One and only SNARC? ~~The Flexibility of Spatial-Numerical Associations.~~

A Registered Report on the SNARC Effect's Range Dependency

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Hans-Christoph Nuerk[1,4,5*], 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:

hc.nuerk@uni-tuebingen.de

Abstract

Numbers are associated with space, but it is unclear how flexible these associations are. In this study, we will investigate whether the SNARC effect (Spatial-Numerical Association of Response Codes; Dehaene et al., 1993), which describes faster responses to small/large number magnitude with the left/right hand, respectively, is fully flexible (~~and~~ depend~~s~~ing only on relative magnitude within a stimulus set), or not (~~and~~ depend~~s~~ing on absolute magnitude as well). Evidence for relative-magnitude dependency comes 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 important conclusion was drawn solely from the absence of evidence for absolute-magnitude dependency in frequentist analysis in underpowered studies. A closer inspection of those descriptive data suggests absolute magnitude ~~might~~ also matters. Hence, we will conduct a close replication of Dehaene et al.'s (1993) Experiment 3 and a conceptual replication considering recent advances in SNARC research, investigating absolute- and relative-magnitude dependency by comparing response patterns to critical numbers, intercepts and SNARC slopes across ranges with Bayesian statistics. To achieve a ~~power~~ probability of .90 for detecting moderate evidence (Bayes Factor above 3 for Cohen's $d = 0.15$ or below 1/3 for $d = 0$)~~ for Cohen's $d = 0.15$~~, we will conduct each experiment online with maximum 800 participants~~, but run sequential analyses with~~ (optional stopping at moderate evidence). We hypothesize that both absolute and relative magnitude influence spatial-numerical associations, suggesting the SNARC effect operates on flexible and absolute number representations simultaneously.

Keywords: spatial-numerical associations, SNARC effect, mental number line, replication, online experiment~~, high statistical power~~

One and only SNARC? ~~The Flexibility of Spatial Numerical Associations.~~

A Registered Report on the SNARC Effect's Range Dependency

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 – for models 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 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. ~~, whereas with~~ 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).

The 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 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) go 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 the 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) demonstrated the 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 pooled standard deviation 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, 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 ~~power~~ calculations in both the frequentist and the Bayesian framework, see 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 et al., 2019). Typically, it is done by using the number set 1, 2, 3, 4, 6, 7, 8, 9, which importantly also excludes zero (see below).

Second, using the number zero is problematic due to its special status shown in several studies: Reading time for zero 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 zero 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 zero (Levenson et al., 2007). Using zero also turned out problematic in SNARC studies: The RTs and dRTs for the number zero 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 zero from the stimuli set (e.g., Cipora 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 zero might have confounded the results of the previous studies (see Table ~~2~~1). Therefore, in addition to the replication that we will conduct as close as possible to the original studies by Dehaene et al. (1993) and Fias et al. (1996), we will also conduct a conceptual replication using a suitable stimulus set to disentangle these potential confounds and tackle all the above-mentioned limitations.

Can the SNARC effect operate 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 in previous work (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.

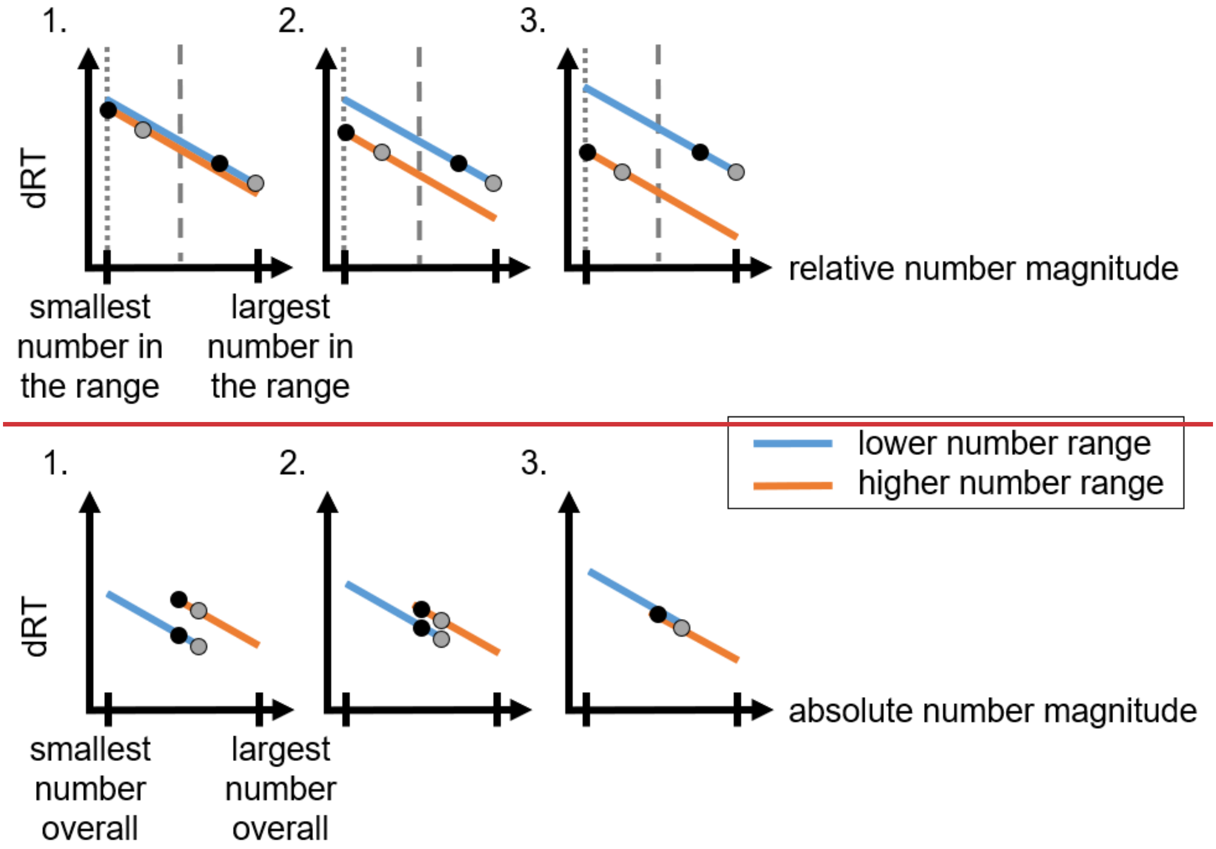
~~**How could absolute magnitude affect the SNARC effect?**~~

~~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 1 and 2 and Table 1, for detailed elaboration of these scenarios):~~

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 1

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

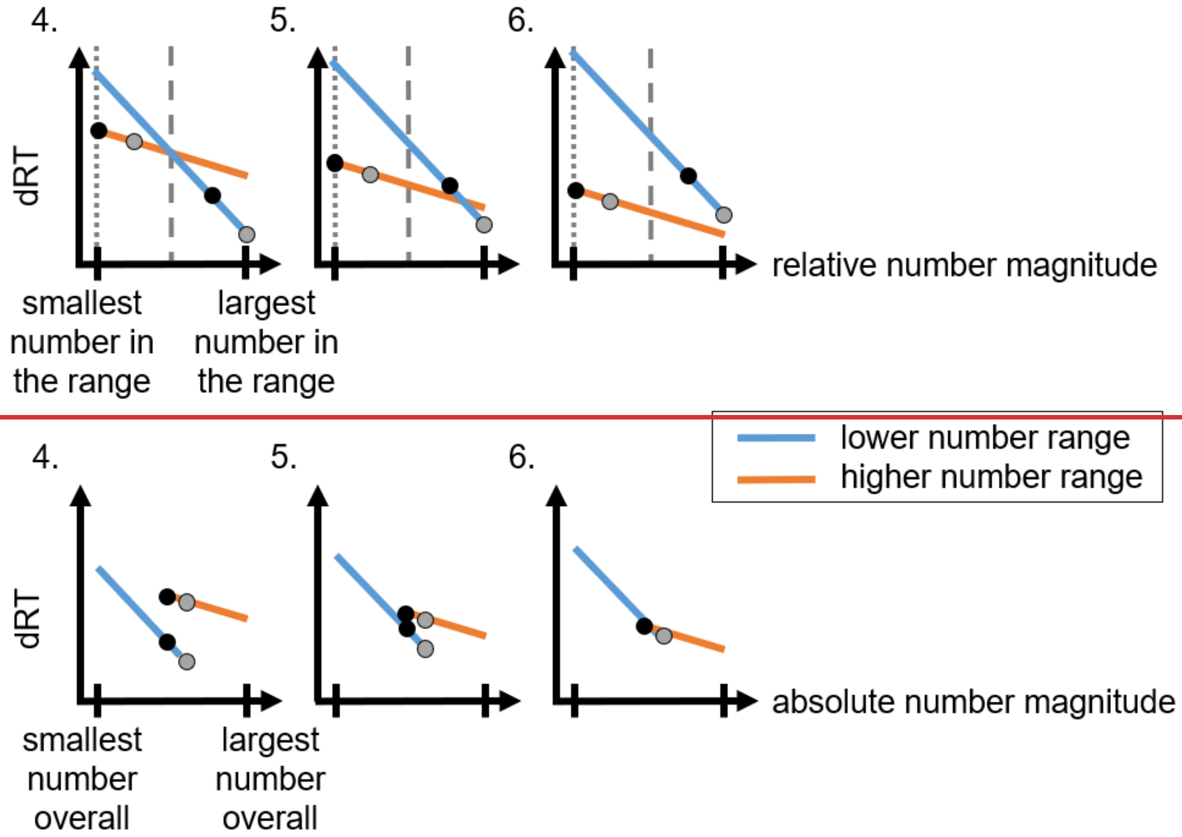


Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 1, 2, and 3, with the regression lines for the lower and higher number ranges being represented in blue and orange, respectively. In the upper part of the figure, relative number magnitudes are used for the 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 2

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 1.

Table 1

Possible Scenarios of RMdependency and AMdependency of the SNARC Effect

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 1 and 2. The crucial distinction consists in whether dRTs, intercepts and slopes differ between the two ranges in both experiments.

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 will 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 but rather complex elaboration of six possible scenarios combining different parameters of (i) and (ii), see Figures S1 and S2 in our Supplementary Material (https://doi.org/10.17605/OSF.IO/Z43PM) .

FirstTo answer the research question, we will first replicate Experiment 3 by Dehaene et al. (1993), which has also been replicated in Experiment 1 by Fias et al. (1996), where we will also use the number ranges from 0 to 5 and from 4 to 9. Second, we will conduct a conceptual replication, which is meant to address confounds due to the unequal distribution of odd and even numbers and due to the presence of zero in both stimuli sets, where we will use 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. Moreover, the critical numbers that appear in both ranges are then the same in both experiments, namely 4 and 5. Table ~~2~~1 gives an overview of the number ranges we will use and of confounds between number parity and number magnitude in Experiment 1 that will be avoided in Experiment 2.

In both of our replication experiments, a high statistical power will be 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 for lack of such differences, we will use the Bayesian instead of frequentist approach. ~~(f~~For the interpretation of different values for the Bayes Factors, we will follow the recommendations by Dienes~~,~~ (2021): A BF_{10} greater than 3 or 10 will be treated as moderate or strong evidence for the alternative hypothesis, while a BF_{10} smaller than 1/3 or 1/10 will be 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 et al., 2019; Gökaydin et al., 2018; Koch et al., 202~~1~~3). 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 expect to replicate the findings by Dehaene et al. (1993) and by Fias et al. (1996) as concerns RMdependency. However, we also expect 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. More precisely, we ~~expect to observe Scenario 4 or 5 (see Figure 2 and Table 1) and~~ hypothesize:

1. A SNARC effect in both (a) the lower and (b) the higher ~~all used~~ number ranges in each experiment. (a) The SNARC effect in the lower range serves as a manipulation check and is considered as a prerequisite for testing Hypotheses 2 and 3 in the respective experiment. Both (a) and (b) aim 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 would be 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 would be reflected by dRTs for these critical numbers being equal between ranges, and by dRTs for the smallest number in each range (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 would mean 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 would be 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.

S~~tatistical power considerations and s~~ample size ~~determination~~considerations

In this study we decided to power for defined Cohen's $d = 0.15$ as the minimal effect size of interest in Hypothesis 3, because the most crucial aim of the present study is 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 will be able to find evidence for or against the SNARC slope differences between number ranges that were descriptively reported in the original studies that we wish to replicate here. 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), but the slope difference of 2.99 ms with a pooled 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 is also valid for testing Hypotheses 1 and 2, which require one-sample *t*-tests. The reason is 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 should be large enough to find at least moderate evidence in case these hypotheses are false.

To ensure a statistical powerprobability of .90 for finding at least moderate evidence for a true underlying effect (i.e., BF_{10} greater thanabove 3, according to Dienes, 2021) for anwith a minimally relevant effect size of Cohen's $d = 0.15$ for in one-sample or paired *t*-tests, the sample needs to consist of 800 participants (for power calculations, see https://doi.org/10.17605/OSF.IO/Z43PM). The sample size of 800 participants is 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 are randomly drawn from a normal distribution around the minimally relevant effect size of $d = 0.15$ are simulated (for a similar approach, see Kelter, 2021).

Following the same procedure, we found that the sample would need 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 is smaller than the initial sample size of 200 that will be collected in the SBF+maxN approach. For ~~this~~ 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 deviation in our previous color judgment experiments were only $SD = 4.2$ ms and $SD = 3.9$ ms. Hence, our calculations are rather conservative, and the ~~statistical power~~probability to find evidence for a true underlying effect thus is 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, low rates of misleading evidence (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 will make use of the "Sequential Bayes Factor with maximal n" (SBF+maxN) approach as described by Schönbrodt & Wagenmakers (2018) and define an optional stopping threshold to make our data collection more efficient. Namely, we use moderate evidence in favor of ~~Hypothesis 3~~all hypotheses ($BF_{10} > 3$) or against ~~it~~ them ($BF_{10} < 1/3$) as thresholds. More precisely, for each experiment, we will first recruit 200 participants and compute 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} does not reach any of the two thresholds ~~yet~~for all hypotheses, we ~~will~~ want to collect another 20 datasets and recalculate the BF_{10} ~~until we reach moderate evidence~~. If no threshold is 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 will stop the sequential recruiting of participants in any case.

Participants

For each experiment, adults will be recruited via the recruiting platform Prolific. To comply with our ethics proposal, they must 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 will be conducted in English, participation is only possible for native English speakers (as per Prolific's screening based on self-reports). Participation will take approximately 20 minutes and will be compensated with £5 (partial payment for partial participation).

Design and experimental task

In the parity judgment task with binary response-key setup, participants will have to indicate as fast and as accurately as possible whether the number presented on the screen is 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 will assign 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 will be used in the lower number range and the numbers from 4 to 9 will be used in the higher number range. In Experiment 2 (conceptual replication), the numbers from 1 to 5 (excluding 3) will be used in the lower and the numbers from 4 to 8 (excluding 6) in the higher number range, eliminating confounds between number parity and number magnitude (see Table ~~2~~1) and special influences of zero.

Table ~~2~~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 gives an overview of the stimulus set we will use 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 will avoid it in both number ranges of Experiment 2, such that number parity and number magnitude are uncorrelated (i.e., they are orthogonal to each other as predictors in regression models). Number parity is typically contrast-coded with –0.5 for odd and +0.5 for even numbers when measuring the MARC effect. The number 0 is included in Experiment 1, but we will not use 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, are present in each of the number ranges.

In both experiments, we will use 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 leads to a total of 600 trials for Experiment 1 and 400 trials for Experiment 2. In each experiment, the trials will be equally divided into four blocks (one per combination of number range and response-key assignment), and a break of minimum 30 seconds must be taken between them. Participants will be randomly assigned to one of four block orders (see Figure ~~3~~1). The order of stimulus presentation within blocks will be fully randomized. Each trial will start with a square (extended ASCII 254 with the font size 72px) as eye fixation point (300 ms). Then the number (Open Sans font, size 72px) will be presented until a response is given. A blank screen (500 ms) will conclude the trial. Stimuli as well as fixation squares will be presented in black (0, 0, 0 in RGB notation), while the background remains gray (150, 150, 150 in RGB notation). The time course of an exemplary trial is illustrated in Figure ~~4~~2. Each block will be preceded by a short practice session in which each number will be presented twice (i.e., 12 practice trials before each block in Experiment 1 and eight practice trials before each block in Experiment 2, respectively). Accuracy feedback will appear during practice sessions only.

Figure ~~3~~1

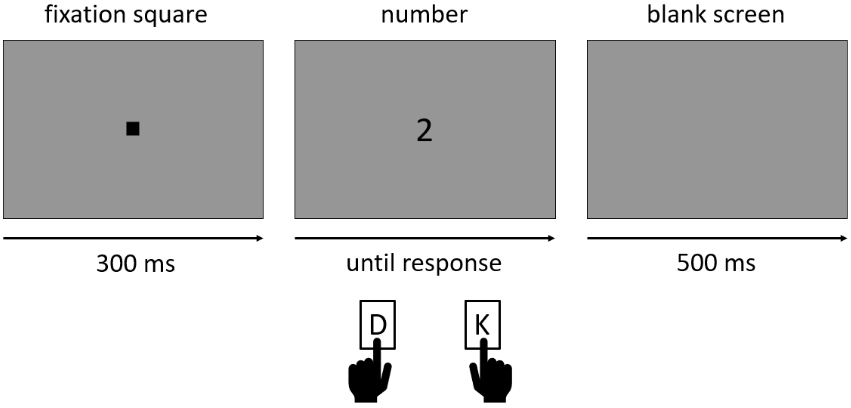
Counterbalancing block orders in Experiments 1 and 2

	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 will be preceded by two repetitions per number as practice trials (12 trials for Experiment 1 and eight trials in Experiment 2), consist of 25 repetitions per number as experimental trials (150 trials for Experiment 1 and 100 trials in Experiment 2) and be followed by a break.

Figure 42

Time course of an exemplary trial



Procedure

The experiments have been 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 (for preregistrations, see https://doi.org/10.17605/OSF.IO/F2GB8, and https://doi.org/10.17605/OSF.IO/VBA7N) 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) will be applied and participants will be asked whether they want to participate seriously. Participants will be asked to take part only if they wish to give their informed consent, if they use a desktop computer or laptop, and if they are between 18 and 40 years old. Then, participants will be 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*, or *does not* *know or no preference*; and stability: *always, usually, does not know or no preference*; in order to replicate findings by Hohol et al., 2022). 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 may choose response keys for the experimental task which are to be located in the same row and about one hand width apart from each other on their keyboard. Then, instructions will be displayed before, and the first block of the experimental task will start with its practice trials. For instance, the instructions will be 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 will be asked to self-rate their math skills compared to people of their age on a visual analogue scale from *very bad* to *very good*. Next, data quality will be 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 will ask participants whether they have 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 will be 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 will equip 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 will restrict participation to devices over 600 pixel screen width. In addition, to exclude multiple submissions we will perform checks based on User-Agents and IP addresses during data evaluation.

Data preprocessing

We will use the same analysis pipeline as in another of our studies, except for not applying any color vision check (for preregistrations, see https://doi.org/10.17605/OSF.IO/F2GB8, and https://doi.org/10.17605/OSF.IO/VBA7N). 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. Specifically, only datasets of participants who indicate to be between 18 and 40 years old and to seriously participate will be analyzed. Datasets will be excluded if participants describe their environment as very/extremely noisy, if they report multiple major distractions, or if they report that they were not using their left/right index finger for the left/right response key, respectively. Practice trials and incorrectly answered trials will not be analyzed in the main analysis. Only trials with RTs of minimum 200 ms will be 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 will be 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. Further outliers will be removed in an iterative trimming procedure for each participant separately, such that only RTs that are maximum 3 SDs above or below the individual mean RT of all remaining trials will be considered. This procedure permits to exclude RTs that are unlikely for each given participant and accounts for the right-skewed distribution of RTs, where the means would otherwise be largely overestimated. Only datasets of participants with at least 75% valid remaining trials will be included in the analysis. Finally, only datasets of participants without any empty experimental cell (number magnitude per response side) in both number ranges will be considered, because an empty cell causes a missing dRT, which in turn makes the calculation of the SNARC slope problematic.

Data analysis

All data analyses will be performed in the statistical computing software R (R Core Team, 2022). An overview of all hypotheses, corresponding tests, and interpretations of possible outcomes is given in the Study Design Table. Instead of frequentist analysis, we decided to take the Bayesian approach. For this, we will determine the BF_{10} associated with the corresponding Bayesian *t*-test to obtain evidence for both null and alternative hypotheses (using the R package *BayesFactor* by Morey et al., 2015, with a default *r*-scale of 0.707 as uninformed prior using Cauchy distribution). More specifically, we will calculate Bayesian *t*-tests and extract the respective BF_{10}. Importantly, considering a BF_{10} larger than 3 as evidence against the null hypothesis is more conservative 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 will apply the SBF+maxN approach for sequential data analysis with optional stopping in case of at least moderate evidence for or against ~~Hypothesis 3~~all hypotheses.

The key dependent variable will be the mean difference between RTs of the right hand minus left hand (dRT), which will be calculated for each number separately per participant and per number range. RTs will be 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). One regression will be calculated for each participant and for each number range. Our first dependent measure will be 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 will calculate smallest-number intercepts and mean-number intercepts (when relative magnitude of the numbers in both ranges is 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 are part of both number ranges (i.e., 4 and 5). An overview of how the following hypothesis tests can help 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 Table 1 Figures S1 and S2 and Table S1 (see Supplementary Material).

First, to we will test 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 in terms of 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) will be used as manipulation check and prerequisite for following investigations (Hypotheses 1b, 2 and 3). The obtained, SNARC slopes will be 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) and accounts for the within-subject design. Although we do not expect reversed, but instead regular SNARC effects reflected by negative slopes (as in each of the six scenarios described above and shown in Figures 1 and 2 and Table 1), we will want to use two-sided tests here to stay consistent within this study. Evidence for the SNARC effect in all ranges would replicate findings from the two seminal 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 finding evidence against it in one of the four ranges would speak against its robustness, but we consider this to be is highly unlikely because the SNARC effect in parity judgment has been shown in numerous studies using different number ranges within the interval from 0 to 9. 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 will test whether dRTs for critical numbers (i.e., 4 and 5) differ 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 not fully flexible to some degree (as in Scenarios 3 and 6 in Figures S1 and S2).

ThirdNext, to test AMdependency of the number mapping on the MNL, we will test whether the smallest-number intercepts differ 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 flexibleRMdependent. Evidence against a difference would indicate that the SNARC effect and the MNL are at least partly ~~flexible~~ RMdependent (as in Scenarios 1 and 4 in Figures S1 and S2).

~~Fourth~~Third, to investigate AMdependency of the strength of the SNARC effect, we will compare 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 a such difference would indicate that the strength of the SNARC effect ~~and of the spatial mental representation~~ does not differ between number ranges (as in Scenarios 1, 2, and 3 in Figures S1 and S2). Once the data is collected, results can be interpreted with the help of Table S1 to see which scenario most likely underlies the mental representation of number magnitude.

~~Positive controls and m~~Manipulation checks

To control the data quality in our study, we have 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 will only analyze trials that reflect mental processes in correctly executed parity judgment, we will exclude incorrectly answered trials and trim RTs (as described in the data preprocessing pipeline). Also, we will exclude full datasets of participants with less than 75% valid trials to only build our results on participants who have understood and followed the task instructions. Moreover, we assess whether participant comply with the instructions to use their left and right index fingers for the left and right response keys, respectively, and only include their datasets into our analysis if they comply with the instructions. Finally, the test of the SNARC effect in the lower number ranges (Hypothesis 1a) will serve as a manipulation check. Importantly, we will only proceed with the testing of other hypotheses if we can find the SNARC effect in the lower number range in the respective experiment.

~~Last, we will check for the presence of the Odd Effect (Hines, 1990; i.e., overall faster reactions to even than to odd numbers, irrespective of the response side). The Odd Effect is quite robust in the parity judgment task, but independent from the SNARC effect (as it is independent from number magnitude and from its mapping onto space and only considers parity). Therefore, we can consider its investigation as a manipulation check, and in case of its presence we will have a positive control for our experiment. For this, we will subtract the average RT for even numbers from the average RT for odd numbers per participant and test the differences (one per participant) against zero in two-sided Bayesian one-sample *t*-tests (one per number range, with positive estimates indicating the Odd Effect).~~

Possible limitations and unexpected outcomes

Finding evidence against the SNARC effect in ~~one of~~ the ~~four~~ lower ranges (Experiment 1: 0 to 5~~ and 4 to 9~~; Experiment 2: 1 to 5 [excluding 3]~~ and 4 to 8 [excluding 6]~~) would be an unexpected outcome which we would not have any explanation for. However, because the SNARC effect in the parity judgment task has been shown in plenty of studies (including online setups) using different number ranges within the interval from 0 to 9 and because our large sample and a high number of repetitions ensure a high ~~statistical power~~ probability to ~~detect~~ find evidence even for small effects, it seems highly unlikely not to observe a SNARC effect in ~~every of the four~~ the lower ranges. We therefore chose to use the presence of the SNARC effect in the lower ranges as a manipulation check and prerequisite for further hypothesis tests. ~~In any case, all further hypothesis tests will be meaningful even if the SNARC effect is not found in all ranges.~~

Even though our Experiment 1 aims to be 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 increase statistical power and measurement precision (Luck, 2019); second, we follow methodological recommendations (Cipora & Wood, 2017); and third, we ensure the comparability with our conceptual replication in Experiment 2. However, because of this methodological improvement, our experiment is therefore strictly speaking not a direct replication.

Just as the original two experiments, our Experiment 1 would have the limitation of the MARC effect being confounded with the SNARC effect because number parity and number magnitude are not orthogonal predictors in the regression model. Therefore, we can only calculate the MARC effect for the data resulting from our Experiment 2. Moreover, because of the special features and an irregular mental representation of the number zero, including it in the stimulus set could drive responses in our Experiment 1. However, we tackle these limitations in our Experiment 2 by using another stimulus set.

Further procedure

Data collection is estimated to last less than one month. Data analysis is expected to be finished within two months after data collection. We plan to write up the full article within three further months for the stage 2 submission.

Data and code availability

Anonymized data and analysis scripts will be available via the Open Science Framework (https://doi.org/10.17605/OSF.IO/Z43PM).

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, U.-D. Reips, L. Roth; *Funding Acquisition*: K. Cipora, H.-C. Nuerk, U.-D. Reips.; *Investigation*: K. Cipora, H.-C. Nuerk, 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, 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, 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) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips supporting Lilly Roth as well as John Caffier, with the assistance of Krzysztof Cipora as a cooperation partner. The authors would like to thank Sebastian Sandbrink for English proofreading of this Registered Report.

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

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

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

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

~~Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory & Cognition*, 18, 40-46. https://doi.org/10.3758/BF03202644~~

Hohol, M., Wołoszyn, K., & Cipora, K. (2022). No fingers, no SNARC? Neither the finger counting starting hand, nor its stability robustly affect the SNARC effect. *Acta Psychologica, 230*, 103765. https://doi.org/10.1016/j.actpsy.2022.103765

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

~~Koch, N., Huber, J., Lohmann, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2021). Mental number representations are spatially mapped both by their magnitudes and ordinal positions. PsyArXiv. https://doi.org/10.31234/osf.io/p89h3~~

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

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. (2015). BayesFactor: Computation of Bayes factors for common designs (version 0.9.12-4.4). 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. O. 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

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

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

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

PCI Study Design Table: **~~How Flexible are Spatial-Numerical Associations?~~One and only SNARC?**

A Registered Replication Report on the SNARC Effect's Range Dependency by L. Roth, J. Caffier, U.-D. Reips, H.-C. Nuerk, and K. Cipora

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
Can a SNARC effect be observed in all number ranges?	*Hypothesis 1 (and manipulation check)*: A robust SNARC effect is expected in ~~all used~~ both (a) the lower and (b) the higher number ranges, i.e., we expect to find at least moderate evidence for 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	To reach the desired ~~statistical power~~ probability of .90 for finding moderate evidence in favor of a true underlying effect (i.e., BF_{10}* > 3) with ~~for~~ an effect size of Cohen's *d* = 0.15 in two-sided Bayesian one-sample *t*-tests or in two-sided Bayesian paired *t*-tests, 800 participants need to be tested (for power calculations and sample size estimations, see https://doi.org/10.17605/OSF.IO/Z43PM). The required sample size for	Four regressions of dRTs on number magnitude followed by f~~F~~our two-sided Bayesian one-sample *t*-tests of SNARC slopes against zero ~~in each number range separately~~ (Experiment 1: 0 – 5 and 4 – 9; Experiment 2: 1 – 5 [excluding 3] and 4 – 8 [excluding 6])	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 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	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) and higher (~~our h~~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).	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 Figure 1 and Table 1 in the supplementary materials: https://doi.org/10.17605/OSF.IO/Z43PM~~manuscript~~). We therefore expect to find at least moderate evidence for it in all four number ranges. Finding at least moderate evidence against

	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, 2 and 3).	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 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		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 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		the SNARC in any of the four number ranges would be highly surprising ~~given that it is a robust effect in the parity judgment task~~, 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 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		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)		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	Evidence for RMdependency[1] would indicate flexibility of the MNL[3], such that its resolution adapts to the context and that relative magnitude

	range, respectively. We hypothesize RMdependency[1] (and possibly AMdependency[2] as well, see ~~below~~Hypothesis 2b) of the number mapping on the MNL[3].	after each added 20 participants. In case the BF_{10} reach~~es the~~ a threshold of 1/3 or of 3 (i.e., moderate evidence for or against Hypotheses 1, 2, and 3~~the null hypothesis~~) before getting to the sample size of 800 participants, we will stop recruiting earlier.	(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.)	find at least moderate evidence in case Hypotheses 1 and 2 are false.	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].	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. 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 Figure 1.
Does the mapping of numbers on the MNL[3] depend on	*Hypothesis 2b:* A left-/right-hand advantage could be		Two two-sided paired Bayesian *t*-tests of smallest-		Finding moderate or even strong evidence for	Evidence for AMdependency[1] would indicate that

whether they are small vs. high numbers in absolute terms?	observed for small/large numbers in absolute terms, respectively (on top of RMdependency[2], see ~~above~~Hypothesis 2a). However, we cannot derive any clear hypothesis from the literature about whether dRTs are lower for the smallest number in a higher than in a lower range (as observed by Dehaene et al., 1993, but not by Fias et al., 1996).		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 range of the respective experiment, see Hypothesis 1a, which serves as a manipulation check.)		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 against different smallest-number intercepts would indicate RMdependency[2] of the number mapping on the MNL[3].	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 a role as well (see above). Evidence for RMdependency[2] would indicate that the MNL[3] is at least partly flexible.
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		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	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

			evidence for a SNARC effect in the lower number range of the respective experiment, see Hypothesis 1a, which serves as a manipulation check.)		number range than in the other.	numbers (as in Scenarios 4, 5, 6 in Figure 1). If the SNARC effect does not differ between 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).
~~*Positive control* or *manipulation check*: Can we observe the Odd Effect (Hines, 1990), irrespective of the response side?~~	~~We expect to observe the Odd Effect, that is at least moderate evidence for the differences in RTs between odd and even numbers (i.e., average RT for odd numbers minus average RT for even numbers per participant) to be positive.~~		~~Four two-sided Bayesian one-sample *t*-tests of differences in RTs between odd and even numbers against zero for each number range separately (Experiment 1: 0 – 5 and 4 – 9; Experiment 2: 1 – 5 excluding 3 and 4 – 8 excluding 6)~~		~~Finding moderate or even strong evidence for the Odd Effect would provide evidence for our hypothesis and be in line with results from previous studies. What is even more, we consider this as a *positive control* or *manipulation check*, such that evidence for the Odd Effect would indicate that~~	~~The Odd Effect is quite robust in the parity judgment task, and we therefore expect to find at least moderate evidence for it in all four number ranges. Finding at least moderate evidence against the Odd Effect in any of the four number ranges would be highly surprising.~~

					~~response patterns we observe are typical for the parity judgment task.~~	

[1]**RMdependency [Relative-Magnitude dependency]**: The SNARC effect dynamically adapts to the stimulus set used in the task and is determined by the relative magnitude of the numbers within the set.

[2]**AMdependency [Absolute-Magnitude dependency]**: The SNARC effect depends on the absolute magnitude of the numbers.

[3]**MNL [Mental Number Line]**: The MNL has been proposed as the first explanation for the SNARC effect.

* The BF_{10} is the Bayes Factor defined as the probability of the obtained data under the alternative hypothesis compared to their probability under the null hypothesis. A resulting BF_{10} greater than 3 or 10 will be treated as moderate or strong evidence for the alternative hypothesis compared to the null hypothesis, respectively, while a resulting BF_{10} smaller than 1/3 or 1/10 will be treated as moderate or strong evidence for the null hypothesis compared to the alternative hypothesis, respectively (Dienes, 2021).

One and only SNARC? A Registered Report on the SNARC Effect's Range Dependency - Power simulations and sample size estimations

Lilly Roth

Version 3: 20th November 2023

This script provides all power calculations that we have run for our Registered Report on the flexibility of spatial-numerical associations and the SNARC's range dependency. It includes Monte-Carlo power simulations from Wickelmaier (2022) for the power-determination analysis and for the effect-size sensitivity approach described by Giner-Sorolla et al. (2020) within the frequentist framework applied to the two seminal studies by Dehaene et al. (1993) and Fias et al. (1996).

Moreover, in parallel to power simulations from the frequentist framework, we have run simulations for the probability to find at least moderate evidence in favor of a true underlying effect and against a truly absent effect within the Bayesian framework to calculate the sample sizes required for our study. We calculated Bayes Factors with the R package *BayesFactor* by Morey et al. (2015, https://CRAN.R-project.org/package=BayesFactor) for all relevant *t*-tests.

At the end of this script, we provide an illustration of the probability of evidence for a true underlying effect depending on the used sample size and the true effect size within a plot.

This script was created with the R packages *rmarkdown* by Allaire et al. (2023) and *knitR* by Xie (2023). This script (Version 3: November 5th, 2023) as well as the two previous versions (Version 2: July 17th, 2023, and Version 1: November 28th, 2022) can be downloaded from https://doi.org/10.17605/OSF.IO/Z43PM.

```
rm(list = ls())
library("rmarkdown")
library("knitr")
library("BayesFactor")
set.seed(123)
```

Power simulations for the original studies

Most parameters were taken from from Fias et al. (1996): sample size (**n.Fias**), standard deviations for SNARC slopes in the low (0, 1, 2, 3, 4, and 5: **SD.Fias.low**) and high (4, 5, 6, 7, 8, and 9: **SD.Fias.high**) number ranges, and standard deviation of the SNARC slope difference between both ranges (**SD.Fias.diff**)

For the missing parameter of Pearson product-moment correlation within participants between two blocks, we chose $r = .05$. We observed this value in our two SNARC automaticity experiments in color judgment tasks (for preregistrations, see https://doi.org/10.17605/OSF.IO/F2GB8 and https://doi.org/10.17605/OSF.IO/VBA7N), which was surprisingly low and might be higher in the parity judgment task. We prefer to take a rather conservative value not to overestimate the power. Please note that if the correlation turns out to be higher, the standard deviation is lower (see formula for **SD.Fias.diff** below), so that the corresponding power will be higher than estimates provided here.

```r
r <- .05
n.Fias <- 24
SD.Fias.low <- 15.1
SD.Fias.high <- 11.2
SD.Fias.diff <- sqrt(SD.Fias.low^2 + SD.Fias.high^2 - 2*r*SD.Fias.low*SD.Fias.high)
obs.diff.Fias <- 7.19-10.18 # will be used for power-determination analysis
obs.diff.Dehaene <- 10.9-20.1 # lacking data for power-determination analysis
rep_n <- 5000 # number of repetitions for our simulations
```

Power-determination analysis (Giner-Sorolla et al., 2019)

applied to Fias et al. (1996)

Given the sample size used by Fias and colleagues (i.e., 24 participants), what is the power to detect a given population effect size (e.g., a difference of 10, 5, or 1 in the SNARC slopes)?

Note that we run the following calculations both within the *Bayesian framework*, to ensure comparability between the simulations for the original studies and for our current study, and within the *frequentist framework*, because the original studies were run within the frequentist framework.

```r
BF.10 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = 10, sd = SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 10
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.10.Bayes <- round(mean(BF.10 > 3), 3)

pval.10 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean=10, sd=SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 10
  t.test(d, mu=0, alternative = "two.sided", conf.level = .95)$p.value
})
power.10.freq <- round(mean(pval.10 < .05), 3)
```

0.553 probability to find moderate evidence (BF10 > 3, Bayesian framework) and 0.731 power to detect a significant effect (p < .05, frequentist framework) for a difference of 10 in the SNARC slopes (i.e., increase of right- hand advantage in ms per magnitude unit) between ranges in a *t*-test with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

```r
BF.5 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = 5, sd = SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 5
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.5.Bayes <- round(mean(BF.5 > 3), 3)

pval.5 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean=5, sd=SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 5
  t.test(d, mu=0, alternative = "two.sided", conf.level = .95)$p.value
})
power.5.freq <- round(mean(pval.5 < .05), 3)
```

0.132 probability to find moderate evidence (BF10 > 3, Bayesian framework) and 0.24 power to detect a significant effect (p < .05, frequentist framework) for a difference of 5 in the SNARC slopes (i.e., increase of right- hand advantage in ms per magnitude unit) between ranges in a *t*-test with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

```
BF.1 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = 1, sd = SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 1
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.1.Bayes <- round(mean(BF.1 > 3), 3)

pval.1 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean=1, sd=SD.Fias.diff)
  # d = random sample of 24 differences d from normal distribution around 1
  t.test(d, mu=0, alternative = "two.sided", conf.level = .95)$p.value
})
power.1.freq <- round(mean(pval.1 < .05), 3)
```

0.022 probability to find moderate evidence (BF10 > 3, Bayesian framework) and 0.062 power to detect a significant effect (p < .05, frequentist framework) for a difference of 1 in the SNARC slopes (i.e., increase of right- hand advantage in ms per magnitude unit) between ranges in a *t*-test with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

To sum up, with the standard deviations observed by Fias et al. (1996), their sample was not large enough to find evidence for SNARC slope differences of 10 (probability of 0.553 in a Bayesian and power of 0.731 in a frequentist analysis), 5 (0.132 Bayesian and 0.24 frequentist), or 1 (0.022 Bayesian and 0.062 frequentist) between the number ranges. Note that SNARC slope differences of 10 and even 5 are rather unlikely given that the SNARC slopes themselves rarely exceed -10 and are usually closer to -5.

Effect-size sensitivity approach (Giner-Sorolla et al., 2019)

applied to Fias et al. (1996)

Given the sample size used by Fias and colleagues (i.e., 24 participants) and a desired probability/power level (e.g., 0.80, 0.90, or 0.95), what is the minimum population effect size that can be detected?

Note that we run the following calculations both within the *Bayesian framework*, to ensure comparability between the simulations for the original studies and for our current study, and within the *frequentist framework*, because the original studies were run within the frequentist framework.

For this, we started with an effect size in ms that was plausible for **mean.BF.80**, **mean.pval.80**, **mean.BF.90**, **mean.pval.90**, **mean.BF.95**, and **mean.pval.95**, ran the simulations, adapted the effect, reran the simulations, etc., until the respective desired probability/power was reached (with precision of 0.1 ms).

```
mean.BF.80 <- 12.8
BF.80 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = mean.BF.80, sd = SD.Fias.diff)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.Fias.80.Bayes <- round(mean(BF.80 > 3), 3)
```

```r
mean.pval.80 <- 11.0
pval.80 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean = mean.pval.80, sd = SD.Fias.diff)
  t.test(d, mu = 0, alternative = "two.sided", conf.level = .95)$p.value
})
power.Fias.80.freq <- round(mean(pval.80 < .05), 3)
```

0.80 probability/power to find moderate evidence (BF10 > 3, Bayesian framework) for a difference of 12.8 ms (i.e., $d = 0.698$) or to find a significant effect (p < .05, frequentist framework) for a difference of 11 ms (i.e., $d = 0.6$) in the SNARC slopes between ranges with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

```r
mean.BF.90 <- 14.6
BF.90 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = mean.BF.80, sd = SD.Fias.diff)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.Fias.90.Bayes <- round(mean(BF.90 > 3), 3)

mean.pval.90 <- 12.7
pval.90 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean = mean.pval.90, sd = SD.Fias.diff)
  t.test(d, mu = 0, alternative = "two.sided", conf.level = .95)$p.value
})
power.Fias.90.freq <- round(mean(pval.90 < .05), 3)
```

0.90 probability/power to find moderate evidence (BF10 > 3, Bayesian framework) for a difference of 14.6 ms (i.e., $d = 0.796$) or to find a significant effect (p < .05, frequentist framework) for a difference of 12.7 ms (i.e., $d = 0.692$) in the SNARC slopes between ranges with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

```r
mean.BF.95 <- 16.0
BF.95 <- replicate(rep_n, {
  d <- rnorm(n = n.Fias, mean = mean.BF.95, sd = SD.Fias.diff)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
prob.Fias.95.Bayes <- round(mean(BF.95 > 3), 3)

mean.pval.95 <- 14.1
pval.95 <- replicate(rep_n, {
  d <- rnorm(n.Fias, mean = mean.pval.95, sd = SD.Fias.diff)
  t.test(d, mu = 0, alternative = "two.sided", conf.level = .95)$p.value
})
power.Fias.95.freq <- round(mean(pval.95 < .05), 3)
```

0.95 probability/power to find moderate evidence (BF10 > 3, Bayesian framework) for a difference of 16 ms (i.e., $d = 0.872$) or to find a significant effect (p < .05, frequentist framework) for a difference of 14.1 ms (i.e., $d = 0.769$) in the SNARC slopes between ranges with $n = 24$ and $sd = 15.1$ ms for the lower and $sd = 11.2$ ms for the higher range

To sum up, with the standard deviations observed by Fias et al. (1996) and with the sample size they used, only unreasonably large SNARC slope differences (i.e., larger than typical

SNARC slopes themselves) could have been detected at probability/power levels of .80 (12.8 ms in a Bayesian and 11 ms in a frequentist analysis), .90 (14.6 ms Bayesian and 12.7 ms frequentist), and .95 (16 ms Bayesian and 14.1 ms frequentist).

Sample size estimations for the current study

Parameters for Monte-Carlo simulations

In the following, we will simulate the probability to find at least moderate Bayesian evidence for a true underlying and minimally relevant effect (BF10 > 3) and against a truly absent effect (BF10 < 1/3) for different sample sizes and different SNARC slope differences between number ranges.

r = Pearson product-moment correlation of unstandardized SNARC slopes between two blocks of around .05, as in our two SNARC automaticity experiments in color judgment tasks (for preregistrations, see https://doi.org/10.17605/OSF.IO/F2GB8 and https://doi.org/10.17605/OSF.IO/VBA7N)

```
r <- .05
```

s = standard deviation for slopes between participants in each range

In our two SNARC automaticity experiments, the standard deviations for slopes were 4.21 and 3.93, and in Fias et al. (1996), the pooled standard deviation for slopes in the lower and higher ranges was 13.29. Although we do not think that there generally is a higher variability of slopes in the parity judgment task as compared to color judgment tasks, and although our planned online study will have high measurement precision, so that we expect rather small standard deviations in the current study, we use the pooled standard deviation from Fias et al. (1996) here, which is more conservative.

For this, we started with sample sizes that were plausible for **necessary_n.H1** and **necessary_n.H0**, ran the simulations, adapted the sample sizes, reran the simulations, etc., until the respective desired probability/power was reached (with precision of 10 participants).

```
s <- sqrt( (SD.Fias.low^2 + SD.Fias.high^2) / 2 )
```

sxy = covariance of unstandardized SNARC slopes between two blocks

```
sxy <- r*s*s
```

necessary_n.H1 = sample size to detect a true slope difference

necessary_n.H0 = sample size to detect no slope difference

to be determined for the current study and to be varied for illustrating different scenarios in a plot

effect (SNARC slope difference between ranges) detected by Fias et al. (1996) and corresponding effect size as Cohen's d

```
E.Fias <- 7.19-10.18
ES.Fias <- abs(E.Fias/SD.Fias.diff)
```

mean_d = minimal effect size of interest, decision based on effect size observed by Fias et al. (1996):

```
mean_d <- 0.15
```

Sample size calculation for the current study

Probability for finding moderate evidence for a true underlying effect

```
necessary_n.H1 <- 800
BF.H1 <- replicate(rep_n, {
  d <- rnorm(n = necessary_n.H1, mean = mean_d, sd = 1)
  # as we set mean to a standardized value (Cohen's d), sd must be set to 1
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
p.samplesize.H1 <- round(mean(BF.H1 > 3), 3)
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 > 3) for the minimally relevant effect size of $d = 0.15$, 800 datasets need to be collected.

Probability for finding moderate evidence against a non-existent effect

```
necessary_n.H0 <- 180
BF.H0 <- replicate(rep_n, {
  d <- rnorm(n = necessary_n.H0, mean = 0, sd = 1)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
p.samplesize.H0 <- round(mean(BF.H0 < 1/3), 3)
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 < 1/3) against a non-existent difference in slopes between the ranges and for d = 0, 180 datasets need to be collected.

To sum up, 800 participants must be tested to reach a probability of 0.90 to find evidence for an true effect of minimally $d = 0.15$ or against a truly absent effect of $d = 0$.

Illustration of further simulations with different sample sizes and effect sizes for BF10 > 3

In the following, we simulate the probability to find moderate evidence (BF10 > 3) for various possibly underlying differences in SNARC slope (i.e., between -15 and 15 in steps of 0.25 ms per number magnitude) and with different sample sizes (i.e., 20, 40, 80, 160, 320, 640, 800). We will afterwards illustrate the simulated results in a plot.

```
difference <- seq(from = -15, to = 15, by = 0.25)
data.frame <- data.frame(matrix(ncol = 3, nrow = length(difference)))
colnames(data.frame) <- c("samplesize", "difference", "simulatedprob")

prob.simulation <- function(samplesize){
for (i in seq_along(difference)){
  data.frame$samplesize[i] <- samplesize
  data.frame$difference[i] <- difference[i]

  BF <- replicate(rep_n, {
```

```r
    d <- rnorm(n = samplesize, mean = difference[i],
               sd = sqrt(s^2 + s^2 - 2*sxy))
    extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
    })
  data.frame$simulatedprob[i] <- mean(BF > 3)
}
  return(data.frame)
}

prob020 <- prob.simulation(samplesize =  20)
prob040 <- prob.simulation(samplesize =  40)
prob080 <- prob.simulation(samplesize =  80)
prob160 <- prob.simulation(samplesize = 160)
prob320 <- prob.simulation(samplesize = 320)
prob640 <- prob.simulation(samplesize = 640)
prob800 <- prob.simulation(samplesize = 800)

prob.all <- rbind(prob020, prob040, prob080, prob160, prob320, prob640, prob800)

# setwd("...")
write.table(prob.all, sep = "\t", dec = ".", quote = FALSE, row.names = FALSE,
  file = "RegisteredReport_Study3_SNARC-Flexibility_Roth_BF-probability_v3.txt")
```

Plot: probability depending on effect size and sample size

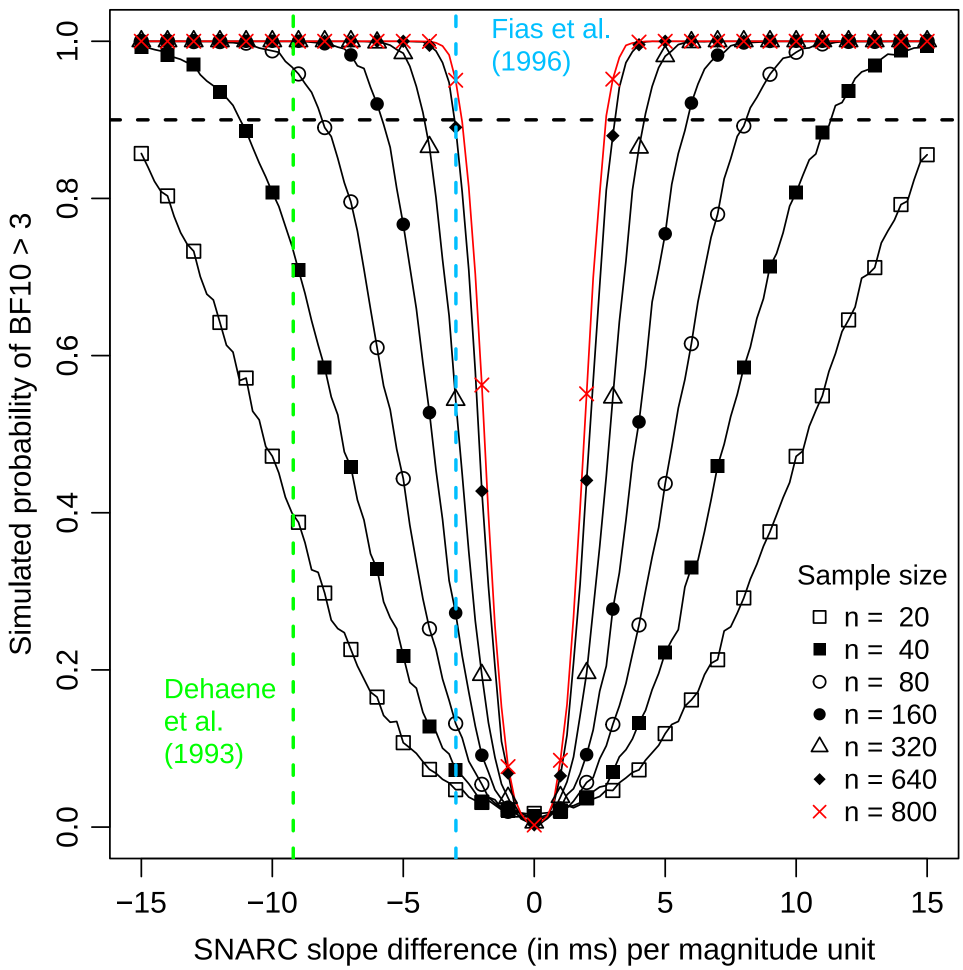
In the following, we create a plot illustrating the results we obtained in the above probability simulations for finding at least moderate Bayesian evidence for a true underlying effect (i.e., slope difference between ranges, BF10 > 3).

We also insert the range differences that were descriptively observed in the two seminal studies by Dehaene et al. (1993, green) and Fias et al. (1996, blue).

```r
rm(list = ls())
prob.all <- read.table("RegisteredReport_Study3_SNARC-Flexibility_Roth_BF-probability_v3.txt",
                       header = TRUE)

prob020 <- prob.all[prob.all$samplesize ==  20,]
prob040 <- prob.all[prob.all$samplesize ==  40,]
prob080 <- prob.all[prob.all$samplesize ==  80,]
prob160 <- prob.all[prob.all$samplesize == 160,]
prob320 <- prob.all[prob.all$samplesize == 320,]
prob640 <- prob.all[prob.all$samplesize == 640,]
prob800 <- prob.all[prob.all$samplesize == 800,]

pdf("RegisteredReport_Study3_SNARC-Flexibility_Roth_BF-probability_v3.pdf",
    height = 6, width = 6, pointsize = 13)
par(mgp = c(2, .7, 0), mai = c(.8, .8, .1, .1))

plot(simulatedprob ~ difference, prob020,
     type = "l", lty = 1,
```

```r
     xlim = c(-15, 15), ylim = c(0,1),
     xlab = "SNARC slope difference (in ms) per magnitude unit",
     ylab = "Simulated probability of BF10 > 3")

points(simulatedprob ~ difference, data = prob040, type = "l", lty = 1)
points(simulatedprob ~ difference, data = prob080, type = "l", lty = 1)
points(simulatedprob ~ difference, data = prob160, type = "l", lty = 1)
points(simulatedprob ~ difference, data = prob320, type = "l", lty = 1)
points(simulatedprob ~ difference, data = prob640, type = "l", lty = 1)
points(simulatedprob ~ difference, data = prob800, type = "l", lty = 1, col = "red")

points(simulatedprob[seq(from = 1, to = nrow(prob020), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob020), by = 4)],
       data = prob020, type = "p", pch = 0)
# only draw points for every fourth simulated point to make the graph not too crowded

points(simulatedprob[seq(from = 1, to = nrow(prob040), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob040), by = 4)],
       data = prob040, type = "p", pch = 15)
points(simulatedprob[seq(from = 1, to = nrow(prob080), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob080), by = 4)],
       data = prob080, type = "p", pch = 1)
points(simulatedprob[seq(from = 1, to = nrow(prob160), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob160), by = 4)],
       data = prob160, type = "p", pch = 16)
points(simulatedprob[seq(from = 1, to = nrow(prob320), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob320), by = 4)],
       data = prob320, type = "p", pch = 2)
points(simulatedprob[seq(from = 1, to = nrow(prob640), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob640), by = 4)],
       data = prob640, type = "p", pch = 18)
points(simulatedprob[seq(from = 1, to = nrow(prob800), by = 4)]
       ~ difference[seq(from = 1, to = nrow(prob800), by = 4)],
       data = prob800, type = "p", pch = 4, col = "red")

# desired probability level of 0.90
abline(h = 0.9, lty = 2, lwd = 2)

# difference in SNARC slopes (high range - low range) descriptively observed by
# Fias et al. (1996)
obs.diff.Fias <- 7.19-10.18
abline(v = obs.diff.Fias, lty = 2, lwd = 2, col = "deepskyblue")
# difference in SNARC slopes (high range - low range) descriptively observed by
# Dehaene et al. (1993)
obs.diff.Dehaene <- 10.9-20.1
abline(v = obs.diff.Dehaene, lty = 2, lwd = 2, col="green")

legend(x = -3.5, y = 1.08, expression("Fias et al. \n(1996)"), title = "",
       bty = "n", text.col = "deepskyblue", cex = 0.9, pt.cex = 1)
legend(x = -16, y = 0.24, expression("Dehaene \net al. \n(1993)"), title = "",
       bty = "n", text.col = "green", cex = 0.9, pt.cex = 1)

# probability of inconclusive/misleading evidence (BF10 < 3) despite true effect
```

```
abline(h = 3, lty = 2)

legend(x = 9.8, y = 0.35,
       expression("n =  20", "n =  40", "n =  80", "n = 160",
                  "n = 320", "n = 640", "n = 800"),
      title = "Sample size", pch = c(0, 15, 1, 16, 2, 18, 4),
      col = c(rep("black", 6), "red"),
      bty = "n", cex = 0.9, pt.cex = 0.9)
dev.off()
```



This plot shows the simulated probability (y-axis) to find moderate evidence (BF10 > 3) for different SNARC slope differences (x-axis) depending on sample size (20, 40, 80, 160, 320, 640, 800), while assuming a pooled standard deviation of $s = 13.29$ like in Fias et al. (1996) and a Pearson product-moment correlation of unstandardized SNARC slopes between two blocks of around $r = .05$ as in our two SNARC automaticity experiments in color judgment tasks (for preregistrations, see https://doi.org/10.17605/OSF.IO/F2GB8 and https://doi.org/10.17605/OSF.IO/VBA7N).

The sample size of n = 800 that we have determined above and we will use as maximal sample size for the SBF+maxN approach is illustrated in red.

The observed effect sizes found in the two original studies are shown as green (Dehaene et al., 1993) and blue (Fias et al., 1996) dashed vertical lines. The desired probability level of 0.90 is shown as dashed horizontal line.

To sum up, the sample sizes used by Dehaene et al. (1993) and by Fias et al. (1996) were not large enough to detect plausible SNARC slope differences. To achieve a probability of 0.90 to find at least moderate evidence for a true minimally relevant effect size of $d = 0.15$ (i.e., a high statistical power in the frequentist terminology) or against an effect in case the null hypothesis is true, we will collect data from 800 participants.

References

Allaire, J. J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2023). *Dynamic Documents for R* (Version 2.25). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/rmarkdown/index.html

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

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

Giner-Sorolla, R., Montoya, A. K., Reifman, A., Carpenter, T., Lewis, N. A., Aberson, C. L., Bostyn, D. H., Conrique, B. G., Ng, B. W., Schoemann, A. M., & Soderberg, C. (2023). *Power to Detect What? Considerations for Planning and Evaluating Sample Size.* Preprint on Open Science Framework. https://doi.org/10.31234/osf.io/rv3kw

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

Xie, Y. (2023). *A General-Purpose Package for Dynamic Report Generation in R* (Version 1.45). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/knitr/index.html

One and only SNARC?

A Registered Report on the SNARC Effect's Range Dependency

Supplementary Material

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Hans-Christoph Nuerk[1,4,5*], 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:

hc.nuerk@uni-tuebingen.de

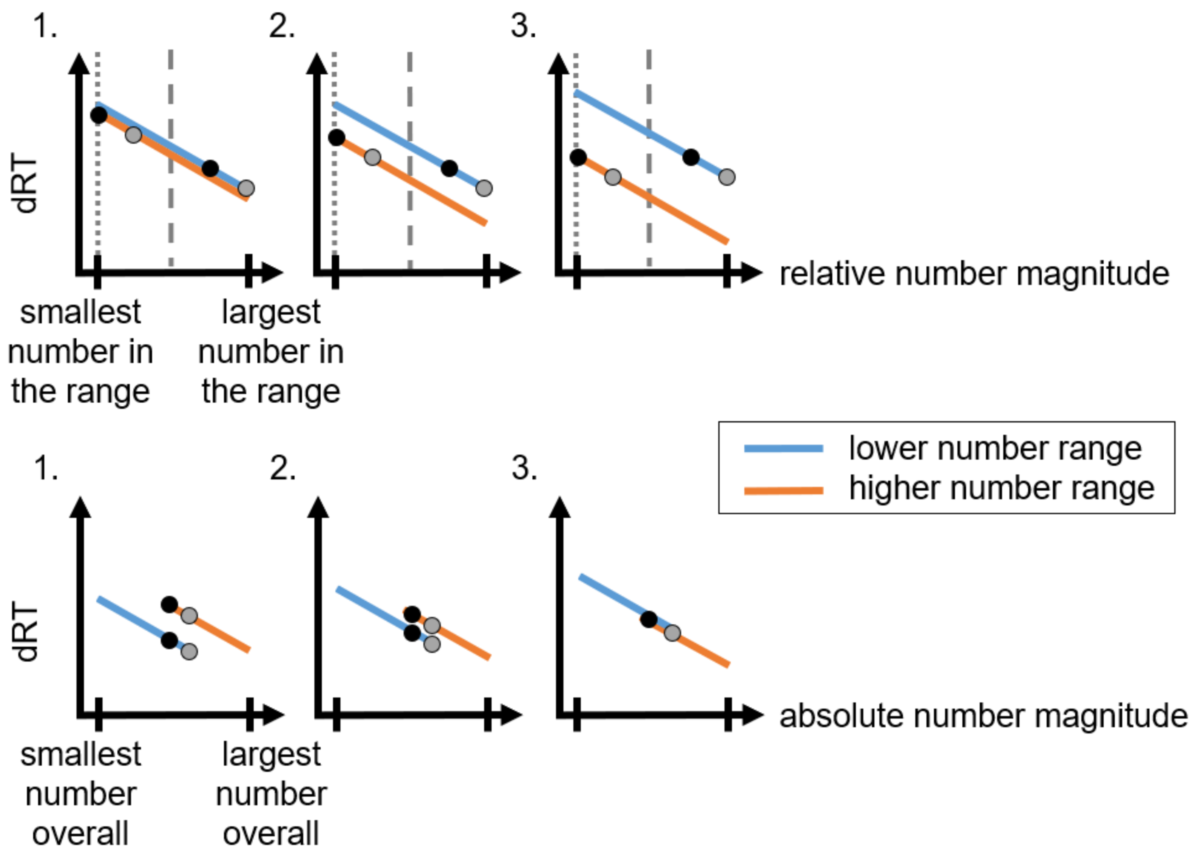
Possible scenarios of RMdependency and AMdependency

In this supplementary material, 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

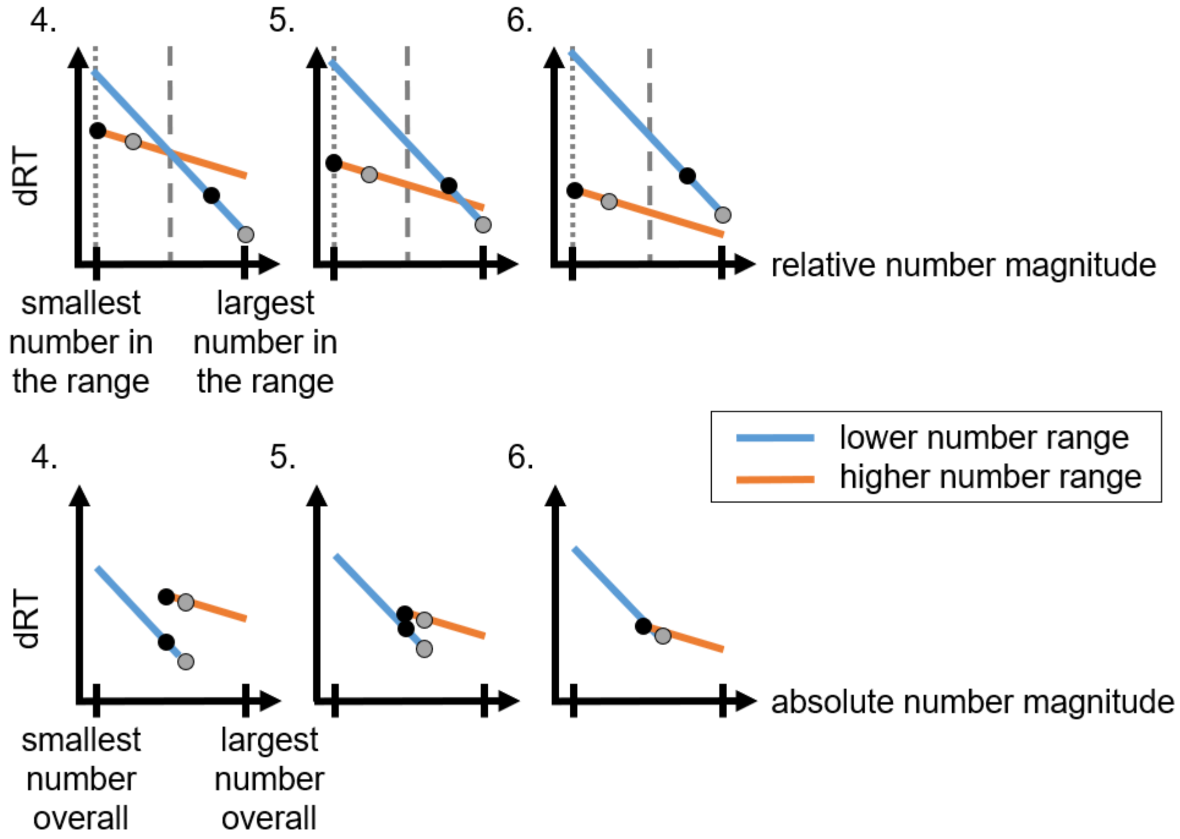


Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 1, 2, and 3, with the regression lines for the lower and higher number ranges being represented in blue and orange, respectively. In the upper part of the figure, relative number magnitudes are used for the 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. Once the data for the study is collected, results from the Bayesian hypothesis tests can be interpreted by looking at this table to see which of the six possible scenarios is most likely to underlie the mental representation of number magnitude.

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 expect to observe Scenarios 4 or 5 (for reasons, see main manuscript).