1] Note that Prolific claims that only one participation from each household is allowed per study. Nevertheless, we received two different complete individual datasets with the same User-Agent, IP address, and Prolific ID. The analyses reported in the current paper are based on both datasets. However, we reran the analyses while excluding the second of these two individual datasets and also both of these individual datasets. Importantly, apart from small changes in the decimals of the obtained Bayes Factors, the results and interpretations did not change substantially in either of the two reanalyses.

participant separately, such that only RTs that are maximum 3 SDs above or below the individual mean RT of all remaining trials were considered. This procedure permitted to exclude RTs that were unlikely for each given participant and accounts for the right-skewed distribution of RTs, where the means would otherwise have been largely overestimated. After these exclusions at the trial level, only data of participants with at least 75% valid remaining trials (after excluding errors and outlier RTs) were included in the analysis at the participant level. Finally, only datasets of participants without any empty experimental cell (number magnitude per response side) in both number ranges were considered, because an empty cell would have caused a missing dRT, which in turn would have made the calculation of the SNARC slope problematic. Only complete individual datasets were included in the analysis (and none of the incomplete individual datasets fulfilled the inclusion criteria listed above).

Data analysis

All data analyses were performed in the statistical computing software R (Version 4.3.3; R Core Team, 2024), using the R packages *BayesFactor* (Morey et al., 2024) *data.table* (Barrett et al., 2024), *dplyr* (Wickham, François, et al., 2023), *GeneNet* (Schaefer et al., 2021), *ggplot2* (Wickham et al., 2024), *neatStats* (Lukács et al., 2022), *plyr* (Wickham, 2023), *tidyr* (Wickham, Vaughan, et al., 2023). An overview of all research questions with corresponding hypotheses, the targeted sample size and planned analyses with a rationale, as well as the interpretations of potential outcomes and theoretical conclusions is given in the Study Design Table (see Table 5). Instead of frequentist analysis, we decided to take the Bayesian approach. For this, we determined the BF_{10} associated with the corresponding Bayesian *t*-test to obtain evidence for both null and alternative hypotheses (using a default *r*-scale of 0.707 as uninformed prior using Cauchy distribution). More specifically, we calculated Bayesian *t*-tests and extracted the respective BF_{10}. Considering a BF_{10} larger than 3 as evidence against the null hypothesis is less likely than rejecting a null hypothesis with a conventional significance level of $\alpha = .05$ in the frequentist approach (Wetzels et al., 2011). As explained above, we applied the SBF+maxN

approach for sequential data analysis with optional stopping in case of at least moderate evidence for or against all hypotheses.

The key dependent variable was the mean difference between RTs of the right hand minus left hand (dRT), which was calculated for each number separately per participant and per number range. RTs were measured as the time from the onset of the number presentation on the screen until the participant's response. A potential SNARC effect can be determined by regressing dRTs on the number magnitude (Fias et al., 1996). For each participant and for each number range one regression was calculated. Our first dependent measure were SNARC slopes resulting from the regression of dRTs on number magnitude, which represent the change in relative advantage of right-hand compared to left-hand responses in ms per increase by one in the number magnitude (the more negative the slope, the stronger the SNARC effect). Moreover, we calculated smallest-number intercepts (when relative magnitude of the numbers in both ranges was matched, i.e., predicted dRTs for 0 in the 0 – 5 range vs. 4 in the 4 – 9 range in Experiment 1, and 1 in the 1 – 5 range [excluding 3] vs. 4 in the 4 – 8 range [excluding 6] in Experiment 2) as well as dRTs for critical numbers that were part of both number ranges (i.e., 4 and 5). An overview of how the tests described in the following helped us distinguish the six scenarios with different number representation shapes, depending on the number mapping on the MNL and the strength of the SNARC effect, is given in Figures S1 and S2 and Table S1 (see Supplementary Material).

First, we tested the presence of the SNARC effect on group level in both number ranges separately in each experiment (Hypothesis 1). As described in the Introduction, the SNARC effect seems to be stronger in the lower than in the higher number range, resulting in a more negative slope. As the SNARC effect is very robust especially for lower ranges and possibly stronger than in higher ranges (see Hypothesis 3), the SNARC effect in lower ranges (Hypothesis 1a) was a manipulation check and prerequisite for following investigations (Hypotheses 1b, 2 and 3). The obtained SNARC slopes were tested against zero with two-sided

Bayesian one-sample *t*-tests in each number range in each experiment. This procedure corresponds to the repeated-measures regressions described by Lorch and Myers (1990) and applied to the SNARC effect by Fias et al. (1996). It accounts for the within-subject design, where each participant completes trials for each digit in each response-to-key assignment. Although we did not expect the SNARC effect to be reversed at the group level, we preregistered two-sided tests here to stay consistent within this study. Evidence for the SNARC effect in all ranges would replicate findings from the two studies by Dehaene et al. (1993) and Fias et al. (1996). The lack of conclusive evidence as regards the SNARC effect in the lower ranges (Hypothesis 1a) with our maximal sample of 800 participants or even evidence against it was highly unlikely in our view. Evidence against the SNARC effect in the higher ranges (Hypothesis 1b) combined with evidence for the SNARC effect in the lower ranges (Hypothesis 1a) would provide support for AMdependency of the strength of the SNARC effect (Hypothesis 3).

Second, to investigate RMdependency of the number mapping on the MNL, we tested whether dRTs for critical numbers (i.e., 4 and 5) differed between the lower and the higher number range (Hypothesis 2a) with one two-sided paired Bayesian *t*-test per number in each experiment. Evidence for a difference would imply that the SNARC effect and the MNL are (at least partly) flexible and adapt to the number range used in a task (as in Scenarios 1, 2, 4, and 5 in Figures S1 and S2 in the Supplementary Material). This would be in line with the literature claiming that numbers 4 and 5 are associated with the right side in the number range from 0 to 5 and with the left side in the number range from 4 to 9. However, this finding would not fully rule out AMdependency. Evidence against a difference would indicate that the SNARC effect and the MNL are AMdependent at least to some degree (as in Scenarios 3 and 6 in Figures S1 and S2).

Next, to test AMdependency of the number mapping on the MNL, we tested whether the smallest-number intercepts differed between the lower and the higher number range

(Hypothesis 2b) with one two-sided paired Bayesian *t*-test in each experiment. Evidence for a difference would lead to the conclusion that small/large numbers are overall shifted to the left/right on the MNL, respectively (as in Scenarios 2, 3, 5, and 6 in Figures S1 and S2). In other words, this would imply that the SNARC effect and the MNL are not fully RMdependent. Evidence against a difference would indicate that the SNARC effect and the MNL are at least partly RMdependent (as in Scenarios 1 and 4 in Figures S1 and S2).

Third, to investigate AMdependency of the strength of the SNARC effect, we compared SNARC slopes between the number ranges (Hypothesis 3) with one two-sided paired Bayesian *t*-test in each experiment. Evidence for steeper SNARC slopes in the lower than in the higher number range can be interpreted as stronger SNARC effect within (in absolute terms) smaller than larger numbers (as in Scenarios 4, 5, and 6 in Figures S1 and S2). This result would lead to the conclusion that the spatial mental representation seems to be more pronounced for small than for large numbers. Evidence against such a difference would indicate that the strength of the SNARC effect does not differ between number ranges (as in Scenarios 1, 2, and 3 in Figures S1 and S2). Once the data was collected, results could be interpreted with the help of Table S1 in the Supplementary Material to see which scenario most likely underlies the mental representation of number magnitude.

Additionally to all effects reported in the unit of interest, we provide effect sizes in terms of Cohen's *d* for all Bayesian *t*-tests. Effects of $d \geq 0.2$, $d \geq 0.5$, or $d \geq 0.8$ will be interpreted as small, medium, or large effect sizes, respectively.

Manipulation checks

To control the data quality in our study, we implemented a seriousness check (Aust et al., 2013; Reips, 2009; review in Reips, 2021) as well as a self-assessment of noise, distractions, and other difficulties. To make sure that we only analyzed trials that reflected mental processes in correctly executed parity judgment, we excluded incorrectly answered trials and trimmed RTs (as described in the data preprocessing pipeline in our Stage-1 Registered Report). Also,

we excluded data of participants with less than 75% valid trials to only build our results on participants who understood and followed the task instructions. Moreover, we assessed whether participants complied with the instructions to use their left and right index fingers for the left and right response keys, respectively, and only included their datasets into our analysis if they did so. Then, as a manipulation check, we tested the SNARC effect in the lower number ranges (Hypothesis 1a). Importantly, we only proceeded with the testing of other hypotheses because we could find the SNARC effect in the lower number range in both experiments.

Possible limitations and unexpected outcomes

Finding evidence against the SNARC effect in the lower number ranges (Experiment 1: 0 to 5; Experiment 2: 1 to 5 [excluding 3]) would have been an unexpected outcome. The SNARC effect in the parity judgment task has been shown in plenty of studies (including online experiments) using different number ranges within the interval from 0 to 9. Our large sample and a high number of repetitions ensured a high probability to find evidence even for small effects. The presence of the SNARC effect in the lower ranges thus is a manipulation check and prerequisite for further hypothesis tests.

Even though our Experiment 1 was a direct replication of Dehaene et al.'s (1993) and Fias et al.'s (1996) study, we decided to use 25 instead of 15 repetitions per experimental cell. First, we thereby increased statistical power and measurement precision (Luck, 2019); second, we followed methodological recommendations for investigating the SNARC effect (Cipora & Wood, 2017); and third, we ensured the comparability with our conceptual replication in Experiment 2. However, because of this methodological improvement, our experiment was therefore strictly speaking not a direct replication.

Just as the original two experiments, Experiment 1 had the limitation of the MARC effect being confounded with the SNARC effect because number parity and number magnitude were no orthogonal predictors in the regression model. Therefore, we only calculated the

MARC effect for the data resulting from Experiment 2. Moreover, because of the special characteristics of the number 0 regarding its mental representation, including it in the stimulus set might have driven responses in our Experiment 1. However, we tackled these limitations in Experiment 2 by using another stimulus set.

Results

The data collection as well as the confirmatory data analyses were conducted as described in the peer-reviewed Stage-1 Registered Report, which received in-principle acceptance by the Peer Community In (PCI) on December 3rd, 2023 (https://doi.org/10.17605/OSF.IO/AE2C8). Additional data analyses are referred to as exploratory data analyses in the following. All R scripts for data preprocessing and analysis as well as all anonymized datasets can be found at https://doi.org/10.17605/OSF.IO/Z43PM. A Study Design Table was filled in prior to data collection and provides an overview of all research questions, corresponding hypotheses, the targeted sample size and planned analyses with a rationale, the interpretations of potential outcomes and theoretical conclusions (see Table 5). It also contains the observed outcomes for both experiments, which are additionally illustrated in Figure 5.

Experiment 1: Close replication with 0 to 5 vs. 4 to 9

Data preprocessing

Our final recruited sample size according to the SBF+maxN approach was 200. This final sample size refers only to collected individual datasets that were complete, but initially 208 individuals started participation of which 8 did not complete. Of these, one participant was under 18 and two participants were over 40 years old. Three participants did not fully follow the instructions as concerns response-to-key assignment, and seven participants used other fingers than required in the instructions or switched around. The data of these participants were excluded from the analysis in a first preprocessing step. All recruited individuals wanted to seriously participate (i.e., none of them only wanted to look at the experiment), and none of them reported a very/extremely noisy environment or multiple major distractions. In the next preprocessing step, exclusion criteria were applied at the trial level. That is, 0.01% of the trials were excluded due to missing responses, 5.78% due to incorrect responses, 0.12% due to responses faster than 200 ms, 2.17% due to responses slower than 1500 ms, and 6.14% in the

sequential RT trimming procedure. After these exclusions at the trial level, none of the participants had any empty experimental cell, but 14 participants had less than 75% remaining valid trials and their data were thus excluded from analyses. The remaining data of 173 individuals were analyzed. Descriptive self-reported information about these participants is summarized in Table 2. All of them used the default response keys D and K. The average RT per number in each range can be found in Table 3 and are illustrated in Figures S3 and S4 in the Supplementary Material. Mean RTs ranged from 525.25 ms to 567.35 ms and standard errors (SE) ranged from 6.75 ms to 8.43 ms (note that the descriptively largest SE was observed for number 0, which is in line with previous studies).

Confirmatory data analysis

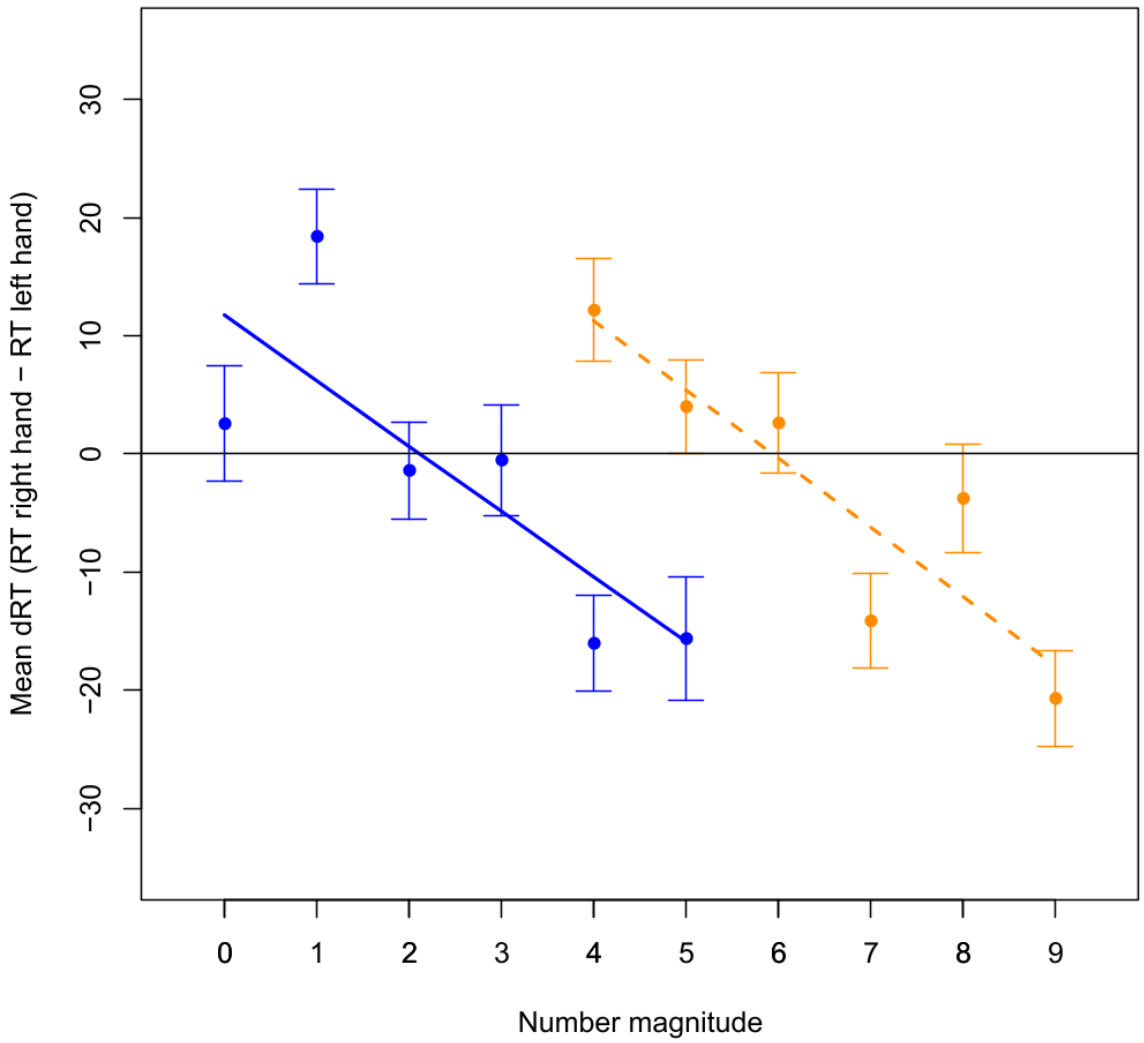
All parameter estimates can be found in Table 4[2]. The average RT for all included trials was 543.44 ms (*SD* = 88.00 ms). The repeated-measures regressions followed by Bayesian one-sample *t*-tests against zero revealed strong evidence for a SNARC effect in the lower range (i.e., for Hypothesis 1a) with BF_{10} = 6956.04 and a slope estimate of -5.53 ms (*SD* = 14.66 ms, *d* = 0.38), as well as in the higher range (i.e., for Hypothesis 1b) with $BF_{10} = 2.63 * 10^6$ and a slope estimate of -5.84 ms (*SD* = 12.33 ms, *d* = 0.47). Hence, the manipulation check confirmed that the manipulation had worked, and the prerequisite for testing all further hypotheses was fulfilled. The SNARC effect is plotted in Figure 3 for both ranges (see also Figure 5, left panel) and looks similar to Scenario 1 illustrated in Figure S1 (see Supplementary Material in "PCI Registered Report Materials" at https://doi.org/10.17605/OSF.IO/Z43PM). Two paired

[2] When excluding number 0 from the lower and number 4 from the higher range in the analyses, the results of the confirmatory data analysis did not change substantially for Hypothesis 1a with $BF_{10} = 2.09 * 10^9$ and for Hypothesis 1b with $BF_{10} = 6.65 * 10^5$, and naturally remained the same for the dRT of number 5 in Hypothesis 2a. However, for Hypotheses 2b and 3, the exclusion of number 0 from the lower and number 4 from the higher range led to inconclusive evidence regarding AMdependency of the number mapping on the MNL and of the strength of the SNARC effect, with BF_{10} = 2.56, and BF_{10} = 0.50, respectively. This stands in contrast with moderate evidence against AMdependency found when including number 0 in the lower and number 4 in the higher range.

Bayesian *t*-tests revealed strong evidence for differences in dRTs for critical numbers between the ranges and thus for RMdependency of the number mapping on the MNL (i.e., for Hypothesis 2a) with BF_{10} = 2.57 * 10^4 (d = 0.40) for number 4 and BF_{10} = 12.86 (d = 0.25) for number 5. Another paired Bayesian *t*-test revealed strong evidence against a difference in smallest-number intercepts between the ranges and thus against AMdependency of the number mapping on the MNL (i.e., against Hypothesis 2b) with BF_{10} = 0.09. The last paired Bayesian *t*-test revealed strong evidence against a difference in SNARC slopes between the ranges and thus against AMdependency of the strength of the SNARC effect (i.e., against Hypothesis 3) with BF_{10} = 0.09.

Figure 3

Mean dRTs per number averaged across all trials from all participants in the lower (blue, solid line) and higher (orange, dashed line) number ranges for Experiment 1, with error bars representing ± 1 standard error for the respective number and regression lines representing slope estimates for the respective range.



Exploratory data analysis

In addition to the confirmatory data analyses and in order to disentangle the possible scenarios illustrated in Figures S1 and S2 and Table S1 (see Supplementary Material), the mean-number intercepts were compared between ranges. The mean number in the lower range (0 to 5) was 2.5 with a dRT estimate of -2.09 ms, and the mean number in the higher range (4 to 9) was 6.5 with a dRT estimate of -3.29 ms. A two-sided paired Bayesian *t*-test revealed moderate evidence against a difference in mean-number intercepts between the ranges and thus against AMdependency of the number mapping on the MNL with $BF_{10} = 0.10$.

Moreover, we tested whether there was a correlation between the SNARC slopes in the lower and the higher range. The data revealed moderate evidence against a correlation with $BF_{10} = 0.19$.

Experiment 2: Conceptual replication with 1 to 5 (excluding 3) vs. 4 to 8 (excluding 6)

Data preprocessing

Our final recruited sample size according to the SBF+maxN approach was 300[3]. This final sample size refers only to collected individual datasets that were complete, but initially 310 individuals started participation of which 10 did not complete. Of these, two individuals did not want to seriously participate and instead only look at the experiment, one was over 40 years old, two participants reported a very/extremely noisy environment, and three participants reported multiple major distractions. Further, 17 participants did not fully follow the instructions as concerns response-to-key assignment, and 10 participants used other fingers than required in the instructions or switched around. The data of these participants was excluded from the analysis in a first preprocessing step. No participant was under 18 and needed to be removed. In the next preprocessing step, exclusion criteria were applied at the trial level. 0.37% of the trials were excluded due to missing responses, 5.52% due to incorrect responses, 0.47% due to responses faster than 200 ms, 1.22% due to responses slower than 1500 ms, and 5.49% in the sequential RT trimming procedure. After these exclusions at the trial level, none of the participants had any empty experimental cell, but 15 participants had less than 75% remaining valid trials and their data was thus excluded from analyses. The remaining data of 255 individuals were analyzed. Descriptive self-reported information about these participants is

[3] Note that due to a computing mistake, data collection was only stopped with 300 participants, although the stopping criterion for the SBF+maxN procedure (i.e., at least moderate Bayesian evidence for or against each hypothesis) was already reached with 200 participants. Importantly, using data from 300 instead of only 200 participants did not change any results substantially and mostly made the Bayesian evidence stronger (see next footnote).

summarized in Table 2. All participants used the default response keys D and K. The average RT per number in each range can be found in Table 3 and are illustrated in Figures S5 and S6 in the Supplementary Material. Mean RTs ranged from 501.59 ms to 522.44 ms and SEs ranged from 5.07 ms to 5.31 ms.

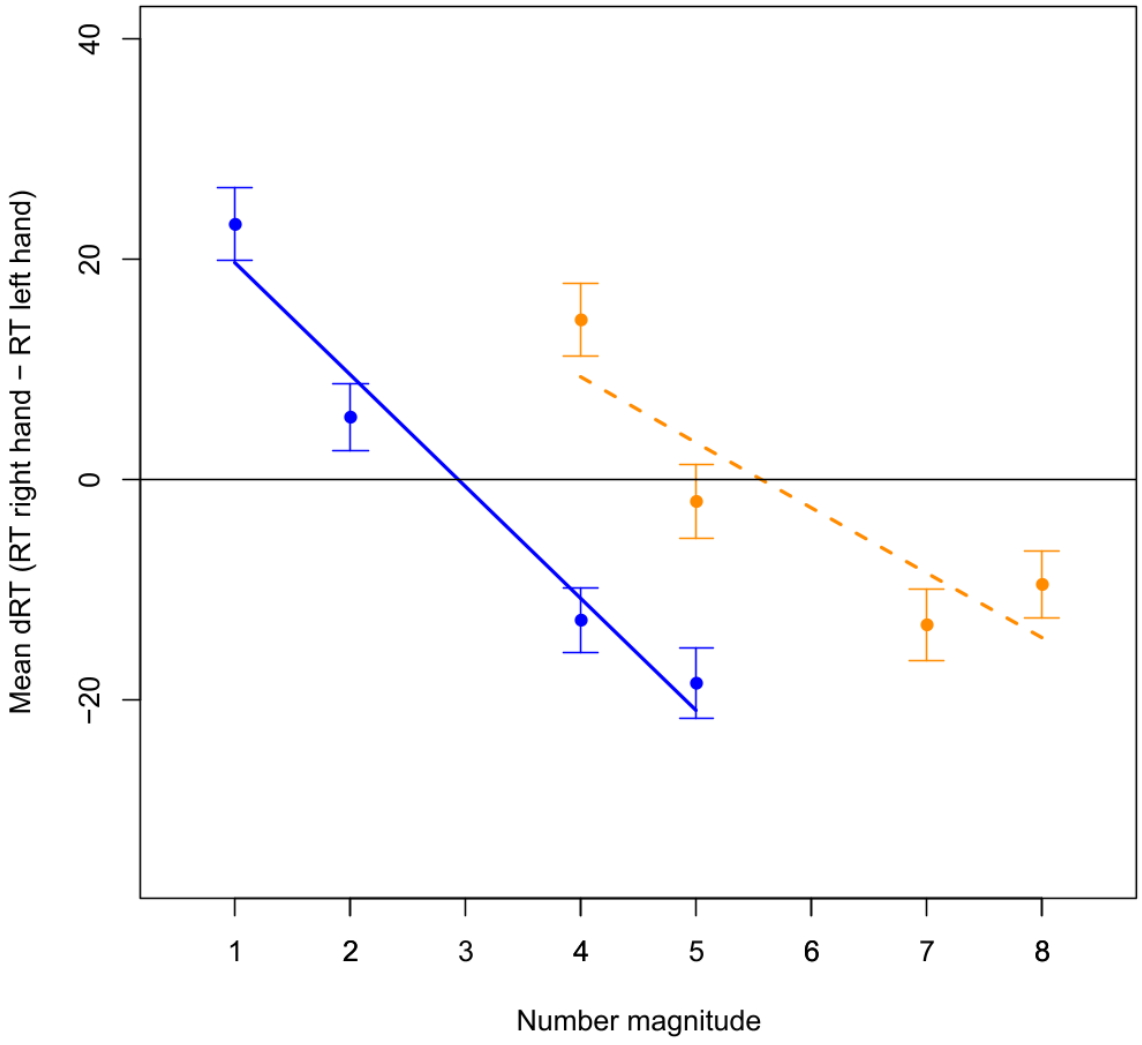
Confirmatory data analysis

All parameter estimates can be found in Table 4. The average RT for all included trials was 512.77 ms (*SD* = 74.46 ms). The repeated-measures regressions followed by Bayesian one-sample *t*-tests against zero revealed strong evidence for a SNARC effect in the lower range (i.e., for Hypothesis 1a) with $BF_{10} = 1.61 * 10^{21}$ and a slope estimate of -10.17 ms (*SD* = 14.32 ms, *d* = 0.71), as well as in the higher range (i.e., for Hypothesis 1b) with $BF_{10} = 1.38 * 10^{12}$ and a slope estimate of -5.92 ms (*SD* = 14.32 ms, *d* = 0.52). Hence, the manipulation check worked (i.e., evidence for Hypothesis 1a was obtained), and the prerequisite for testing all further hypotheses was fulfilled. The SNARC effect is plotted in Figure 4 for both ranges (see also Figure 5, right panel) and seems to correspond to Scenario 5 presented in Figure S2. Two paired Bayesian *t*-tests revealed strong evidence for differences in dRTs for critical numbers between the ranges and thus for RMdependency of the number mapping on the MNL (i.e., for Hypothesis 2a) with $BF_{10} = 6.64 * 10^{7}$ (*d* = 0.42) for number 4 and $BF_{10} = 64.64$ (*d* = 0.24) for number 5. Another paired Bayesian *t*-test revealed strong evidence for a difference in smallest-number intercepts between the ranges and thus for AMdependency of the number mapping on the MNL (i.e., for Hypothesis 2b) with $BF_{10} = 546.98$ (*d* = 0.27). The last paired Bayesian *t*-test revealed strong evidence for a difference in SNARC slopes between the ranges and thus for AMdependency of the strength of the SNARC effect (i.e., for Hypothesis 3) with $BF_{10} = 1271.17$ (*d* = 0.28)[4].

[4] If data collection had been stopped with a sample size of 200 participants, results would have been similar with $BF_{10} = 4.13 * 10^{13}$ (*d* = 0.70) for Hypothesis 1a, $BF_{10} = 2.34 * 10^{9}$ (*d* = 0.57) for Hypothesis 1b, $BF_{10} = 5563.96$ (*d* = 0.37) for number 4 and $BF_{10} = 75.13$ (*d* = 0.29) for number 5 for Hypothesis 2a, $BF_{10} = 33.56$ (*d* = 0.27) for Hypothesis 2b, and $BF_{10} = 19.46$

Figure 4

Mean dRTs per number averaged across all trials from all participants in the lower (blue, solid line) and higher (orange, dashed line) number ranges for Experiment 2, with error bars representing ± 1 standard error for the respective number and regression lines representing slope estimates for the respective range



Exploratory data analysis

(d = 0.26) for Hypothesis 3. Thus, evidence would also have been strong and in favor of each hypothesis with 200 participants, just like with 300 participants.

As for Experiment 1, to disentangle the possible scenarios illustrated in Figures S1 and S2 and Table S1 (see Supplementary Material) in Experiment 2, the mean-number intercepts were compared between ranges. The mean number in the lower range (1 to 5, excluding 3) was 3 with a dRT estimate of -0.57 ms, and the mean number in the higher range (4 to 8, excluding 6) was 6 with a dRT estimate of -2.51 ms. A two-sided paired Bayesian *t*-test revealed moderate evidence against a difference in mean-number intercepts between the ranges and thus against AMdependency of the number mapping on the MNL with $BF_{10} = 0.18$.

As in Experiment 1, we tested whether there was a correlation between the SNARC slopes in the lower and the higher range. In contrast to the data of Experiment 1, the data of Experiment 2 revealed strong evidence for a moderate correlation with an estimate of $r = 0.34$ and $BF_{10} = 5.22 * 10^5$.

In contrast to Experiment 1, number parity and number magnitude were orthogonal in Experiment 2 (i.e., the mean number magnitude was equal for odd and even numbers in each range). Therefore, we were also able to test the MARC effect in Experiment 2. A two-sided Bayesian one-sample *t*-test of the MARC slopes against zero revealed moderate evidence against a MARC effect in the lower range with $BF_{10} = 0.16$, and inconclusive evidence regarding a MARC effect in the higher range with $BF_{10} = 0.51$.

Table 2

Descriptive self-reported information about the samples in both experiments (N = 173 in Experiment 1 and N = 255 in Experiment 2 after exclusions)

Demographic item	Answer	Experiment 1	Experiment 2
Gender	Woman	79 (45.7%)	121(47.5%)
	Man	92 (53.2%	133 (52.2%)
	Diverse	2 (1.2%)	1 (0.4%)

	No answer	0 (0.0%)	0 (0.0%)
Age	Mean (SD)	30.10 (5.52)	30.20 (5.86)
Native language	English	171 (98.8%)	253 (99.2%)
	Others	2 (1.2%)	2 (0.8%)
Handedness	Right-handed	150 (86.7%)	224 (87.8%)
	Left-handed	15 (8.7%)	24 (9.4%)
	Ambidextrous	8 (4.6%)	7 (2.7%)
Finger counting habit	Right-Starters	82 (47.4%)	128 (50.2%)
	Left-Starters	64 (37.0%)	95 (37.3%)
	Does not know or no preference	27 (15.6%)	32 (12.5%)
Finger counting stability	Always	57 (32.9%)	99 (38.8%)
	Mostly	72 (41.6%)	103 (40.4%)
	Slightly more often	15 (8.7%)	15 (5.9%)
	Does not know or no preference	29 (16.8%)	38 (14.9%)
Math skills (0-400)	Mean (SD)	240.80 (86.94)	235.80 (94.32)

Table 3

Average RT in ms per number in each range (with SEs in parentheses), for plots see Figures S3, S4, S5, and S6 in our Supplementary Material

Experiment 1				Experiment 2			
Lower range		Higher range		Lower range		Higher range	
Number	Mean RT (SE RT)	Number	Mean RT (SE RT)	Number	Mean RT (SE RT)	Number	Mean RT (SE RT)

0	566.63	4	525.25	1	517.87	4	513.47
	(8.43)		(6.75)		(5.19)		(5.15)
1	567.35	5	544.53	2	512.08	5	519.53
	(7.58)		(6.97)		(5.26)		(5.07)
2	538.28	6	540.12	4	514.18	7	508.70
	(7.61)		(7.40)		(5.24)		(5.12)
3	547.82	7	529.12	5	522.44	8	501.59
	(7.59)		(7.00)		(5.31)		(5.12)
4	539.00	8	529.96				
	(7.11)		(6.92)				
5	550.15	9	548.12				
	(7.52)		(7.28)				

Table 4

Parameter estimates for both experiments in the lower and higher number ranges (calculated per participant and averaged across them), with one asterisk indicating moderate evidence and with two asterisks indicating strong evidence for H0 (i.e., no difference between ranges) or for H1 (i.e., difference between ranges)

	Experiment 1			Experiment 2		
	Lower	Higher	Evidence	Lower	Higher	Evidence
SNARC intercept	11.72	34.68	H1**	29.93	33.00	H0*
SNARC slope	-5.53	-5.84	H0*	-10.17	-5.92	H1**
dRT for number 4	-16.03	12.18	H1**	-12.72	14.52	H1**
dRT for number 5	-15.63	4.02	H1*	-18.44	-1.95	H1**
Smallest-number intercept	11.72	11.31	H0*	19.76	9.32	H1**

Mean-number intercept	-2.09	-3.29	H0*	-0.57	-2.51	H0*
MARC slope	-	-	-	-5.88	10.07	Inconcl.

Table 5

This Study Design Table contains all research questions with corresponding hypotheses, the targeted sample size and planned analyses with a rationale, as well as the interpretations of potential outcomes and theoretical conclusions. All entries apply to both Experiment 1 (direct replication using 0 to 5 and 4 to 9) and Experiment 2 (conceptual replication using 1 to 5 excluding 3 and 4 to 8 excluding 6). The template for the Study Design Table was taken from PCI-RR and filled in before data collection started. Only the rightmost column "Observed outcome" has been added after data was collected and analyses were run.

Question	**Hypothesis**	**Sampling plan**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test** for confirming or disconfirming the hypothesis	**Interpretation** given different outcomes	**Theory that could be shown wrong** by the outcomes	**Observed outcome**
Can a SNARC effect be observed in all number ranges?	*Hypothesis 1 (and manipulation check):* A robust SNARC effect is expected in both (a) the lower and (b) the higher number ranges, i.e., we expect to find at least moderate evidence for	To reach the desired probability of .90 for finding moderate evidence in favor of a true underlying effect (i.e., $BF_{10}* > 3$) with an effect size of Cohen's $d = 0.15$ in two-sided Bayesian	Four regressions of dRTs on number magnitude followed by four two-sided Bayesian one-sample *t*-tests of SNARC slopes against zero (Experiment 1: 0 – 5 and 4 – 9; Experiment 2:	The most crucial aim of the present study is to find out whether AMdependency of the strength of the SNARC effect exists (*Hypothesis 3*). The minimally relevant effect size of $d = 0.15$	Finding moderate or even strong evidence for a SNARC slope smaller than 0 in a Bayesian *t*-test in each number range would provide evidence for a SNARC effect in both the lower (Hypothesis 1a)	The SNARC effect in the parity judgment task has been shown in numerous studies using different number ranges within the interval from 0 to 9 (as in all scenarios, see Figures S1 and	**Experiment 1:** Strong evidence for a SNARC effect in the lower range with $BF_{10} = 6956.04$ ($d = 0.38$) and in the higher range with $BF_{10} = 2.63 * 10^6$ ($d = 0.47$) **Experiment 2:** Strong evidence for a SNARC effect in the lower

	SNARC slopes (one per participant and per number range, calculated by regressing dRTs on number magnitude) to be smaller than zero in each number range. As the SNARC effect is very robust especially for lower ranges and possibly stronger than in higher ranges, the SNARC effect in lower ranges (Hypothesis 1a) will be used as manipulation check and prerequisite for following investigations (Hypotheses 1b, 2a, 2b, and 3).	one-sample *t*-tests or in two-sided Bayesian paired *t*-tests, 800 participants need to be tested in each experiment (for power calculations and sample size estimations, see https://doi.org/10.17605/OSF.IO/Z43PM). The required sample size for finding moderate evidence against a truly absent effect (i.e., $BF_{10} < 1/3$) for $d = 0$ is only 180. By ensuring our design is sensitive to find evidence for $d = 0.15$, we will be able to detect a slope difference of the size found by Fias et al. (1996), as	1 – 5 [excluding 3] and 4 – 8 [excluding 6])	was chosen because it corresponds to the SNARC slope difference of 2.99 ms between number ranges (with a pooled standard deviation of 18.34 ms) that was descriptively found but remained non-significant in the original study by Fias et al. (1996) that we wish to replicate here. Note that due to the lacking report of standard deviations, it is not possible to calculate Cohen's *d* for the slope difference of 9.2 ms found by Dehaene et al. (1993). Importantly, a smaller effect size than $d = 0.15$	and higher (Hypothesis 1b) number ranges and be in line with results from previous studies (e.g., the two seminal studies by Dehaene et al., 1993, and by Fias et al., 1996).	S2 and Table S1 in the supplementary materials: https://doi.org/10.17605/OSF.IO/Z43PM). We therefore expect to find at least moderate evidence for it in all four number ranges. Finding at least moderate evidence against the SNARC in any of the four number ranges would be highly surprising, especially in the lower number ranges. Evidence against the SNARC effect in the higher ranges (Hypothesis 1b) combined with evidence for the SNARC effect in the lower ranges (Hypothesis 1a) would provide	range with $BF_{10} = 1.61 * 10^{21}$ ($d = 0.71$) and in the higher range with $BF_{10} = 1.38 * 10^{12}$ ($d = 0.52$) **Summary:** Manipulation check successful in both experiments, replication of previous results, prerequisite for following investigations fulfilled

		predicted by Hypothesis 3, and a smaller effect size would not be meaningful for Hypotheses 1 and 2 either. However, we will employ the SBF+maxN approach as described by Schönbrodt & Wagenmakers (2018). More precisely, we will first recruit 200 participants and then calculate the BF_{10} for all *t*-tests after each added 20 participants. In case the BF_{10} reach a threshold of 1/3 or of 3 (i.e., moderate evidence for or against Hypotheses 1, 2, and 3) before getting to the		would not be meaningful for the SNARC effect (Hypothesis 1) or for RMdependency and AMdependency of the number mapping on the MNL (Hypothesis 2) either. Similarly, the chosen maximal sample size should be large enough to find at least moderate evidence in case Hypotheses 1 and 2 are false.		support for AMdependency of the strength of the SNARC effect (Hypothesis 3).	
Does the number mapping on the MNL[3] depend on whether it is the lowest vs. highest number in the current number range?	*Hypothesis 2a:* For the same critical number, a left-/right-hand advantage is expected when it is the lowest/highest number in the current number range, respectively. We hypothesize RMdependency[1] (and possibly AMdependency[2] as well, see Hypothesis 2b) of the number mapping on the MNL[3].		Four two-sided paired Bayesian *t*-tests of dRTs for the same number in lower vs. higher number range (i.e., for 4 and 5 in each experiment) (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number range of the respective experiment, see Hypothesis 1a, which serves as a manipulation check.)		Finding moderate or even strong evidence for a different pattern for numbers that appear in both number ranges in the lower and the higher number range in a *t*-test would provide evidence for RMdependency[1] of the SNARC effect. Finding moderate or even strong evidence against a different dRT pattern would indicate AMdependency[2] of the number mapping on the MNL[3].	Evidence for RMdependency[1] would indicate flexibility of the MNL[3], such that its resolution adapts to the context and that relative magnitude plays a role for spatial-numerical associations. However, this does not rule out the possibility that absolute magnitude plays a role as well (see below). Evidence for AMdependency[2] would indicate that the MNL[3] is at least not fully flexible.	**Experiment 1:** Strong evidence for differences in dRTs for critical numbers between the ranges with $BF_{10} = 2.57 * 10^4$ ($d = 0.40$) for number 4 and $BF_{10} = 12.86$ ($d = 0.25$) for number 5 **Experiment 2:** Strong evidence for differences in dRTs for critical numbers between the ranges with $BF_{10} = 6.64 * 10^7$ ($d = 0.42$) for number 4 and $BF_{10} = 64.64$ ($d = 0.24$) for number 5 **Summary:**

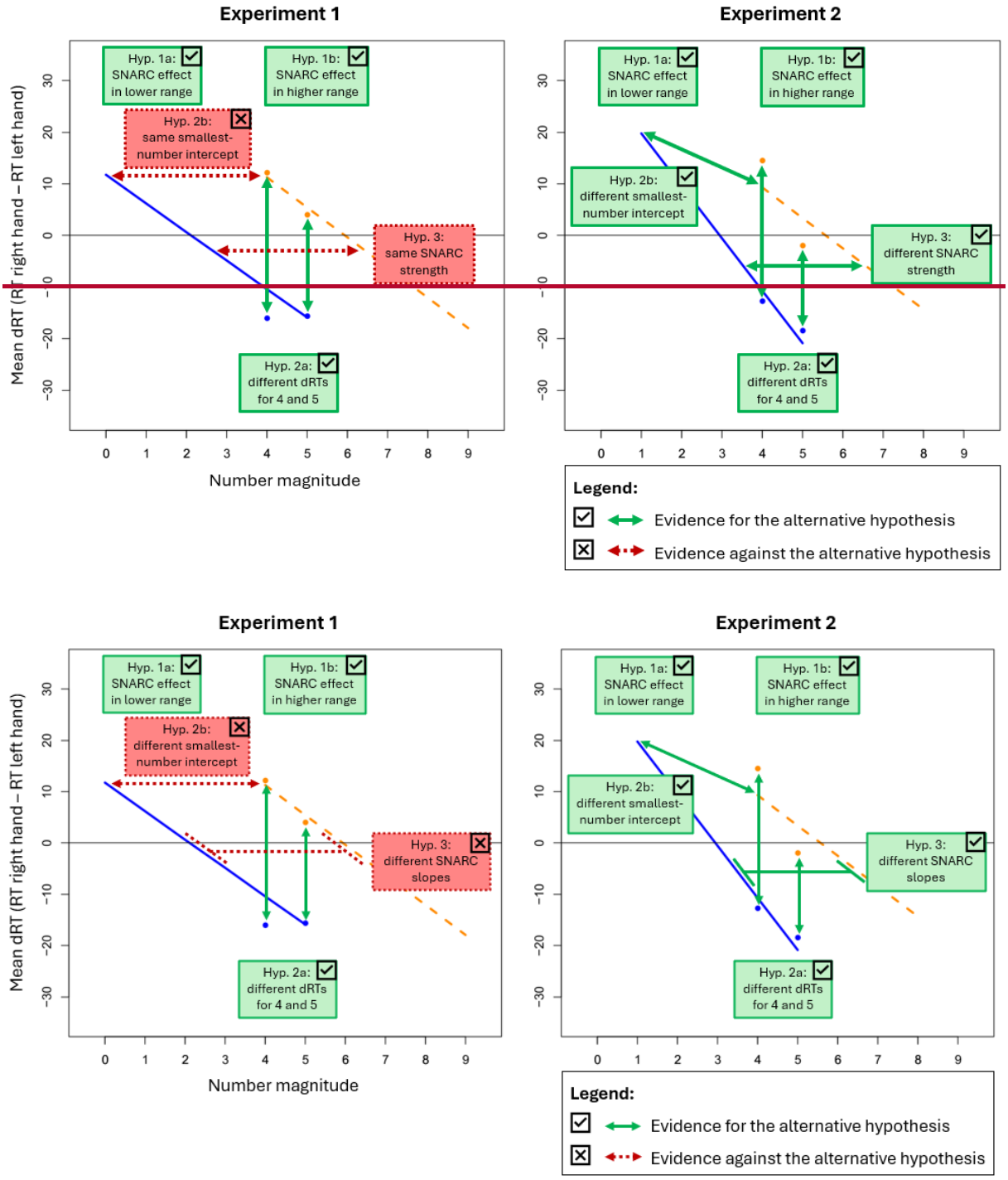
		sample size of 800 participants, we will stop recruiting earlier in the respective experiment.				Full RMdependency is illustrated in Scenarios 1 and 4, full AMdependency is shown in Scenarios 3 and 6, and a combination of both corresponds to Scenarios 2 and 5 in Figures S1 and S2.	Evidence for RMdependency and thus flexibility of the number mapping on the MNL in both experiments
Does the mapping of numbers on the MNL[3] depend on whether they are small vs. high numbers in absolute terms?	*Hypothesis 2b:* A left-/right-hand advantage could be observed for small/large numbers in absolute terms, respectively (on top of RMdependency[2], see Hypothesis 2a). However, we cannot derive any clear hypothesis from the literature about whether dRTs are lower for the		Two two-sided paired Bayesian *t*-tests of smallest-number intercept in lower vs. higher number range (one test per experiment) (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number		Finding moderate or even strong evidence for different smallest-number intercepts in the lower and the higher number range in a Bayesian *t*-test would indicate AMdependency[1] of the number mapping on the MNL[3]. Finding moderate or even strong	Evidence for AMdependency[1] would indicate that the MNL[3] and the SNARC effect are not fully flexible and that absolute magnitude plays a role for spatial-numerical associations. However, this does not rule out the possibility that relative magnitude plays	**Experiment 1:** Strong evidence against a difference in smallest-number intercepts between the ranges with BF_{10} = 0.09 **Experiment 2:** Strong evidence for a difference in smallest-number intercepts between the ranges with BF_{10} = 546.98 (d = 0.27) **Summary:**

	smallest number in a higher than in a lower range (as observed by Dehaene et al., 1993, but not by Fias et al., 1996).		range of the respective experiment, see Hypothesis 1a, which serves as a manipulation check.)		evidence against different smallest-number intercepts would indicate RMdependency[2] of the number mapping on the MNL[3].	a role as well (see above). Evidence for RMdependency[2] would indicate that the MNL[3] is at least partly flexible.	Evidence against AMdependency of the number mapping on the MNL in Experiment 1, but evidence for it in Experiment 2, suggesting that the number mapping on the MNL is not fully flexible when certain stimulus sets are used
Does the strength of the SNARC effect depend on absolute number magnitudes in the used range?	*Hypothesis 3:* The SNARC effect is expected to be stronger in the lower than in the higher number ranges.		Two two-sided paired Bayesian *t*-tests of SNARC slopes in lower vs. higher number range (one test per experiment) (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number range of the respective		Finding moderate or even strong evidence for a more negative SNARC slope in one of the two number ranges would indicate that the SNARC effect seems to be stronger in this number range than in the other.	Finding the SNARC effect to be stronger in the lower than in the higher number range, would indicate that the spatial mental representation of small numbers is more pronounced than for large numbers (as in Scenarios 4, 5, 6 in Figure S1). If the SNARC effect does not differ between	**Experiment 1:** Strong evidence against a difference in SNARC slopes between the ranges with BF_{10} = 0.09 **Experiment 2:** Strong evidence for a difference in SNARC slopes between the ranges with BF_{10} = 1271.17 **Summary** Evidence against AMdependency of the strength of

			experiment, see Hypothesis 1a, which serves as a manipulation check.)			number ranges, no evidence can be provided for the strength of the SNARC effect to depend on absolute number magnitudes (as in Scenarios 1, 2, 3).	the SNARC effect in Experiment 1, but evidence for it in Experiment 2, suggesting that the spatial mental representation of small numbers is more pronounced than for large numbers when certain stimulus sets are used

Figure 5

Summary of all tested hypotheses and outcomes in the plot for the linear regression of mean dRTs on number magnitude separately for the lower (blue, solid line) and higher (orange, dashed line) number ranges for Experiments 1 (left panel, see Figure 3) and 2 (right panel, see Figure 4)



Note. The figure only includes the mean dRTs for the critical numbers 4 and 5, which appear in both the lower and the higher number ranges. Hypotheses 1a and 1b were tested with one-sample *t*-tests, whereas Hypotheses 2a, 2b, and 3 were comparisons tested with paired *t*-tests and are illustrated with two-sided arrows. Green boxes with a solid outline and a checkmark as well as green solid arrows indicate Bayesian evidence for the alternative hypothesis (i.e., $BF_{10} > 3$). Red boxes with a dotted outline and a cross as well as red dotted arrows indicate Bayesian evidence against the alternative hypothesis (i.e., $BF_{10} < 1/3$).

Discussion

The goal of the present study was to determine the degree of flexibility of SNAs. More precisely, we wanted to find out whether the SNARC effect is entirely flexible and depends only on relative magnitude (RMdependency), or whether it is less flexible than previously assumed and also depends on absolute magnitude (AMdependency). Importantly, the concepts of RMdependency vs. AMdependency can be differentiated in two ways: (i) The number mapping on the MNL (e.g., dRT for number 4) can be RMdependent, AMdependent, or both, and (ii) the strength of the SNARC effect in terms of the relative increase of right-hand advantage per increase in magnitude (i.e., the SNARC slope) can be AMdependent or not. To summarize, the aim of the study was to determine whether SNAs operates on fixed and flexible number representations simultaneously.

RMdependency and AMdependency of SNAs

In the two seminal studies by Dehaene et al. (1993; Experiment 3) and by Fias et al. (1996; Experiment 1), numbers 4 and 5 were associated with the right when presented in the range from 0 to 5, but with the left when presented in the range from 4 to 9. These results are often quoted in the literature as evidence for pure RMdependency, although the studies were underpowered for small effects and descriptive data suggests AMdependency as well. In Experiment 1, which was run online with the same stimulus sets as in the two original studies,

we replicated the findings from the original studies. Specifically, we found strong evidence for a small-sized SNARC effect in the lower (average slope of -5.53) and higher (average slope of -5.84) range separately (Hypothesis 1), as well as for small dRT differences for critical numbers (i.e., 4 and 5) between the ranges (Hypothesis 2a), in line with the original studies. Moreover, we observed strong evidence against smallest-number intercept (Hypothesis 2b) and SNARC slope differences (Hypothesis 3) as well as moderate evidence against mean-number intercept differences (exploratory analysis) between the lower and the higher range. To conclude, the results from our investigation are entirely in line with the results from the original studies, and in fact, without further consideration (see below), they seem to support the conclusion of full RMdependency drawn by Dehaene et al. (1993) and Fias et al. (1996): Whereas their data hinted descriptively towards AMdependency as pointed out in the Introduction, we even observed moderate evidence against it.

Crucially, the second experiment in the present study yielded different results. In Experiment 2, we ran a conceptual online replication considering recent advances in SNARC research with the number ranges from 1 to 5 (excluding 3) and from 4 to 8 (excluding 6). As opposed to the original stimulus set, this modified stimulus set avoids potential confounds with the MARC effect thanks to the equal number of odd and even numbers in the original stimulus set and parity being orthogonal to magnitude. Also, number 0 was excluded because it has special properties and deviates in its dRT pattern from all other numbers (see (Nuerk et al., 2004, for detailed analysis). In line with the literature, the dRT pattern for number 0 observed in the current study did not align well with the regression line, and the RT variation was descriptively larger for number 0 than for all other numbers. Therefore, one would not want to base a general statement about the flexibility of the number line on number 0, because the relatively low dRT estimate for 0 might considerably attenuate the SNARC effect in the lower range. However, these two number ranges used in Experiment 2 still include the same critical numbers that are part of both number ranges in the original studies, namely 4 and 5. As in

Experiment 1, we found strong evidence for the SNARC effect in the lower (average slope of -10.17) and higher (average slope of -5.92) range separately (Hypothesis 1). Note that in contrast to the small effect size in both ranges in Experiment 1 (d = 0.38 and d = 0.47), the effect size was medium in both ranges in Experiment 2 (d = 0.71 and d = 0.52). Thus, our results support the claim that the inclusion of number 0 in the stimulus set or a potential confound with the MARC effect due to an unequal number of odd and even numbers might have decreased the SNARC effect in the seminal studies by Dehaene et al. (1993) and Fias et al. (1996). Moreover, as in Experiment 1, strong evidence was found for dRT differences for critical numbers (i.e., 4 and 5) between ranges (Hypothesis 2a). Further, as in Experiment 1, moderate evidence was found against mean-number intercept differences between ranges (exploratory analysis). However, the support for RMdependency is not the entire story.

Importantly, the data revealed strong evidence for differences in smallest-number intercept (Hypothesis 2b) and SNARC slopes (Hypothesis 3) between the ranges. These differences indicating AMdependency are only small (d = 0.27 and d = 0.28 for Hypotheses 2b and 3, respectively), but so are the differences indicating RMdependency (d = 0.42 and d = 0.24 for Hypothesis 2a). Hence, in contrast to the pure RMdependency observed in Experiment 1, the results from Experiment 2 suggest AMdependency both for (i) the number mapping on the MNL, and (ii) the strength of the SNARC effect (in line with Scenario 5 in the Supplementary Material).

How can these results be reconciled with Experiment 1, where we observed evidence against AMdependency? As outlined above, we conducted Experiment 2 (among other reasons) to see whether 0 drives results for the number ranges from 0 to 5 and from 4 to 9 used in the original experiments by Dehaene et al. (1993) and Fias et al. (1996). Therefore, we reanalyzed the data in an exploratory way without 0. Note this is not the same as if the experiment had been run without number 0. It was part of the stimulus set, even when it was not analyzed later, and the number range used in the experiment might still influence results. When excluding

number 0 from the stimuli set in Experiment 1 post-hoc (see Footnote 1), the evidence against AMdependency disappeared. Instead, the evidence was now inconclusive regarding smallest-number intercept differences (Hypothesis 2b) and SNARC slope differences (Hypothesis 3). Since number 0 was still part of the range of Experiment 1, we did not expect the same results as in Experiment 2, which was run without 0. However, when the results of Experiment 1 are analyzed without 0, there is at least no conflicting evidence anymore. This change suggests that the inclusion of 0 in some range plays a major role for the eventual outcome.

Notably, the dRT was also rather high for number 1 in both experiments and pulled the regression line upwards. That is, number 1 seems to be more strongly associated with the left than what would be predicted based on the regression slope alone. This observation fits with the results from our two previously conducted color judgment tasks, which do not require semantic number processing at all (Roth, Caffier et al., 2024). Number 1 seemed to be strongly associated with the left in these experiments as well, providing further support for AMdependency of the MNL.

In the present study, Bayesian analyses permitted to interpret and quantify evidence for the null hypotheses. Moreover, these findings can be considered as trustworthy, because the sample sizes were large enough (173 datasets analyzed out of 200 recruited participants in Experiment 1 and 255 datasets analyzed out of 300 recruited participants in Experiment 2). Thanks to the SBF+maxN approach, an optimally sized sample was recruited in each experiment. These samples were much larger than in the original studies, where a difference between number ranges might just have stayed undetected due to lacking statistical power.

To sum up, although the picture is blurred by methodological issues, especially the inclusion of number 0 in one but not in the other range, in the original studies and in our Experiment 1, the findings from Experiment 2 together with the reanalysis of Experiment 1 without 0 seem to suggest that there is not the “one and only SNARC”. The spatial mapping of numerical magnitude onto space seems not to be fully flexible and dependent on the used range.

Null effects of absolute magnitude can only be found when number 0 is included. We therefore conclude that the SNARC effect seems to operate on multiple number representations and on multiple spatial reference frames simultaneously, namely on both flexible and absolute ones.

RMdependency and AMdependency from a theoretical point of view

As outlined in the Introduction, different predictions regarding the SNARC effect's flexibility can be derived from the models that have been proposed to account for the origin of the SNARC effect. The working memory account (Fias & van Dijck, 2016; van Dijck & Fias, 2011) postulates that the SNARC effect is constructed during task execution. In the present study, the currently used stimulus set (i.e., the lower or the higher number range) was stored in working memory, which was reflected by the differential patterns for critical numbers 4 and 5. Namely, when these critical numbers were the highest (i.e., in the lower ranges in both experiments), they were associated with the right, whereas they were associated with the left when they were the lowest (i.e., in the higher ranges in both experiments). This observed RMdependency is clearly in line with the working memory account. In contrast, although the MNL (Dehaene et al., 1993) is claimed to dynamically adapt to task demands as well, such that zooming in and out is possible (Pinhas et al., 2013), it can be considered as a mental representation in long-term memory. Similarly, the verbal-spatial coding account (Gevers et al., 2010) and polarity correspondence account (Proctor & Cho, 2006) postulate number representations stored in long-term memory. Crucially, long-term representations hardly justify the SNARC effect's flexibility (Ginsburg & Gevers, 2015; van Dijck et al., 2015). Instead, they are in line with the AMdependency observed in Experiment 2, reflected by (i) the number mapping on the MNL in terms of smallest-number intercept differences between ranges, and (ii) the degree of spatialization in terms of differences between ranges regarding the strength of the SNARC effect. A potential explanation for this AMdependency is that small numbers are more frequently used than large numbers, which might lead to a more fixed and stronger spatial mental representation. Importantly, RMdependency and AMdependency can coexist (in line

with Ginsburg et al., 2014; Koch et al., 2023; van Dijck et al., 2015) and multiple spatial reference frames can be activated simultaneously (Weis et al., 2018).

A mental number line account, in which spatial associations are partly flexible (aka zoomed in) and partly fixed (i.e. always more left for absolutely smaller numbers) can also explain the current data (see Koch et al., 2023). However, any account, which wishes to explain the current data, needs a fixed and a flexible component. Any fully flexible account is in our view not consistent with these data.

Correlation of the SNARC effect between ranges

Apart from the SNARC effect at the group level, one can also investigate the phenomenon at the individual level (Cipora, van Dijck, et al., 2019; Roth, Jordan, et al., 2024). Assessing parity judgment in both stimulus ranges in a within-subjects design permits to investigate the correlation of SNARC effects between ranges (exploratory analyses), which was not tested in the original studies. Surprisingly, Experiment 1 revealed moderate evidence against a correlation of SNARC slopes between ranges, whereas Experiment 2 revealed strong evidence for a moderate correlation, with an estimate of $r = 0.34$. A hypothetical explanation could be that the inclusion of number 0 in the stimulus set of Experiment 1 might have blurred a true underlying correlation[5]. That is, number 0 showed descriptively stronger variations in RTs than other numbers did, which is in line with the literature (see Figure 4 in Nuerk et al., 2004), thus introducing a large error term for the dRT. The dRT for number 0 is rather uncorrelated with dRTs for other numbers (Nuerk et al., 2004), so slopes built on a stimulus set including 0 are also more likely to be uncorrelated with other slopes built on stimulus sets excluding 0. This large variation in turn results in increased noise for the SNARC slopes in the

[5] Excluding number 0 from the lower and number 4 from the higher range in Experiment 1 in the analysis did not affect the results regarding the correlation: Moderate evidence against a correlation was found, with $BF_{10} = 0.22$. Note that this does not contradict the strong evidence for a correlation in Experiment 2, because number 0 was part of the stimulus set when running Experiment 1 and might have undermined the potential correlation.

lower number range in Experiment 1. To summarize, our results from Experiment 2 show a positive relationship between the strength of the spatial mapping in small and large numbers within participants.

Conclusion

The current study demonstrates that the spatial mental representation of numbers is not entirely flexible. Mental spatialization can be adapted to the context and depends largely on relative number magnitude, but at the same time it is also influenced by absolute number magnitude. Relative and absolute number magnitude play a role for (i) the association of specific numbers with horizontal directional space, such that numbers that are small in relative or absolute terms are associated with the left, whereas numbers that are large in relative terms are associated with the right. At the same time, relative and absolute magnitude play a role for (ii) the strength of the spatialization, such that the association of small with left and large with right is stronger within a lower than within a higher number range. To conclude, the spatial representation of number magnitude seems to be partly flexible and partly fixed.

Competing interests

The authors declare no financial or non-financial conflicts of interest with the content of this article.

Author contributions

All the authors have full access to all the data and take responsibility for the integrity of the data and the accuracy of the data analysis. *Conceptualization*: K. Cipora, H.-C. Nuerk, U.-D. Reips; *Data Curation*: K. Cipora, H.-C. Nuerk, U.-D. Reips, L. Roth; *Formal Analysis*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Funding Acquisition*: K. Cipora, H.-C. Nuerk, U.-D. Reips.; *Investigation*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Methodology*: K. Cipora, H.-C. Nuerk, U.-D. Reips, L. Roth; *Project Administration*: H.-C. Nuerk, U.-D. Reips, L. Roth; *Resources*: H.-C. Nuerk, U.-D. Reips; *Software*: J. Caffier, A. T. Overlander, U.-D. Reips; *Supervision*: K. Cipora, H.-C. Nuerk, U.-D. Reips; *Validation*: K. Cipora, H.-C. Nuerk, U.-D. Reips, L. Roth; *Visualization*: L. Roth; *Writing – original draft*: L. Roth; *Writing – review and editing:* J. Caffier, K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips.

Acknowledgements

This research was supported by the DFG project “Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)” (NU 265/8-1 and RE 2655/3-1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips supporting Lilly Roth, John Caffier, with the assistance of Krzysztof Cipora as a cooperation partner. Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1). The authors would like to thank Sebastian Sandbrink for English proofreading of the stage 1 Registered Report. Moreover, the authors wish to thank PCI-RR and especially the recommender Robert McIntosh for their helpful feedback at both stages 1 and 2.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). Rmarkdown: Dynamic Documents for R (version 2.18). R package. https://github.com/rstudio/rmarkdown

Antoine, S., & Gevers, W. (2016). Beyond left and right: Automaticity and flexibility of number-space associations. *Psychonomic Bulletin & Review*, *23*(1), 148-155. https://doi.org/10.3758/s13423-015-0856-x

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., …, & Czekanski, M. (2024). data.table: Extension of „data.frame“ (version 1.15.2). R package. https://CRAN.R-project.org/package=data.table

Ben Nathan, M., Shaki, S., Salti, M., & Algom, D. (2009). Numbers and space: Associations and dissociations. *Psychonomic Bulletin & Review*, *16*(3), 578-582. https://doi.org/10.3758/PBR.16.3.578

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*, 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Brysbaert, M. (1995). Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, *124*(4), 434-452. https://doi.org/10.1037/0096-3445.124.4.434

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H. (2018). Pwr: Basic functions for power analysis (version 1.3-0). R package. https://CRAN.R-project.org/package=pwr

Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: why related? *Annals of the New York Academy of Sciences, 1477*(1), 5-19. https://doi.org/10.1111/nyas.14355

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. Hubbard (Ed.), *Spatial Biases in Perception and Cognition* (pp. 41–59). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316651247

Cipora, K., Schroeder, P. A., Soltanlou, M., & Nuerk, H.-C. (2018). More space, better mathematics: Is space a powerful tool or a cornerstone for understanding arithmetic? In K. S. Mix & M. T. Battista (Eds.), *Visualizing Mathematics. Research in Mathematics Education* (pp. 77-116). Springer, Cham. https://doi.org/10.1007/978-3-319-98767-5_4

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods, 51*(4), 1676-1692. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). A Minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect. PsyArXiv. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20∗20 Monte Carlo investigation. *Frontiers in Psychology, 8*, 1194. https://doi.org/10.3389/fpsyg.2017.01194

Cleland, A. A., & Bull, R. (2019). Automaticity of access to numerical magnitude and its spatial associations: The role of task and number representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(2), 333-348. https://doi.org/10.1037/xlm0000590

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

Deng, Z., Chen, Y., Zhu, X., & Li, Y. (2017). The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Memory & Cognition, 45*(3), 428-441. https://doi.org/10.3758/s13421-016-0676-x

Di Giorgio, E., Lunghi, M., Rugani, R., Regolin, L., Dalla Barba, B., Vallortigara, G., & Simion, F. (2019). A mental number line in human newborns. *Developmental Science, 22*(6), e12801. https://doi.org/10.1101/159335

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9–26. https://doi.org/10.1037/cns0000258

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

Fias, W., & van Dijck, J.-P. (2016). The temporary nature of number—space interactions. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale, 70*(1), 33–40. https://doi.org/10.1037/cep0000071

Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition – From single digits to arithmetic. *Quarterly Journal of Experimental Psychology, 67*(8), 1461-1483. https://doi.org/10.1080/17470218.2014.927515

Funke, F., & Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods, 24*(3), 310–327. https://doi.org/10.1177/1525822X12444061

Gevers, W., Santens, S., Dhooge, E., Chen, Q., Van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number–space interactions. *Journal of Experimental Psychology: General, 139*(1), 180-190. https://doi.org/10.1037/a0017688

Giner-Sorolla, R., Aberson, C. L., Bostyn, D. H., Carpenter, T., Conrique, B. G., Lewis, N. A., & Soderberg, C. (2019). Power to detect what? Considerations for planning and evaluating sample size. Open Science Framework. https://osf.io/jnmya/

Ginsburg, V., & Gevers, W. (2015). Spatial coding of ordinal information in short-and long-term memory. *Frontiers in Human Neuroscience, 9,* 1-10. https://doi.org/10.3389/fnhum.2015.00008

Ginsburg, V., van Dijck, J.-P., Previtali, P., Fias, W., & Gevers, W. (2014). The impact of verbal working memory on number–space associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 976-986. https://doi.org/10.1037/a0036378

Giurfa, M., Marcout, C., Hilpert, P., Thevenot, C., & Rugani, R. (2022). An insect brain organizes numbers on a left-to-right mental number line. *Proceedings of the National Academy of Sciences, 119*(44), e2203584119. https://doi.org/10.1073/pnas.2203584119

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential effects in SNARC. *Scientific Reports, 8*(1), 1-13. https://doi.org/10.1038/s41598-018-29337-2

Jenadeleh, M., Zagermann, J., Reiterer, H., Reips, U.-D., Hamzaoui, R., & Saupe, D. (2023). *Relaxed forced choice improves performance of visual quality assessment methods.* Proceedings of the 15th International Conference on Quality of Multimedia Experience (QoMEX), Ghent, June 2023. ArXiv. https://doi.org/10.48550/arXiv.2305.00220

Kelter, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics, 36*(2), 1263–1288. https://doi.org/10.1007/s00180-020-01034-7

Koch, N. N., Huber, J. F., Lohmann, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2023). Mental Number Representations Are Spatially Mapped Both by Their Magnitudes and Ordinal Positions. *Collabra: Psychology*, *9*(1), Article 67908. https://doi.org/10.1525/collabra.67908

Levenson, E., Tsamir, P., & Tirosh, D. (2007). Neither even nor odd: Sixth grade students' dilemmas regarding the parity of zero. *The Journal of Mathematical Behavior, 26*(2), 83-95. https://doi.org/10.1016/j.jmathb.2007.05.004

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149-157. https://doi.org/10.1037/0278-7393.16.1.149

Luck, S. (2019, February 19). Why experimentalists should ignore reliability and focus on precision. Luck Lab. Retrieved from https://lucklab.ucdavis.edu/blog/2019/2/19/reliability-and-precision

Lukács, G., Kleinberg, B., & Doorn, J. van. (2022). neatStats: Neat and Painless Statistical Reporting (version 1.13.3). R package. https://cran.r-project.org/web/packages/neatStats/index.html

Mitchell, T., Bull, R., & Cleland, A. A. (2012). Implicit response-irrelevant number information triggers the SNARC effect: Evidence using a neural overlap paradigm. *Quarterly Journal of Experimental Psychology*, *65*(10), 1945-1961. https://doi.org/10.1080/17470218.2012.673631

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2024). BayesFactor: Computation of Bayes factors for common designs (version 0.9.12-4.7). R package. https://CRAN.R-project.org/package=BayesFactor

Nieder, A. (2016). Representing something out of nothing: The dawning of zero. *Trends in Cognitive Sciences, 20*(11), 830-842. https://doi.org/10.1016/j.tics.2016.08.008

Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *The Quarterly Journal of Experimental Psychology Section A, 57*(5), 835-863. https://doi.org/10.1080/02724980343000512

Patro, K., Nuerk, H.-C., Cress, U., & Haman, M. (2014). How number-space relationships are assessed before formal schooling: A taxonomy proposal. *Frontiers in Psychology*, *5*, 419. https://doi.org/10.3389/fpsyg.2014.00419

Pinhas, M., & Tzelgov, J. (2012). Expanding on the mental number line: Zero is perceived as the "smallest". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1187-1205. https://doi.org/10.1037/a0027390

Pinhas, M., Pothos, E. M., & Tzelgov, J. (2013). Zooming in and out from the mental number line: Evidence for a number range effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 972-976. https://doi.org/10.1037/a0029527

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 (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89-117). Academic Press. https://doi.org/10.1016/B978-012099980-4/50005-8

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49*(4), 243-256. https://doi.org/10.1026/1618-3169.49.4.243

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. Human Vision and Electronic Imaging XIV, *Proceedings of SPIE, 7240*, Article 724008. https://doi.org/10.1117/12.823416

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers, 34*(2), 234-240. https://doi.org/10.3758/BF03195449

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

Roth, L., Caffier, J. P., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2024). *True colors SNARCing: Automaticity of the SNARC effect – evidence from color judgment tasks* [Preprint on OSF]. https://doi.org/10.31234/osf.io/aeyn8

Roth, L., Jordan, V., Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.-P., & Cipora, K. (2024). Don't SNARC me now! Intraindividual variability of cognitive phenomena – Insights from the Ironman paradigm. *Cognition, 248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science, 347*(6221), 534-536. https://doi.org/10.1126/science.aaa1379

Schaefer, J., Opgen-Rhein, R., & Strimmer, K. (2021). GeneNet: Modeling and inferring gene networks (version 1.2.16). R package. https://CRAN.R-project.org/package=GeneNet

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schwarz, W., & Keus, I. M. (2004). Moving the eyes along the mental number line: Comparing SNARC effects with saccadic and manual responses. *Perception & Psychophysics, 66*(4), 651-664. https://doi.org/10.3758/BF03194909

Stieger, S., Reips, U.-D., & Voracek, M. (2007). Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. *Journal of the American Society for Information Science and Technology, 58*, 1653-1660. http://doi.org/10.1002/asi.20651

Tlauka, M. (2002). The processing of numbers in choice-reaction tasks. *Australian Journal of Psychology*, *54*(2), 94-98. https://doi.org/10.1080/00049530210001706553

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

Tzelgov, J., Zohar-Shai, B., & Nuerk, H.-C. (2013). On defining quantifying and measuring the SNARC effect. *Frontiers in Psychology*, *4*, 302. https://doi.org/10.3389/fpsyg.2013.00302

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

van Dijck, J.-P., Ginsburg, V., Girelli, L., & Gevers, W. (2015). Linking numbers to space: From the mental number line towards a hybrid account. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 89–105). Oxford University Press.

Weis, T., Nuerk, H.-C., & Lachmann, T. (2018). Attention allows the SNARC effect to operate on multiple number lines. *Scientific Reports*, *8*(1), 1-13. https://doi.org/10.1038/s41598-018-32174-y

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wickelmaier, F. (2022). *Simulating the Power of Statistical Tests: A Collection of R Examples.* ArXiv. https://doi.org/10.48550/arXiv.2110.09836

Wickham, H. (2023). plyr: Tools for splitting, applying and combining data [R package, version 1.8.9]. https://CRAN.R-project.org/package=plyr

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & van den Brand, T. (2024). ggplot2: Create elegant data visualisations using the grammar of graphics [R package, version 3.5.0]. https://cran.r-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation [R package, version 1.1.4]. https://CRAN.R-project.org/package=dplyr

Wickham, H., Vaughan, D., Girlich, M., & Ushey, K. (2023). tidyr: Tidy messy data [R package, version 1.3.1]. https://CRAN.R-project.org/package=tidyr

Wühr, P., & Richter, M. (2022). Relative, not absolute, stimulus size is responsible for a correspondence effect between physical stimulus size and left/right responses. *Attention, Perception, & Psychophysics, 84*(4), 1342-1358. https://doi.org/10.3758/s13414-022-02490-7

Xie Y. (2022). knitr: A General-Purpose Package for Dynamic Report Generation in R (version 1.41). R package. https://yihui.org/knitr/

Yu, S., Li, B., Zhang, S., Yang, T., Jiang, T., Chen, C., & Dong, Q. (2018). Does the spatial-numerical association of response codes effect depend on digits' relative or absolute magnitude? Evidence from a perceptual orientation judgment task. *The Journal of General Psychology*, *145*(4), 415-430. https://doi.org/10.1080/00221309.2018.1532391

Zohar-Shai, B., Tzelgov, J., Karni, A., & Rubinsten, O. (2017). It does exist! A left-to-right spatial–numerical association of response codes (SNARC) effect among native Hebrew speakers. *Journal of Experimental Psychology: Human Perception and Performance, 43*(4), 719–728. https://doi.org/10.1037/xhp0000336

One and only SNARC?

Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude

Supplementary Material

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Hans-Christoph Nuerk[1,4,5], Annika Tave Overlander[2], Krzysztof Cipora[3*]

[1]Department of Psychology, University of Tübingen, Germany

[2]Department of Psychology, University of Konstanz, Germany

[3]Centre for Mathematical Cognition, Loughborough University, United Kingdom

[4]LEAD Graduate School & Research Network, University of Tübingen, Germany

[5]German Center for Mental Health (DZPG)

*Corresponding author:

k.cipora@lboro.ac.uk

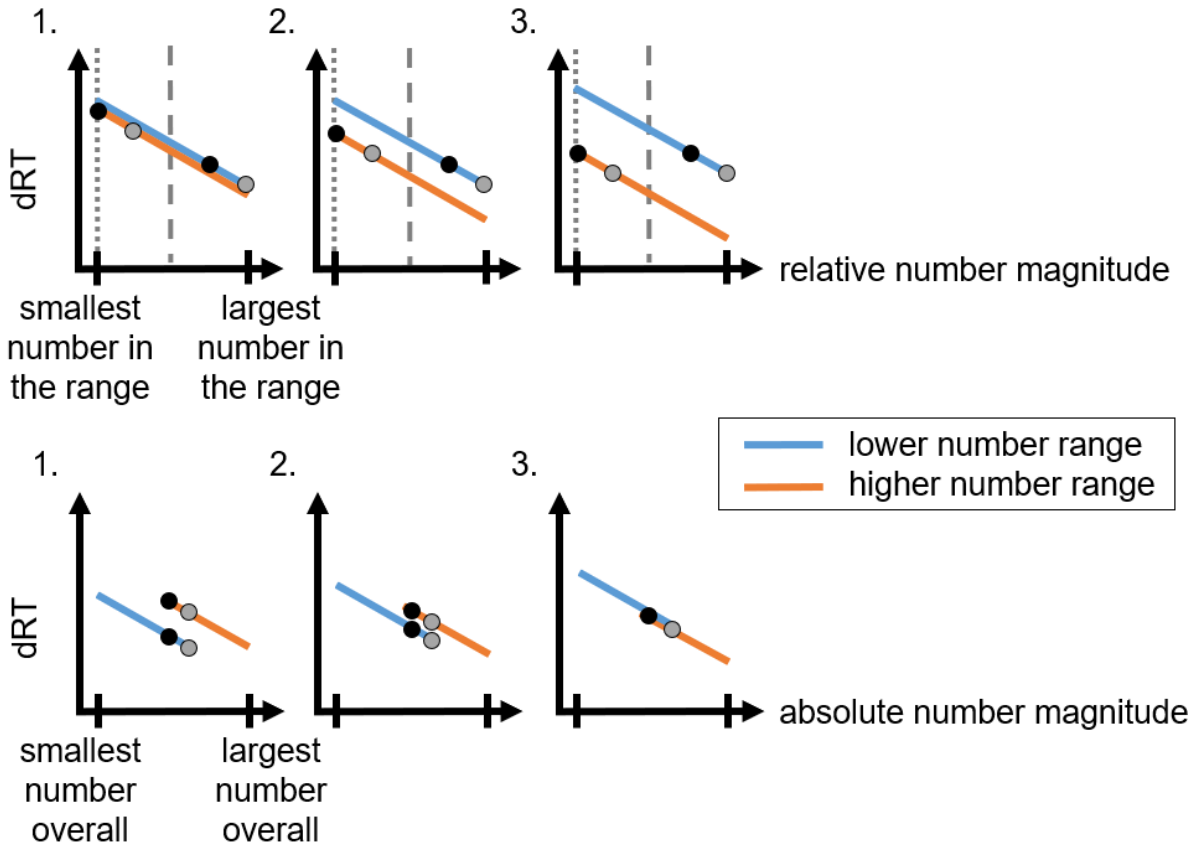
Possible scenarios of RMdependency and AMdependency

In the following, we want to present six possible scenarios regarding RMdependency and AMdependency of both the number mapping on the MNL and the strength of the SNARC effect. Apart from the regression slope that quantifies the strength of the SNARC effect, the smallest-number intercept (when relative magnitude of the numbers in both ranges is matched, i.e., the predicted dRT for 0 and 4 in Experiment 1 and for 1 and 4 in Experiment 2) and the mean-number intercept (i.e., the predicted dRT for 2.5 and 6.5 in Experiment 1 and for 3 and 6 in Experiment 2) can be determined in order to investigate the number mapping on the MNL. When discussing RMdependency and AMdependency of the SNARC effect, the following scenarios are possible (see Figures S1 and S2 and Table S1):

1. RMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different dRTs of critical numbers that are part of both number ranges, namely 4 and 5)
2. Both RMdependency and AMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different dRTs of critical numbers, different smallest-number intercepts, and different mean-number intercepts)
3. AMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different smallest-number intercepts and different mean-number intercepts) – note that concluding RMdependency of the number mapping on the MNL from finding a significant SNARC effect in both number ranges without testing dRTs of critical numbers is incorrect

Figure S1

Possible scenarios of RMdependency and AMdependency of the number mapping on the MNL

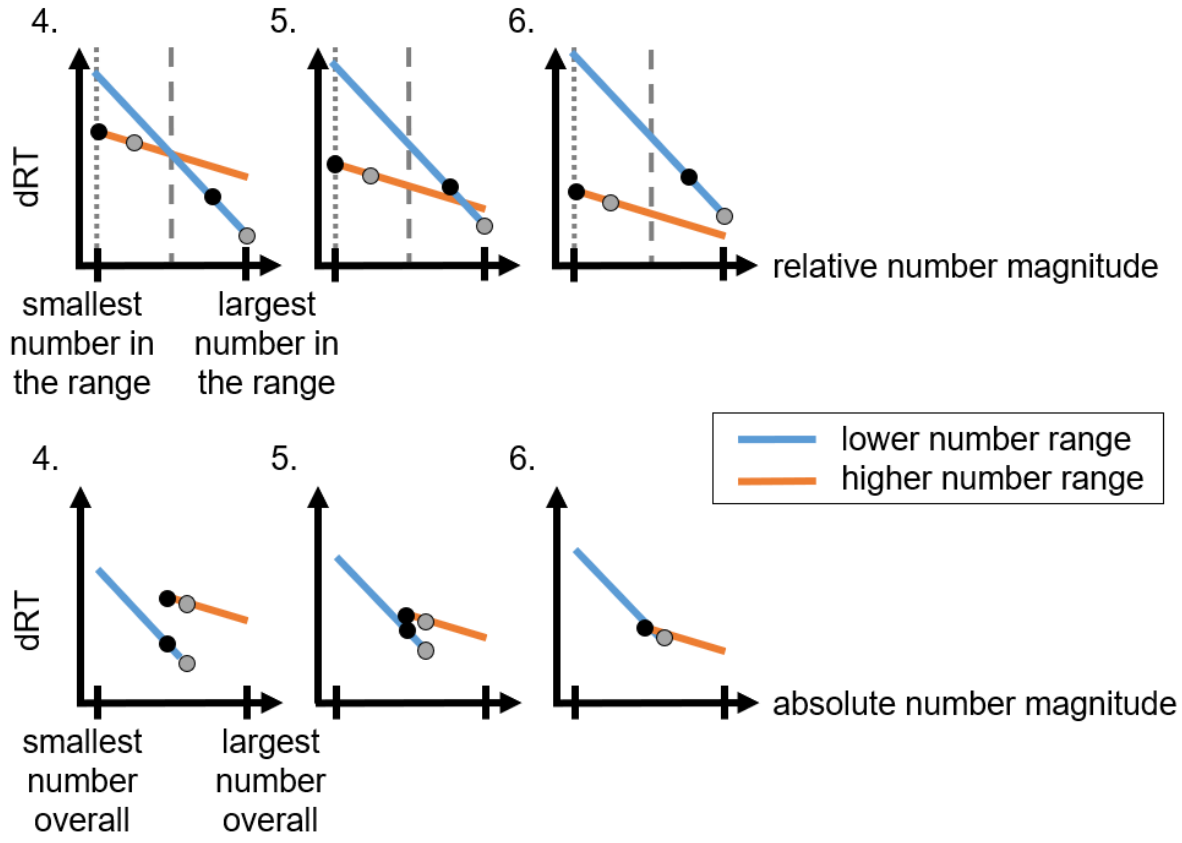


Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 1, 2, and 3, with the regression lines for the lower and higher number ranges being represented in blue and orange, respectively. In the upper part of the figure, relative number magnitudes are used for the x-axis, so that the regression lines for both number ranges start at their smallest and end at their largest number magnitude. For example, in Experiment 1, the dRTs for 0 (smallest number in the lower number range) and 4 (smallest number in the higher number range) are on the very left, and the dRTs for 5 (largest number in the lower number range) and 9 (largest number in the higher number range) are on the very right. In the lower part of the figure, the same scenarios are illustrated, but absolute number magnitudes are used for the x-axis. In our study, the absolute number magnitudes will be 0 to 5 and 4 to 9 in Experiment 1, and 1 to 5 (excluding 3) and 4 to 8 (excluding 6) in Experiment 2. For example, the dRTs for numbers 4 and 5 are on the very same spot of the x-axis for both the lower and the higher range, because they have the same absolute magnitude. The dotted line in the upper part of the figure depicts the intercept for the smallest number magnitude, and the dashed line depicts the intercept for the mean number magnitude in the respective number range. The black and the gray dots indicate the critical numbers being part of both the lower and the higher number range (i.e., 4 and 5).

4. AMdependency of the strength of the SNARC effect, and RMdependency of the number mapping on the MNL (i.e., different SNARC slopes, different dRTs of critical numbers, different smallest-number intercepts), as in Fias et al. (1996)
5. AMdependency of the strength of the SNARC effect, and both RMdependency and AMdependency of the number mapping on the MNL (i.e., different SNARC slopes, different dRTs of critical numbers, different smallest-number intercepts, and mean-number intercepts), as in Dehaene et al. (1993)
6. AMdependency of the strength of the SNARC effect and of the number mapping on the MNL (i.e., different SNARC slopes, different smallest-number intercepts, and different mean-number intercepts)

Figure S2

Possible scenarios of RMdependency and AMdependency of the strength of the SNARC Effect



Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 4, 5, and 6. For an explanation of magnitudes on the x-axis as well as concrete examples for data points, see *Note* of Figure S1.

Table S1

Possible Scenarios of RMdependency and AMdependency of the SNARC Effect

	Scenario					
Characteristic of the scenario	1	2	3	4	5	6
SNARC effect in both ranges	yes	yes	yes	yes	yes	yes
Different dRTs for critical numbers (4 and 5)	yes	yes	no	yes	yes	no
Different smallest-number intercept	no	yes	yes	yes	yes	yes
Different mean-number intercept	no	yes	yes	no	yes	yes
Different SNARC slopes	no	no	no	yes	yes	yes

Note. This table summarizes the characteristics of the six possible scenarios of RMdependency and AMdependency of the SNARC effect, which are described above and illustrated in Figures S1 and S2. The crucial distinction consists in whether dRTs, intercepts and slopes differ between the two ranges in both experiments.

The mean-number intercept that is illustrated by a dashed vertical line in Figures S1 and S2 helps distinguish the scenarios from each other. However, as can be seen in Table S1, it is not necessary to test it against zero in a Bayesian one-sample *t*-test, because the scenarios can be distinguished with the other tests. We expected to observe Scenarios 4 or 5 (for reasons, see main manuscript).

Response times in Experiment 1

The average RTs in ms per number in the two ranges from 0 to 5 and from 4 to 9 are plotted separately per response hand in Figure S3, and together for both response hands with standard errors in Figure S4.

Figure S3

Average RT in ms per number and per response hand (left: squares, right: triangles) in each range (low: blue, high: orange) in Experiment 1

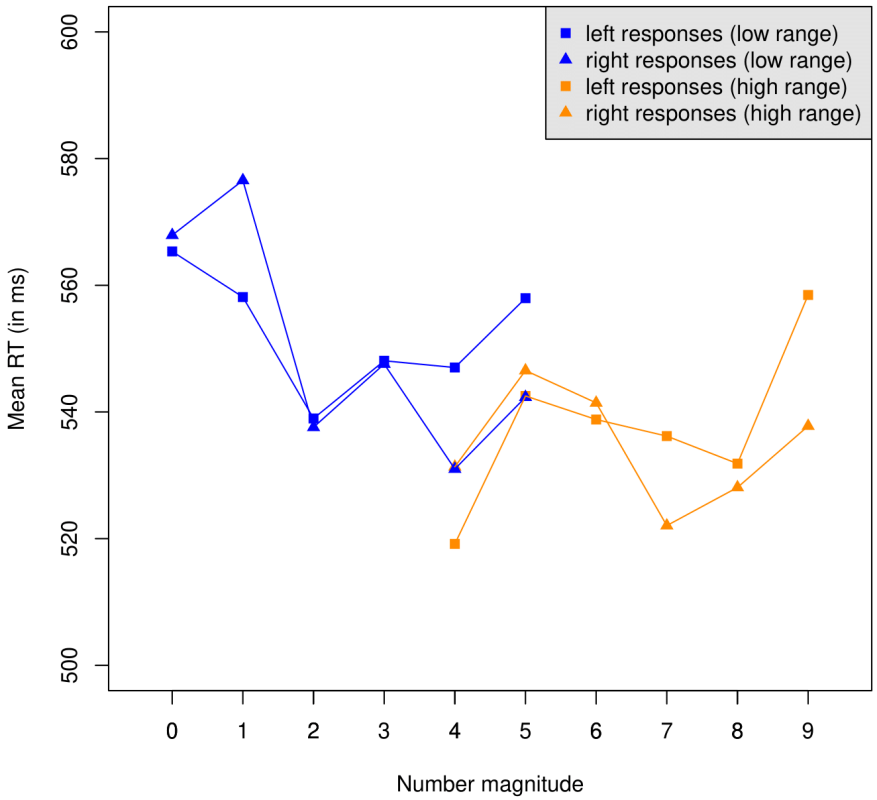
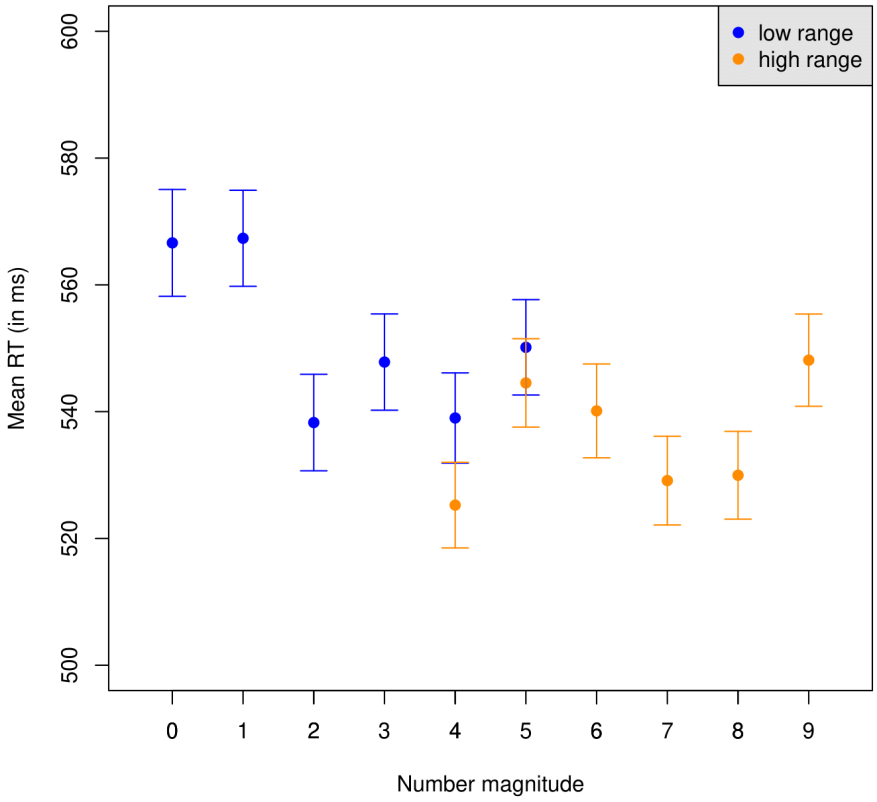


Figure S4

Average RT in ms (with with error bars indicating the respective SE) per number averaged over response hands in each range (low: blue, high: orange) in Experiment 1



Response times in Experiment 2

The average RTs in ms per number in the two ranges from 1 to 5 (excluding 3) and from 4 to 8 (excluding 6) are plotted separately per response hand in Figure S5, and together for both response hands with standard errors in Figure S6.

Figure S5

Average RT in ms per number and per response hand (left: squares, right: triangles) in each range (low: blue, high: orange) in Experiment 2

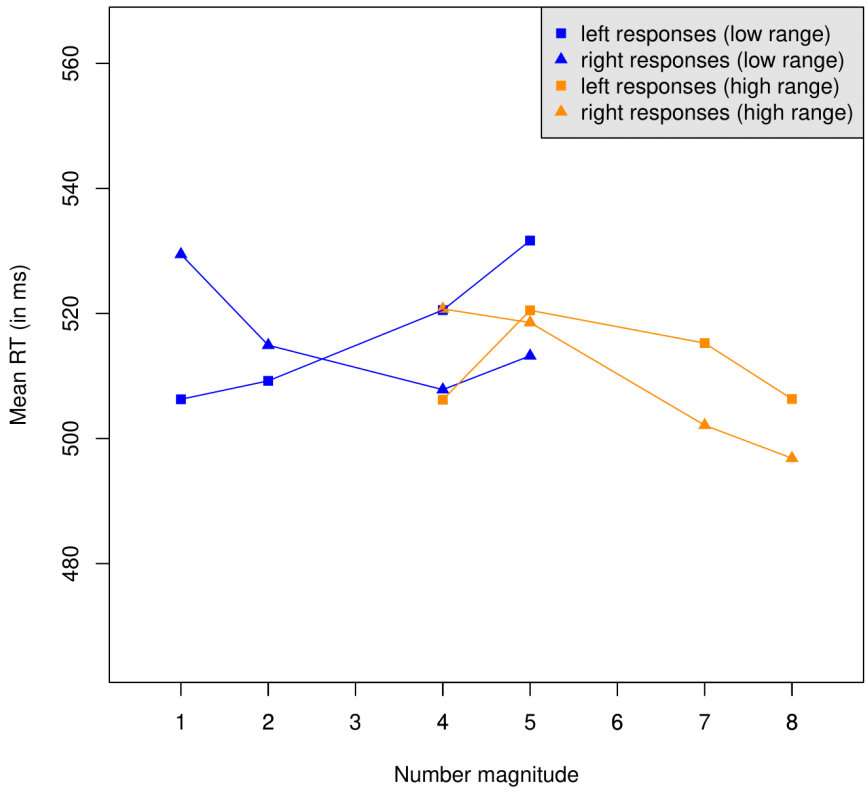


Figure S6

Average RT in ms (with error bars indicating the respective SE) per number averaged over response hands in each range (low: blue, high: orange) in Experiment 2