Shape of SNARC: How task-dependent are Spatial-Numerical Associations?

A highly powered online experiment

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

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

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

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

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

[5]German Center for Mental Health (DZPG)

*Corresponding author:

reips@uni-konstanz.de

Abstract

Spatial-Numerical Associations (SNAs) are fundamental to numerical cognition. They are essential for number representation and mathematics learning. However, SNAs are highly dependent on the experimental situation and task. Understanding this dependency is crucial to understanding SNAs and their impact on mathematical cognition. The hallmark SNA is the Spatial-Numerical Association of Response Codes (SNARC) effect, which denotes faster responses to small/large magnitude numbers on the left/right side, respectively (Dehaene et al., 1993). It is typically measured in magnitude classification (MC), where participants decide whether numbers from 1 to 9 (excluding 5) are smaller or larger than 5, or in parity judgment (PJ), where participants decide whether these numbers are odd or even. Despite their similarity, these tasks differ in the necessity of magnitude processing, compatibility effects being present, and other phenomena. Interestingly, the MC-SNARC seems to be categorical (i.e., same left-hand advantage between 1 and 4, and same right-hand advantage between 6 and 9), whereas the PJ-SNARC is continuous (i.e., increasing right-hand advantage with increasing magnitude). Strikingly, no matter the task, the standard analysis is a continuous linear regression, even though the MC-SNARC data are usually categorical. Only few studies systematically investigate similarities and differences between MC-SNARC and PJ-SNARC, and they often lack statistical power. In this registered report, we propose a highly powered online experiment to thoroughly investigate the shape of the MC-SNARC and the PJ-SNARC as well as ~~their potential correlation and~~ task differences in a within-subjects design with up to 1700 participants.

Keywords: spatial-numerical associations, SNARC effect, magnitude classification, parity judgment, task dependency

Shape of SNARC: How task-dependent are Spatial-Numerical Associations?

A highly powered online experiment

Spatial-numerical associations (SNAs) belong to the fundamental primitives of numerical cognition (Fischer & Shaki, 2014; Toomarian & Hubbard, 2018). They have been implicated as an important underlying representation (e.g., Dehaene et al., 2003) and as means to foster numerical and arithmetic learning (Booth et al., 2008; Dackermann et al., 2017, for an overview on embodied spatial-numerical learning). SNAs can be divided in spatial-extensional SNAs, where a particular number or magnitude is related to a physical extension (i.e., larger number to larger extensions), and directional SNAs, where a particular number is associated with a particular location in space (see Patro et al., 2014). Both SNAs are important and seem to be highly dependent on the experimental situation or task (Cipora, Patro & Nuerk, 2018). Understanding such situational dependencies is key to understanding SNAs and their relation to mathematics as such (Cipora, He, & Nuerk, 2020)

SNAs can refer to explicit or implicit associations of different characteristics of numbers (e.g., cardinality, ordinality, parity) with different aspects of space, namely directions or extensions (Cipora, Haman, et al., 2020; Cipora, Schroeder, et al., 2018; Patro et al., 2014). For instance, the MARC effect (Linguistic Markedness of Response Codes; Nuerk et al., 2004) reflects the association between parity (odd/even numbers) and direction (the left/right side), respectively. The hallmark directional SNA, however, is the Spatial-Numerical Association of Response Codes (SNARC) effect, which denotes that – at least in left-to-right reading cultures – participants respond faster to small/large magnitude numbers on the left/right side, respectively (Dehaene et al., 1993). SNAs are claimed to reflect implicit and explicit mental numerical representations of numbers and processes operating on them (Cipora, Haman, et al., 2020). They therefore belong to the primitives of numerical cognition. The tendency to map characteristics of numbers onto space is considered one of the basic traits of human cognition (Cipora, Patro, & Nuerk, 2018). The SNARC effect has been replicated with stimuli in different

modalities and notations (e.g., visual Arabic numerals, visual number words, auditory number words, visual dice patterns; Nuerk et al., 2004; Nuerk, Wood, & Willmes, 2005) and in different response setups (e.g., manual responses, pedal responses, saccadic eye movements; Schwarz & Keus, 2004; Schwarz & Müller, 2006), offline as well as online (Cipora et al., 2019; Roth, Jordan, et al., 2024). The association of number magnitude and space therefore seems to be highly robust and generalizable across many settings, even though many situational modulations have been described (Cipora et al., 2018).

Importantly ~~for our study~~, the SNARC effect arises in several tasks, which – as we will outline in detail below – have major conceptual differences in the underlying semantic features of the numbers that needs to be processed or that are automatically processed. Two tasks that inquire about semantic numerical attributes of the digits/numbers themselves are by far the most frequently used to investigate the SNARC effect: (i) the magnitude classification (MC) task and (ii) the parity judgment (PJ) task (see Table 2 in Wood et al., 2008). In MC, participants judge whether numbers are smaller or larger than a reference number[1]. In PJ, participants judge whether numbers are odd or even. Although some studies have used other kinds of stimuli (e.g., dice patterns or number words in Nuerk et al.~~, Wood, & Willmes~~, 2005, and multi-digit numbers in Tlauka, 2002; Weis et al., 2018), single-digit Arabic numbers are most used in this task. In both tasks, the instruction is to respond as quickly and as accurately as possible with a left- or right-hand key to numbers presented centrally on the computer screen. Typically, symbolic numbers from 1 to 9 (excluding 5) are used as the stimulus set, with number 5 serving as the reference number in MC. In both tasks, the response-to-key assignment is flipped in the middle of the experiment, so that both left- and right-hand responses are given for each number.

[1] In the current manuscript, tasks where presented numbers are to be compared with a fixed reference number (e.g., comparing whether a presented number between 1 and 9 excluding 5 is smaller or larger than 5) are referred to as magnitude *classification*. In contrast, tasks where the reference number is not fixed but varies between trials are referred to as magnitude *comparison*.

Conceptual differences between MC and PJ

At first sight, apart from the instructions, the MC and PJ tasks seem to be similar: A semantic feature of numbers (i.e., magnitude or parity) ~~must~~ is to be categorized ~~in both tasks~~. In the current study, we will use the most common ~~experimental setup, that is, the~~ bimanual computerized setup with symbolic numbers from 1 to 9 (excluding 5) described above. One might assume that the required cognitive processes, the responses given by participants, and the arising spatial mapping of number magnitude are similar in both tasks. However, as we shall see, the SNARC effects in MC and PJ (called *MC-SNARC* and *PJ-SNARC* in the following) differ, and the relation between them remains unclear. In the following, we describe conceptual differences between MC and PJ to shed light on reasons for the different SNARC effects.

Relevance of number magnitude and number parity

Most obviously, number magnitude is directly task-relevant to MC but not to PJ. This leads to an important difference~~: as concerns the primitives of numerical cognition. Specifically,~~ Tzelgov et al. (2015) distinguish between intentional and automatic processing, where the latter means processing without conscious monitoring according to Bargh (1992). Automatic processing can be measured in tasks where the process in question is not part of the task requirements (Tzelgov, 1997). In contrast, intentional processing is supposed to reflect the task requirements.

For participants to show the SNARC effect, two representations have to be activated, namely the *magnitude* (e.g., two) or *ordinality* (e.g., the second) of a number and its directional association with space (Cipora, He, & Nuerk, 2020). ~~Both representations are primitives in PJ according to Tzelgov et al. (2015), because~~ Importantly, when judging parity~~,~~ neither the processing of magnitude nor of its directional association with space is task-relevant and intentional. Therefore, the *PJ-SNARC* is often referred to as a marker for automatic number magnitude processing in humans and for single-digit Arabic numbers to be primitives (i.e., their meaning can be holistically retrieved from memory without further processing) in Western

cultures (Tzelgov et al., 2015). This is fundamentally different for MC, where the processing of magnitude is intentional, as it is task-relevant. Therefore, the *MC-SNARC* only shows that the directional association with space is ~~a primitive for the MC-SNARC~~ automatic (as it is not needed for the response), but not automatic number-magnitude processing.

Importantly, the two tasks differ not only regarding task-relevance of number magnitude but also in the task-relevance of number parity. More precisely, parity is task-relevant in PJ while being task-irrelevant in MC. Apart from the SNARC effect, a phenomenon to be considered in the current study is the MARC effect (Linguistic Markedness of Response Codes) (Nuerk et al., 2004), which is typically observed in PJ but not in MC (see ~~Manipulation~~Replication Check 2). The MARC effect reflects faster responses to odd/even numbers on the left/right side, respectively. In a similar vein as for the SNARC effect, there are two prerequisites for the MARC effect: the processing of parity and its directional association with space. Crucially, the processing of number parity is less automatic than the processing of number magnitude (Roth, ~~Jordan~~Caffier, Cipora, et al., ~~2024~~in press) and is more consistently found when using number words rather than Arabic digits (Nuerk et al., 2004; Roettger & Domahs, 2015). The processing of number parity ~~cannot be called a primitive~~seems not to be automatic, as typically no MC-MARC is found (Cipora, 2014; Deng et al., 2018). In contrast, the directional association of parity with space can be considered automatic ~~a primitive according to Tzelgov et al. (2015)~~, because it is not required in PJ, yet a PJ-MARC can typically be observed.

Further, RTs increase when numerical magnitude increases, which is referred to as the Numerical Size Effect (Moyer & Landauer, 1967). We expect the effect to arise in both tasks (Hypothesis 3a), although it has mainly been demonstrated for two-digit numbers (Brysbaert, 1995) and single-digit numbers will be used in the present study. Moreover, we expect it ~~but~~ to be stronger in MC than in PJ (~~see~~ Hypothesis 3b) because numerical size (i.e., number magnitude) is only task-relevant in MC. Moreover, RTs increase with increasing numerical

distance between the stimulus and the reference number in MC, which is referred to as the Numerical Distance Effect (Gevers, Verguts, et al., 2006). We expect the might effect to arise in MC (Replication Check 4), but it cannot arise in PJ because there is no criterion that numbers are compared to).

SNARC and MARC compatibility

Importantly, the SNARC and MARC effects can also be considered to be compatibility effects. In line with this, (e.g., see cognitive-control account for the SNARC effect by Zhang et al., (2022) proposed a cognitive-control account for the SNARC effect. According to this account, the automatic spatial mapping of magnitude on a first processing stage is the source of the SNARC effect, as described above. However, after this first processing stage, the authors state that compatibility effects on a second processing stage make the SNARC effect observable in experiments. Specifically, the side of the response key to be pressed according to the task instructions can be compatible (e.g., stimulus "2" with instruction to respond to even numbers with the left hand in PJ) or incompatible (e.g., stimulus "2" with instruction to respond to even numbers with the right hand in PJ). The authors attribute the visibility of the SNARC effect to conflict monitoring and inhibition control processes that arise from this compatibility. Similarly, the MARC effect can be considered a result of the spatial mapping of parity and its compatibility with the response side.

Crucially, MC and PJ differ in with respect to such compatibility effects. In a typical PJ task with two blocks, one block is MARC-compatible (i.e., when the instruction is to respond to odd/even numbers with the left-/right-hand key, respectively) and the other block is MARC-incompatible. At the same time, within each block, half of the trials are SNARC-compatible (i.e., when the response to the parity of small/large numbers is assigned to the left/right, respectively) and half of the trials are SNARC-incompatible. On the contrary, in a typical MC task with two blocks, one block is SNARC-compatible (i.e., when participants are asked to respond to numbers that are smaller/larger than a reference number with the left-/right-hand

key, respectively) and the other is SNARC-incompatible. SNARC-compatible and -incompatible trials alternating within blocks, as is the case in PJ, can elicit Gratton effects (Gratton et al., 1992). In line with this, Pfister et al. (2013) reported reduced SNARC effects after SNARC-incompatible than after SNARC-compatible trials in PJ. In MC, where SNARC-compatibility is grouped by block, such trial-to-trial effects cannot occur.

Apart from trial-to-trial compatibility effects, there can be compatibility effects depending on block order. Van Galen and Reitsma (2008) observed a stronger MC-SNARC in participants who completed the SNARC-compatible followed by a SNARC-incompatible block than in participants who were administered the reverse order, whereas Bulut, Roth, et al. (in press) observed the opposite effect in one of the three tested samples, but no effect in the two remaining samples. No effect of block order on the PJ-SNARC has been found (Bulut, Roth, et al., in press; Cipora, van Dijck, et al., 2019; Roth, Jordan, et al., 2024), where each block consists of half SNARC-compatible and half SNARC-incompatible trials. For the MARC effect, influences of block order have been found in both directions. In a previous study, we found a stronger MARC effect in PJ with the MARC-incompatible-compatible order compared to the compatible-incompatible order (between-subjects design, Cipora, van Dijck, et al., 2020). In another previous study, we found the reversed pattern (within-subjects design, Roth, Jordan, et al., 2024). The difference might be attributable to the design, and as the current study will be run between-subjects ~~as~~ like the study by Cipora, van Dijck, et al. (2020), we expect the same pattern here. A stronger PJ-MARC in the MARC-incompatible-compatible order than in the MARC-compatible-incompatible order seems plausible: Participants need to familiarize themselves with PJ and overcome their natural odd-left and even-right association in the first block, while they are already familiar with PJ and can respond in line with their natural odd-left and even-right association in the second block. Participants therefore have two reasons to be slower in the first block, and the difference (i.e., the MARC effect) between blocks is

therefore especially strong in this block order. We will exploratorily investigate compatibility-order effects in both tasks in the current study (Exploratory 3).

Bae et al. (2013) and Bulut, Çetinkaya, et al. (2024) demonstrated that the response-to-key assignment in MC influences the SNARC effect in subsequently measured PJ. Specifically, they found a regular left-to-right number mapping in PJ after a SNARC-compatible MC block (i.e., small-left and large-right), but a reversed right-to-left number mapping in PJ after a SNARC-incompatible MC block (large-left and small-right). However, in these studies, participants were assigned to only one of two possible response-to-key assignments for MC. Hence, habituation or practice that spilled over from MC to PJ was unidirectional, and furthermore, no MC-SNARC could be determined.

Strength of the SNARC effect in MC and PJ

As outlined above, the processing of number magnitude is highly automatized and single-digit Arabic numbers can therefore be considered as ~~a~~ primitives in Western cultures' numerical cognition (Tzelgov et al., 2015). In contrast, the processing of number parity ~~needs to be executed intentionally, as it is not as highly~~ might be less automatized when only a semantic number feature other than parity is task-relevant (no MC-MARC found by Cipora, 2014, and Deng et al., 2018). Note that evidence has been found for both the MARC effect and the Odd effect (i.e., faster responses to odd than to even numbers, Hines, 1990) when only non-semantic features of numbers were judged (i.e., font color; Roth, Caffier, Cipora, et al., in press), reflecting automatic parity processing to some extent, but the evidence was only weak. Also, number magnitude is more often relevant in daily life than number parity. Hence, the processing of magnitude seems to be more straightforward than the processing of parity. In line with this, average responses are typically faster in MC than in PJ (Kiesel et al., 2007; Saeki & Saito, 2009; descriptively also observed by Fattorini et al., 2015; Fitousi et al., 2009; Gevers, Verguts, et al., 2006; Ito & Hatta, 2004, see also Wood et al., 2008, for a meta-analysis), which we expect to find in the current study as well (see ~~Manipulation~~Replication Check 3).

The processing of magnitude being explicitly required in MC, but not in PJ, might elicit a stronger spatial mapping in MC than in PJ. In line with this assumption, the MC-SNARC has been found to be stronger than the PJ-SNARC (Bae et al., 2009; Cheung et al., 2015; Fitousi et al., 2009; van Dijck et al., 2009). On the other hand, judgments of ~~a primitive such as~~ number magnitude ~~is~~are automatic (Tzelgov et al., 2015) and therefore ~~are~~ naturally faster than judgments of ~~a non-primitive such as~~ number parity. At the same time, the SNARC effect is typically stronger in slower responses in both MC and PJ, both within participants and on the sample level (Cipora, Soltanlou, et al., 2019, Supplementary Material, Table ST4; Didino et al., 2019, Table 3; Gevers, Verguts, et al., 2006, Figure 6). In contrast to the reasoning above, this would lead to the opposite prediction of the PJ-SNARC being stronger than the MC-SNARC, which has been observed by Georges et al. (2017), Gevers, Verguts, et al. (2006), and Ito and Hatta (2004). No difference between the MC-SNARC and the PJ-SNARC was found by Didino et al. (2019) in an independent-samples *t*-test. Taken together, we have two opposing mechanisms: (i) easier and possibly stronger processing of magnitude in MC than in PJ, which should lead to a greater magnitude-space association in MC, and (ii) longer response times in PJ than in MC, which should lead to a greater magnitude-space association in PJ. Both opposing processes seem to be valid and there is no clear picture in the literature. It remains unclear whether the SNARC effect differs in size between tasks, and we will therefore look at this in an exploratory analysis.

Further differences between MC and PJ

Several more differences may exist between MC and PJ. First, the MC-SNARC seems to more strongly involve visuospatial working memory, the PJ-SNARC seems to rely more on verbal working memory (Deng et al., 2017; Herrera et al., 2008; van Dijck et al., 2009). Second, the MC-SNARC and the PJ-SNARC might arise at different processing stages (Basso Moro et al., 2018; Xiang et al., 2022). Third, cognitive mechanisms underlying the MC-SNARC and the PJ-SNARC might differ. Namely, Prpic et al. (2016) claim that ordinality drives the SNARC

effect in *direct* tasks (e.g., MC, where magnitude is response-relevant), whereas cardinality underlies in *indirect* tasks (e.g., PJ, where magnitude is response-irrelevant). Note that Casasanto and Pitt (2019) claim that only ordinality is crucial for both direct and indirect tasks, and that Koch et al. (2023) show that order- and magnitude-related mechanisms are not mutually exclusive. Looking into these differences between the MC-SNARC and the PJ-SNARC is beyond the scope of the current study; however, the current study will provide a better understanding of the two tasks and thereby lay the groundwork for further investigations.~~, some of which will be investigated in the current study. The first is that the MC- and PJ-SNARC seem to require different working memory (WM) resources. Van Dijck et al. (2009) reported a double dissociation between the SNARC effects. The PJ-SNARC disappears under verbal WM load, but the MC-SNARC remains unchanged. At the same time, the MC-SNARC disappears under visuospatial WM load, while the PJ-SNARC remains unchanged (see Herrera et al., 2008 for the same results in MC). Similarly, Deng et al. (2017) found the MC-SNARC to increase under spatial WM load and to decrease under verbal load, whereas they found the PJ-SNARC to disappear completely under both spatial and verbal load. The second difference that cannot be tested in the present study is that the MC-SNARC and the PJ-SNARC might arise at different processing stages. On the one hand, Basso Moro et al. (2018) claim that the SNARC effect occurs at a late response-selection stage rather than at an early semantic-representation stage in both tasks. Concurrently, Xiang et al. (2022) claim that magnitude, especially, is spatially represented on the early semantic-representation stage, therefore causing stronger interference with the response key selection than parity. This debate needs to be resolved in future research. The current study will provide a solid basic understanding of the two tasks and lay the groundwork for such investigations. The third potential difference between the MC-SNARC and the PJ-SNARC might be their underlying cognitive processes. Prpic et al. (2016) claim that ordinal information drives the SNARC effect in so-called *direct* tasks (e.g., in MC), where magnitude is response-relevant, whereas cardinal information (e.g., number 2 has a magnitude

~~of 2) underlies the SNARC effect in so-called *indirect* tasks (e.g., in PJ), where magnitude is response-irrelevant (but see Casasanto & Pitt, 2019, for a different view). Note that order- and magnitude-related mechanisms are not mutually exclusive within the same task (Koch et al., 2023), but that their relative contribution might differ between tasks. However, the role of ordinality and cardinality cannot be distinguished in the present study, because numbers convey both types of information.~~

In summary, several conceptual differences exist between MC and PJ, concerning the task-relevance of number magnitude and number parity, the compatibility of the response-to-key assignment with the SNARC and MARC effects, arising numerical-cognition effects, and underlying cognitive mechanisms. However, both tasks elicit a SNARC effect, and the presence and strength of the MC- and PJ-SNARC in ~~single~~ the same participants might be related to one another, as will be discussed in the next section.

Correlation between the MC- and PJ-SNARC

After having described the similarities and differences of MC and PJ, the question arises whether the MC-SNARC and the PJ-SNARC are correlated. However, both factors at the construct level of SNAs and at the operational level of the two tasks might lead to a null correlation. First, there seem to be high fluctuations in the SNARC effect over time (Roth, Jordan, et al., 2024) that limit the maximum correlation that can be detected. Second, the test-retest reliability of the SNARC effect has been found to be poor for both MC and PJ (correlations $.22 < r < .41$; Cipora & Göbel, 2013; Georges et al., 2013; Hedge et al., 2018; Viarouge et al., 2014). The lower the test-retest reliabilities of the MC-SNARC and PJ-SNARC, the lower is also the maximally observable correlation between the two effects. Third, the split-half reliability of the SNARC effect has been found to be poor for PJ at least in some studies (correlations $.43 < r < .96$; for an overview, see Cipora, van Dijck, et al., 2019, Table 1 there). To conclude, both properties of the SNA construct and its operationalization in experimental tasks influence whether a correlation between the MC-SNARC and the PJ-SNARC will be

found. Possible reasons for a null finding could be low intraindividual stability, low reliability, or low internal consistency, whereas a high correlation between the MC-SNARC and the PJ-SNARC would lead to the conclusion that both MC and PJ reliably measure the same underlying theoretical construct.strength of the SNARC effect in these two tasks is similar within individuals. Crucially, different predictions can be derived from the literature regarding whether there is a relation between the MC-SNARC and the PJ-SNARC, lydepending on the proposed account for the SNARC effect. The first theory proposed to explain the SNARC effect was the Mental Number Line (MNL) account (Dehaene et al., 1993). It suggests that numbers are mentally represented in a spatial format in long-term memory, which is oriented from left to right in Western cultures. According to the MNL account, both MC and PJ should activate the spatial mental representation of number magnitude, as both tasks require semantic number processing. Thus, if the MNL underlies both SNARC effects, but the MNL and hence the SNARC effect differs systematically between subjects (e.g., Cipora et al., 2019), this account predicts a correlation between MC-SNARC and PJ-SNARC (Cheung et al., 2015).

Second, the dual-route model suggests that the SNARC effect arises as a compatibility effect of two routes. On the one hand, a fast unconditional route activates the response associated with the spatial preference for the stimulus, while on the other hand, a slow conditional route identifies the response required by task instructions (Gevers, Ratinckx, et al., 2006). Importantly, the dual-route model would predict a compatibility effect in MC: The fast unconditional route activates left-side responses for small and right-side responses for large number magnitudes, while, depending on the instructions, the slow conditional route activates the same (compatible) or the reversed (incompatible) responses. In contrast, expected compatibility effects in PJ are smaller and differ between trials and not within blocks. This is because the spatial mapping of number parity (i.e., MARC effect) is not as automatized as the spatial mapping of number magnitude (i.e., SNARC effect; see Roth, Jordan, et al., 2024). Therefore, number magnitude taking the fast unconditional route in PJ should not interfere

~~much with number parity taking the slow conditional route. To summarize, the dual-route model does suggest only a small correlation of the SNARC effect between MC and PJ (Didino et al., 2019).~~

~~Third, van Dijck et al. (2014) proposed a WM account to explain the SNARC effect. In this account, the SNARC effect was originally claimed to be a temporary association of numbers and space that is constructed in WM during task execution. Specifically, the WM account claims that items appearing earlier in an ordinal sequence in WM are associated with the left, while items appearing later are associated with the right. The repeated use of common ordinal sequences can lead to spatial associations of WM contents being stored in long-term memory, and WM contents can be currently activated long-term memory representations (Abrahamse et al., 2016). Crucially, different types of WM load (verbal vs. visuospatial) have been shown to increase or decrease the MC-SNARC and the PJ-SNARC in different ways (Deng et al., 2017; Herrera et al., 2008; van Dijck et al., 2009). Hence, different WM resources seem to underlie the MC-SNARC and the PJ-SNARC, which does not speak for a correlation between MC-SNARC and PJ-SNARC.~~

~~Importantly, the theories about the origin of the SNARC effect might not be mutually exclusive and may apply in different cases (e.g., see multiple coding account of Schroeder et al., 2017). In line with this, Prpic et al. (2016) claimed that ordinality drives the SNARC effect in direct tasks such as MC (i.e., WM account), whereas magnitude drives it in indirect tasks such as PJ (i.e., MNL account), although Casasanto and Pitt (2019) disagree. According to this, no correlation of the SNARC effect between tasks is expected.~~

~~To conclude, different accounts for the SNARC effect make different predictions regarding the presence of a correlation between the MC-SNARC and the PJ-SNARC.~~

Several previous studies ~~have investigated the SNARC effect in both MC and PJ in a within-subjects design and~~ did not find any correlation (correlations with 95% confidence intervals and *p*-values: r = -.02 [-0.19, 0.15] and p = .822 for Germans; r = -.08 [-0.26, 0.10]

and p = .386 for Turks; r = .10 [-0.13, 0.32] and p = .402 for Iranians, in Bulut, Roth, et al., in press2024; r = .09 [-0.18, 0.35], and p > .05= .513 in Cipora, 2014; r = 0.06 [-0.30, 0.40] and p = .744 in Didino et al., 2019; r = .18 [-0.07, 0.42] and p = .18 in Fattorini et al., 2015; r = 0.20 [-0.01, 0.39] and p = .07 in Georges et al., 2017). To our knowledge, a significant correlation has only been reported by Cheung et al. (2015; r = 0.25) and by Cipora (2014; r = .50, but only in a unimanual setup). Note that an existing weak correlation between the MC-SNARC and the PJ-SNARC despite limiting factors such as low intraindividual stability, low reliability, or low internal consistency would only be detectable in large samples. , so that most of the previously mentioned studies might have missed a true underlying correlation due to lack of statistical power.

Finding a correlation between the MC-SNARC and the PJ-SNARC speaks in favor of the MNL and dual-route accounts of the SNARC effect and rather against the WM account and the different sources depending on direct vs. indirect tasks. However, finding evidence against a correlation is not as instructive and does not necessarily lead to the opposite conclusions. Crucially, at least two more potential issues can lead to the lack of correlation between the MC-SNARC and the PJ-SNARC apart from the theories described above. First, it could simply be the case that at least one of the two tasks is not an appropriate paradigm for assessing the underlying construct and the operationalization is not valid. Second, it could be due to the high fluctuations in the SNARC effect over time (Roth, Jordan, et al., 2024) and due to poor test-retest or split-half reliabilities for the SNARC effect (Cipora, Soltanlou, et al., 2019) that no correlation is found.

To shed further light on the be able to detect a potential correlation between the MC-SNARC effect in MC and the PJ-SNARC, in the current study, we will administer both tasks to a large sample in a within-subjects design. This will enables us to test the correlation between the MC-SNARC and the PJ-SNARC in an exploratory analysis with high statistical power (Exploratory 5).

Different shapes of the SNARC effect

After having outlined the differences between MC and PJ and after having discussed the potential correlation between the SNARC effect in these two tasks, it is important to note that the shape of the SNARC effect seems to differ systematically between MC and PJ (Wood et al., 2008). While the advantage of the right hand over the left hand increases with number magnitude in a continuous manner in PJ, it seems to be categorical in MC with the same left-hand advantage for all small numbers and the same right-hand advantage for all large numbers. However, the SNARC effect in MC is often modelled as a continuous phenomenon, just as in PJ, as described in the following~~which would then underestimate the SNARC and its fit in MC tasks~~. The main aim of the current study is to thoroughly investigate the SNARC effect in the two most widely used tasks to assess it and to find out how to best statistically model the SNARC effect.

The SNARC effect is usually calculated by subtracting the mean reaction times (RTs) with the left hand from those with the right hand for each number and regressing these differences (dRTs) on magnitude as a continuous predictor in both MC and PJ. A negative regression slope reflects the increasing right-hand advantage for larger numbers and therefore the SNARC effect. To investigate whether the effect is present on group level, regression slopes (one per participant) are then tested against zero in a one-sample *t*-test (repeated-measures regression, adapted by Fias et al., 1996, based on Lorch ~~&~~ and Myers, 1990). Importantly, this analysis method is only suitable for a continuous SNARC effect, reflecting a constant increase in right-hand advantage (reflected by a constant decrease in dRT) per increase of magnitude. Hence, when participants judge whether numbers in the typically used stimulus set from 1 to 9 (excluding 5) are odd or even, the spatial mapping of extreme magnitudes such as 1 and 9 is stronger than for magnitudes closer to the mid of the stimulus set such as 4 and 6. In other words, the association with the left side is stronger for the very small number 1 than for the slightly small number 4. While the PJ-SNARC is linear, the MC-SNARC is typically

categorical (e.g., Gevers, Verguts, et al., 2006), especially in adults (van Galen & Reitsma, 2008): In a typical MC task responses are *equally* faster with the left hand to numbers from 1 to 4 and equally faster with the right hand to numbers from 6 to 9. Therefore, a stepwise model reflects the MC-SNARC better than a continuous model (as reflected by a better model fit in terms of a higher proportion of explained variance). The use of a categorical instead of a linear function for quantifying the MC-SNARC would increase the model fit at the participant level and thereby likely also the precision of the effect size estimate at the sample level.~~has important consequences: First, it correctly avoids systematic underestimation of the effect size itself and an increased likelihood of false null results, and second, it avoids a systematic underestimation of the correlations between the effect and other measures like numerical or spatial skills. This means that a sometimes diagnosed "weak or non-significant" SNARC effect and its underestimated relations to other measures might not be an attribute of the MC task but rather a result of an incorrect statistical data analysis choice.~~

Nevertheless, a linear predictor in the regression of dRTs on number magnitude remained a frequently used analysis of the MC-SNARC (Bachot et al., 2005; Bae et al., 2009; Bull et al., 2005; Cheung et al., 2015; Deng et al., 2017; Han et al., 2017; Herrera et al., 2008; Hoffmann et al., 2013; E. M. Hubbard et al., 2009; Ito & Hatta, 2004; Lohmann et al., 2018; Mourad & Leth-Steensen, 2017; Nathan et al., 2009; Pinto et al., 2021; Schiller et al., 2016; Shaki & Gevers, 2011; van Dijck & Doricchi, 2019; van Dijck et al., 2009; van Dijck et al., 2012; van Galen & Reitsma, 2008; Weis et al., 2018). As the correlation between the linear (1, 2, 3, 4, 6, 7, 8, 9) and the categorical (-0.5, -0.5, -0.5, -0.5, 0.5, 0.5, 0.5, 0.5) magnitude predictor is extremely high (~~namely,~~ r = .913), the model with a linear magnitude predictor fits relatively well both to the continuously and categorically distributed dRTs in MC and PJ (see top panel of Figure 1 in Bae et al., 2009, or Figure 1 in Nathan et al., 2009). In some studies, a two-way ANOVA including magnitude (small vs. large) and response side (left vs. right) as within-subjects factors has been used to quantify the MC-SNARC (Fattorini et al., 2015; Gevers,

Verguts, et al., 2006; Herrera et al., 2008; Hoffmann et al., 2013; Nathan et al., 2009). However, compared to that approach, the repeated-measures regression approach has several advantages (Fias et al., 1996): First, the presence of the SNARC effect is judged by a main effect instead of an interaction effect, which allows a quantification of the size of the effect in milliseconds by the slope. Second, the presence or absence of a SNARC effect can be assessed for each participant individually. A repeated-measures regression with a categorical predictor for the MC-SNARC has only been used in few studies (Bulut, Roth, et al., in press~~2024~~; Cipora, 2014; Didino et al., 2019; Fitousi et al., 2009; Georges et al., 2017; Gevers, Verguts, et al., 2006; Hohol et al., 2020; Nathan et al., 2009; Nuerk, Wood, & Willmes, 2005; Weis et al., 2018; Zorzi et al., 2012). For an overview of all mentioned studies including MC, see Table A1 in Appendix A.

Unfortunately, the suitability of the linear and categorical predictors for dRTs in MC with the stimulus set from 1 to 9 (excluding 5) was assessed by direct comparison in only a few studies. Fitousi et al. (2009) and Nathan et al. (2009) computed two separate regression models, one of which with a categorical and the other with a linear predictor, and in both studies the fit was higher with the categorical (R^2 = .904 in Fitousi et al., 2009; R^2 = .988 in Nathan et al., 2009) than with the linear predictor (R^2 = .775 in Fitousi et al., 2009; R^2 = .891 in Nathan et al., 2009). Similarly, Didino et al. (2019), Gevers, Verguts, et al. (2006) and Nuerk, Bauer, et al. (2005) ran regression analyses including both predictors, and only the categorical predictor turned out to be significant in all three studies. In a study with two-digit numbers where participants performed PJ and MC for the unit digit, Weis et al. (2018) also included both linear and categorical predictors for both unit and decade magnitude concurrently into one regression model. They found only the categorical predictors for units and decades to be significant in MC and only the linear predictors for units and decades to be significant in PJ, providing further evidence for a categorical MC-SNARC and a continuous PJ-SNARC. Importantly, because a linear model fits well even for the categorical MC-SNARC (e.g., see Fitousi et al., 2009; and

Nathan et al., 2009), ~~the power to find~~ evidence for a better fit of a categorical model in MC can only be achieved with sufficient power by using a large sample and a sufficient number of repetitions per experimental cell (resulting from the combination of each stimulus with each response hand per task). However, as outlined above, although the linear model fit to the MC-SNARC might be high in some studies, ~~using~~ the ~~more adequate~~ categorical model seems to be more adequate~~can be decisive in other studies to not underestimate the effect or its relation to covariates~~. We expect to find a better fit of the categorical model in MC (Hypothesis 1) and of the linear model in PJ (Hypothesis 2).

Explanations for the categorical MC-SNARC effect shape

The literature provides several explanations for the different shapes of the SNARC effect, depending on the task. First, numbers are typically roughly classified into small and large numbers (Banks et al., 1976; Tzelgov et al., 1992~~, as cited in Fitousi et al., 2009). Such a gross classification into smaller or larger compared to the reference number is sufficient in MC (i.e., direct task, according to Prpic et al.'s classification, 2016), and participants are not instructed to process the exact number magnitude~~), which is sufficiently precise to perform MC and might lead to the categorical MC-SNARC (Fitousi et al., 2009; Gevers, Verguts, et al., 2006). In contrast, participants are not instructed to process number magnitude at all in PJ ~~(i.e., indirect task)~~, and thus number magnitude processing is not intentional (i.e., slow conditional route according to the dual-route model by Gevers, Ratinckx, et al., 2006) but rather automatic (i.e., fast unconditional route)[2]. ~~Thus, while number magnitude is intentionally processed as either "smaller than the reference" or "larger than the reference" and categorically mapped onto space in MC, the exact number magnitude is automatically processed~~Automatic number magnitude processing seems to be more exact and continuously mapped onto space in PJ. This explanation

[2] As outlined by the stage-1 PCI-RR reviewer Peter Wühr, predictions can be derived from the dual-route model. Importantly, both the automatic and the intentional route are activated in MC, whereas only the automatic route is activated in PJ. Thus, the SNARC effect should be stronger in MC than in PJ because it results from both routes instead of only one route. Moreover, a positive correlation of the MC- and PJ-SNARC can be assumed based on the dual-route model, since both effects are (at least partly) caused by the automatic route.

is in line with the polarity-correspondence account of the SNARC effect by Proctor and Cho (2006), as well as with the application of the markedness principle to number magnitude (~~see~~ Nuerk & Schroeder, 2024; Schroeder et al., 2017). According to these two theories, the SNARC effect arises because both *large* and *right* are associated with the positive or unmarked polarity and both *small* and *left* with the negative or marked polarity. ~~While some theories of markedness (Nuerk & Schroeder, submitted) postulate a graded nature of markedness effects, it is important to note that the association in an MC task is indeed categorical, because the experimental question is usually to decide "larger" or "smaller" than 5 (rather than large or small), which requests a binary comparative decision rather than a graded representation that a number is relatively larger or smaller as in the PJ task.~~ Similarly, this explanation is compatible with the verbal-spatial account of the SNARC effect proposed by Gevers, Verguts, et al. (2006; see also Gevers et al., 2010), stating that verbal categories such as *small* vs. *large* and *left* vs. *right* are responsible for the SNARC effect. ~~Importantly, both t~~These accounts argue for an intermediate classification into small or large numbers (Santens & Gevers, 2008), and the polarities, markedness, or verbal labels are categorical rather than continuous (Bae et al., 2009), which explains the categorical MC-SNARC. Note that it is possible that the PJ-SNARC is linear in the beginning of the task and becomes categorical over the course of the task, so that the continuous shape shifts to a stepwise one. That is, participants might start classifying the stimuli into the two categories "small" and "large" in PJ as soon as they become familiar with the stimulus set because they might notice that the stimulus set consists of two single-digit number sequences (i.e., 1 to 4 and 6 to 9) separated by the missing number 5. We will investigate this possibility in the exploratory analysis.

Second, the Numerical Distance Effect ~~(Moyer & Landauer, 1967)~~ might play a role for the MC-SNARC (Gevers, Verguts, et al., 2006). ~~The Numerical Distance Effect refers to faster reactions with increasing numerical distance between the stimulus and the reference number in MC (see Manipulation Check 4, but it cannot arise in PJ because there is no criterion that~~

~~numbers are compared to).~~ For instance, if the reference number is 5 in MC with the stimulus set from 1 to 9, numbers 4 and 6 need to be processed more intensely than numbers 1 and 9 to discriminate them from number 5 (Wood et al., 2008). Hence, responses are slowest for numbers 4 and 6 and fastest for numbers 1 and 9 with this stimulus set. ~~Importantly, the Numerical Distance Effect demonstrates automatic processing of number magnitude (i.e., fast unconditional route), because it reflects performance differences in the discrimination between numbers and arises although the task instructions do not favor or disfavor the performance for specific stimuli. It thus does not build on the gross classification into smaller and larger numbers described in the previous paragraph (i.e., slow conditional route), but rather on the exact number magnitudes.~~ In combination with the finding that the SNARC effect becomes stronger with increasing RTs, the absolute values of dRTs for number magnitudes that are close to the reference are larger than dRT predictions by the linear SNARC regression slope, resulting in a categorical shape (Didino et al., 2019; Georges et al., 2017; Gevers, Verguts, et al., 2006). Importantly, the Numerical Distance Effect demonstrates automatic processing of number magnitude (i.e., fast unconditional route), because it reflects performance differences in the discrimination between numbers and arises although the task instructions do not favor or disfavor the performance for specific stimuli. It thus does not build on the gross classification into smaller and larger numbers described in the previous paragraph (i.e., slow conditional route), but rather on the exact number magnitudes.

In summary, the rationale for a categorical instead of linear MC-SNARC is twofold: First, the intentional classification into small and large numbers is categorical in MC, and second, the interaction between the Numerical Distance Effect and the positive correlation between the SNARC effect and overall RTs contributes to a step-wise shape. Since statistical models should correspond to scientific models as closely as possible (Westermann & Hager, 2017), the MC-SNARC should therefore be tested with a categorical predictor.

Influence of task order on the SNARC effect

The influence of task order (first MC and second PJ, or reversed) on the MC-SNARC and on the PJ-SNARC has been investigated in only a few studies. Didino et al. (2019) did not find any task-order effects on the SNARC effect. Fattorini et al. (2015) did not observe any task-order effect on the PJ-SNARC but found the MC-SNARC to be weaker after PJ than when MC was the first task. In most other studies including the two tasks, the effects of task order have either not been reported (Cheung et al., 2015; Gevers, Verguts, et al., 2006; Nuerk, Bauer, et al., 2005; Weis et al., 2018), or could not be calculated because task order was not counterbalanced (Bae et al., 2009; Cipora, 2014; Fitousi et al., 2009; Georges et al., 2017; Zorzi et al., 2012) or because different samples completed MC and PJ (Ito & Hatta, 2004; van Dijck et al., 2009). To our knowledge, only Bulut, Roth, et al. (in press; see Supplementary Materials) have tested the influence of task order, and they did not find an influence on the SNARC effect in any of the two tasks in any of three samples (130 German, 112 Turkish, and 75 Iranian participants). In fact, two opposite theoretical predictions can be made. On the one hand, the SNARC effect might be stronger in each task if it is~~We expect a stronger SNARC effect in both MC (Hypothesis 2a) and PJ (Hypothesis 2b) when they are~~ the second~~ task~~, because the processing of number magnitude and its spatial mapping should be stronger ~~when~~ if they have already been activated ~~with~~in a previous task. On the other hand, the SNARC effect might be weaker in each task if it is the second, because RTs typically decrease with practice and faster RTs are typically associated with a weaker SNARC effect (note that both decreasing RT and a decreasing SNARC effect over time in PJ have been found by Roth, Jordan, et al., 2024). If both mechanisms were true, they might cancel out each other and make the influence of task order invisible. Hence, we cannot make any directional prediction and will investigate the potential influence of task order in an exploratory analysis (Exploratory 1).

The current study

In this large-scale online study, we wish to thoroughly investigate whether the MC-SNARC is truly categorical (i.e., better described by a categorical number magnitude

predictor) and the PJ-SNARC continuous (i.e., better described by a continuous number magnitude predictor). Evidence for this systematic difference would suggest that we should not talk about *the* SNARC effect, but instead acknowledge that different SNARC effects exist, which are elicited depending on the task. It is crucial to shed light on this issue ~~both for the conceptual and the practical level~~, because wrong~~ly measuring~~ measurements and ~~interpreting~~ interpretations of the SNARC effect can lead to ~~underestimations~~ misconceptions of SNAs ~~and to wrong assumptions about their relations with covariates~~. Another goal of the present study is to investigate the relationship between the SNARC effect(s) that will be observed in the two tasks.

First, we expect the following replications~~The following replications will serve as manipulation checks~~ in the current study:

1. a SNARC effect in both MC and PJ with the standard analysis of a continuous linear regression (this ~~manipulation check~~positive control will be used as a basis for all further analyses, i.e., finding the SNARC effect with the standard analysis in both tasks is a prerequisite for testing the hypotheses in this study);
2. a MARC effect in PJ, but not in MC, because the activation of parity seems not to be automatic when only a semantic number feature other than parity is task-~~ir~~relevant;
3. shorter RTs in MC than in PJ, because processing magnitude is more straightforward and automatized than processing parity;
4. a Numerical Distance Effect in MC, which is typically found.

To summarize our hypotheses derived above, we expect:

1. ~~(a)~~ a categorical MC-SNARC, i.e., a better fit of the categorical than continuous MC-SNARC model~~, and (b)~~
~~1.~~2. a continuous PJ-SNARC, i.e., a better fit of the continuous than categorical PJ-SNARC model;

~~2. a task-order effect on (a) the MC-SNARC and (b) the PJ-SNARC, such that the SNARC effect is larger when the respective task is the second, because the first task already activates magnitude processing and its mapping onto space;~~

3. (a) a Numerical Size Effect in both tasks, (b) which is stronger in MC than in PJ, because processing magnitude is task-relevant in MC but task-irrelevant in PJ.

Moreover, we will explore whether the following observations can be made (without directional predictions):

1. task-order effects on both (a) the MC-SNARC and (b) the PJ-SNARC;
2. a good model fit when including both continuous and categorical magnitude predictors for (a) the MC-SNARC or for (b) the PJ-SNARC, indicating a mixed shape of the SNARC effect (see Panel C in Figure 2);
3. compatibility-order effects on (a) the MC-SNARC (SNARC slopes in Conditions 1 and 3 versus Conditions 2 and 4) or on (b) the PJ-MARC (MARC slopes in Conditions 1 and 3 versus Conditions 2 and 4);
4. a shape difference of (a) the MC-SNARC or for (b) the PJ-SNARC between earlier and later phases within each task;
5. a correlation between the categorical MC-SNARC slopes and the continuous PJ-SNARC slopes.

~~For this purpose, we~~We will collect data for MC and PJ with the numbers from 1 to 9 (excluding 5) in a within-subjects design using the typical bimanual response setup. Participants will be assigned to one of four conditions differing in block order, and 30 repetitions will provide reliable estimates per experimental cell (number magnitude * response side * task; see Cipora & Wood, 2017). Conducting this study online offers the possibility to test much larger samples than in most previous studies and thus reach high statistical power (Reips, 2000, 2002).

The SNARC effect has been successfully replicated in online settings (Bulut, Roth, et al., in press~~2024~~; Cipora, Soltanlou, et al., 2019; Gökaydin et al., 2018; Koch et al., 2023; Roth, Caffier, Cipora, et al., in press~~2024~~; Roth, Caffier, Reips, et al., in press~~2023~~; Roth, ~~Jordan~~Huber, et al., 2024). The measurement in the online setup showed reliability and a similar magnitude compared to the SNARC effect that is typically observed in lab studies. Further, it seems to be valid regarding correlations with mean RT and standard deviations of RT. We will calculate Bayes Factors (BF_{10}) to be able to quantify evidence both for differences between MC and PJ as well as for the relationship between the SNARC effects in the two tasks, and lack of such differences or such a relationship. This way, we hope to shed more light on the SNARC effect and specifically its shape in the two popular and widely used tasks.

Method

The ethics committee of the University of Tübingen's ~~Institute~~ Department of Psychology has approved of this study.

Sample size considerations

The "Sequential Bayes Factor with maximal n" (SBF+maxN) approach described by Schönbrodt and Wagenmakers (2018)[3] will be applied to make our data collection ~~more~~ efficient. This means that we will run the data analysis with a total of 500 participants in ~~a~~ the first round and recruit further participants sequentially in steps of 50 until ~~our~~ the ~~optional~~ stopping criterion or maximal sample size is reached. Our ~~optional~~ stopping criterion will be ~~the case of~~ moderate evidence regarding all hypotheses, so that ~~our~~ the data either provides evidence in favor ($BF_{10} > 3$) or against ($BF_{10} < 1/3$) each of them.

[3] Note that, apart from a maximum sample size, a minimum sample size needs to be defined as well, which is why it might be more reasonable to term the approach "sequential Bayes Factor with a minimum and maximum N" as done by Witt (2019).

For the SBF+maxN approach, we need to define a maximal sample size. Thus, we will determine the sample size that is necessary to detect evidence for a true underlying effect or against a truly absent effect with a high probability (similar to power analyses in the frequentist framework). This will be done for each hypothesis and the largest required sample size will be chosen as maximal sample size for the SBF+maxN approach. Our main aim of the current study is to determine the shape of the SNARC effect in the two most common tasks. For this, we will compare the fit of a continuous and a categorical statistical model in MC and PJ separately (to test Hypotheses 1~~a~~ and ~~1b~~2). We therefore chose the effect size of interest (ESOI) in a standardized unit, namely Cohen's $d = 0.2$ (although this is not recommended for power simulations, see Correll et al., 2020). The sample size considerations were based on this ESOI for all hypotheses, because smaller effect sizes are not practically meaningful. Specifically, $d = 0.2$ reflects a small effect and corresponds to around 1% of explained variance (calculated according to Ruscio, 2008, using the conversion formula assuming equal-sized groups, see their Table 2). Regarding the detection of a SNARC effect while assuming similar standard deviations as reported in the literature, $d = 0.2$ corresponds to -4 in the continuous MC-SNARC (with $SD = 20$), -10 in the categorical MC-SNARC (with $SD = 50$), and -2 in the continuous PJ-SNARC (with $SD = 10$) in their measured unit (i.e., increase of right-hand advantage per continuous magnitude or categorically for large compared to small numbers in milliseconds). These SNARC slopes are of a typically observed or even small size.

Analogously to statistical power simulations in the frequentist framework, we randomly drew 5000 samples from a distribution around the ESOI ($d = 0.2$) and simulated the probability to obtain at least moderate evidence (i.e., $BF_{10} > 3$) for that effect size by looking at the proportion of Bayesian tests revealing at least moderate evidence for the alternative hypothesis (for a similar approach, see Kelter, 2021; Roth, Caffier, Reips, et al., in press~~2023~~). Similarly, we randomly drew 5000 samples from a distribution with the respective SD around a truly absent effect ($d = 0$) and simulated the probability to obtain at least moderate evidence for the

null hypothesis (see Kelter, 2021). We thereby determined the sample size that is required for a probability of .90 to obtain moderate evidence for and against the six hypotheses with two-sided Bayesian *t*-tests. Paired or one-sample *t*-tests will be used for Hypotheses 1~~a, 1b~~, 2, 3a, and 3b~~; independent-samples *t*-tests will be used for Hypotheses 2a, and 2b~~. Paired or one-sample *t*-tests as well as independent-samples *t*-tests and a Pearson correlation *t*-test will also be used for all replication checks and exploratory analyses. The required sample size was largest for finding at least moderate Bayesian evidence for a true underlying effect of $d = 0.2$ with a probability of .90 in a two-sided Bayesian independent-samples *t*-test ($n = 2 * 850 = 1700$). The required sample sizes for finding evidence against a truly absent effect in an independent-samples *t*-test ($n = 2 * 340 = 680$), for evidence for a true underlying effect in a one-sample or paired *t*-test ($n = 440$), or for evidence against a truly absent effect in a one-sample or paired *t*-test ($n = 160$) were much smaller. We will therefore target $n = 1700$ as a maximal sample size for the SBF+maxN approach. The exact calculations and results for all tests can be found here: https://osf.io/4wpv6/.

Participants

We will sequentially recruit adults aged between 18 and 40 years via the recruiting platform Prolific, which checks participants' demographic variables via objective criteria rather than self-report during signup – an important issue in recruitment for Web-based research (Reips, 2021). As the study will be conducted in English, participation is only possible for native English speakers (as per Prolific's screening based on self-reports). Complete participation will be compensated with £5 (Prolific users receive their payment in this currency), and incomplete participation will be compensated partially.

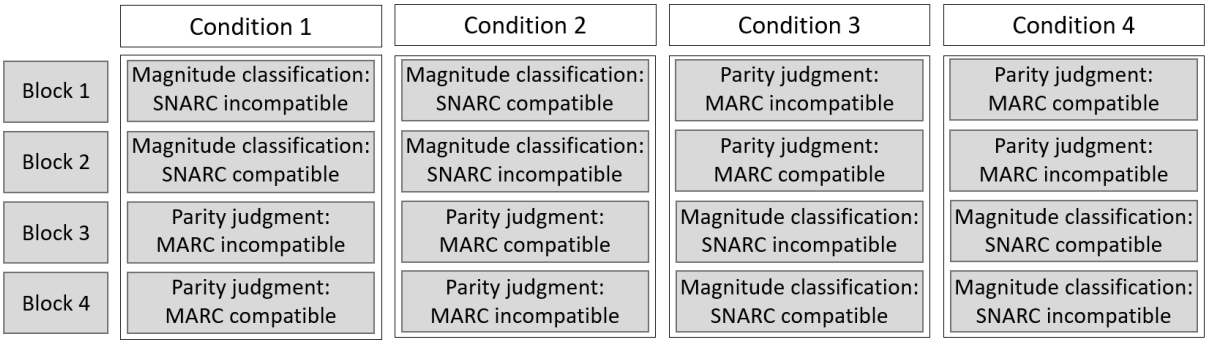
Design and experimental task

The present study follows a 2 (task: MC vs. PJ) * 2 (compatibility: incompatible vs. compatible) within-subjects design, resulting in four experimental blocks per participant. Participants will be randomly assigned to one of four block orders. In Conditions 1 and 2,

participants complete MC in the first and PJ in the second half of the experiment, while the task order is reversed in Conditions 3 and 4. Both blocks of each task will be kept together and presented one after the other to avoid mixing up instructions. Within each task, participants are assigned to the SNARC-/MARC-incompatible block first and to the SNARC-/MARC-compatible block second in Conditions 1 and 3, while the compatibility order is reversed in Conditions 2 and 4 (cf. Figure 1). Given the planned number of trials (see below), each of the two tasks is expected to take 15 minutes, so that the full participation including both tasks and some demographic questions will take approximately 35 minutes.

Figure 1

Within-subjects manipulations and resulting block orders counterbalanced between-subjects



Note. We will randomly assign participants to one of the four conditions ~~illustrated in this figure, which differ in block order. These conditions result~~ resulting from the combination of task order and compatibility order in the 2 (task: MC vs. PJ) * 2 (compatibility: incompatible vs. compatible) within-subjects design. ~~In Conditions 1 and 2, participants will start with MC and end with PJ, whereas in Conditions 3 and 4, participants will start with PJ and end with MC. In Conditions 1 and 3, participants will start with the SNARC-/MARC-incompatible block, whereas in Conditions 2 and 4, participants will start with the SNARC-/MARC-compatible block.~~

A binary response-key setup will be employed, requiring participants to respond as quickly and accurately as possible using a left or right key (defaults: D or K – can be adjusted individually by participants due to large technical variance on the Internet; Reips, 2000, 2021) depending on whether the number presented on the screen is smaller or larger than 5 (MC) or whether it is odd or even (PJ). In each of the experimental conditions resulting from the two within-subjects factors *task* and *compatibility*, number magnitude (1, 2, 3, 4, 6, 7, 8 vs. 9) will be manipulated. Thirty repetitions per experimental cell will lead to 240 SNARC-incompatible and 240 SNARC-compatible trials in MC, as well as 240 MARC-incompatible and 240 MARC-compatible trials in PJ per participant. Participants must take a break of a minimum of 30 seconds between blocks. T ~~and t~~he order of stimulus presentation within blocks will be ~~fully~~ randomized, with the restriction that within each block, each stimulus will be presented for the 1st throughout 15th time before each stimulus will be presented for the 16th throughout 30th time (i.e., each block is divided in two subblocks indistinguishable to the participant, in which each stimulus will be presented 15 times). Each trial will start with a square (extended ASCII 254, size 72px), serving as the eye fixation point (300 ms), presented in the center of the screen. Then the number (Open Sans font, size 72px) will replace the square and remain on the screen 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 color (0, 0, 0 in RGB notation), while the background remains gray (150, 150, 150 in RGB notation) throughout the experiment. A practice session with 16 trials will precede each block, in which each number will be presented twice. Accuracy feedback will appear during practice sessions only.

Procedure

At the very beginning of the experiment, a seriousness check (e.g., Reips, 2009) will be applied (i.e., 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 computer (participation from mobile devices is not possible because a keyboard is required), and if they

are at least 18 years old. Then, participants will be asked to provide basic demographic data, namely age, gender, first native language, and handedness. In each question, participants will have the opportunity to click on "I prefer not to answer." Next, participants may choose response keys for the experimental task that are located on the same height and about one hand width apart from each other on their keyboard, e.g., if this is not the case for the default response keys D and K. These default keys were chosen because they are located on the same height and about one hand width apart from each other on typical keyboards like QWERTZ, QWERTY, and AZERTY. Then, instructions will be displayed, and the first block of the experimental task will start with its practice trials.

After completion of both experimental tasks, ~~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* (with responses being coded between 0 and 400 for data analysis). 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). Participants will be provided with a completion code to be inserted in Prolific and with contact information of our research team.

The experiment has been set up with WEXTOR (https://wextor.eu; Reips & Neuhaus, 2002) in its HTML and JavaScript framework and adapted (see demo version at https://luk.uni-konstanz.de/numcog-2/?demo). Our previous experiments have demonstrated that this software is suitable for detecting the SNARC effect in an online setup (Roth, Caffier, Reips, et al., in press~~2023~~; Roth, Caffier, Cipora, et al., in press~~2024~~). To prevent search engine bots (e.g., Googlebot) from submitting data on our experiment, WEXTOR equips 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 with a screen width of over 600 pixels. Additionally, to exclude multiple submissions from the same devices, we will perform checks based on User-Agents and IP addresses during data evaluation.

Data preprocessing

All data preprocessing steps and all analyses will be performed in the statistical computing software R (R Core Team, 2022). As concerns data preprocessing, we ~~want to~~will stay consistent with our previous studies and apply the same inclusion criteria (Roth, Caffier, Reips, et al., in press~~2023~~; Roth, Caffier, Cipora, et al., in press~~2024~~). Only datasets of participants who indicate to be at least 18 years old and the intention to seriously participate will be analyzed. Datasets will not be included for analyses if participants describe their environment as very/extremely noisy or if they report multiple major distractions. Practice trials and incorrectly answered trials will not be analyzed. Only trials with RTs from 200 to 1500 ms will be included in the analysis. Further outliers will be removed in an iterative trimming procedure for each participant separately, such that only RTs that are a maximum 3 SDs above or below the individual mean RT of all remaining trials will be considered. Finally, only datasets of participants with at least 75% valid remaining trials and without any empty experimental cell (number magnitude * response side * task) will be considered.

Data analysis

Confirmatory data analysis

An overview of all hypotheses, corresponding tests, and interpretations of possible outcomes is given in the Study Design Table (see https://osf.io/4wpv6/). We will calculate Bayes Factors 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). 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). Considering a BF_{10} larger than 3 as evidence against the null hypothesis is more conservative than rejecting a null hypothesis in the frequentist framework with the typical significance level of $\alpha = .05$ (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 all hypotheses.

Reaction times (RTs) will be measured as the time elapsing from the onset of the number presentation on the screen until a response key is pressed (within the limitations that apply in Internet-based research with consumer-grade equipment, see e.g., Garaizar & Reips, 2019). As the dependent variable, we will calculate the mean differences between reaction times (dRTs), which result from subtracting the average RT of the left hand from the average RT of the right hand for each number separately per participant and for each task separately.

Several regression models will be fit for each participant separately. In these regression models, number magnitude will be included as a predictor for dRTs to determine the shape of the SNARC effect in each task separately. First, magnitude will be included as a continuous predictor, which is equal to the actual stimulus that is displayed (e.g., 3 for number 3, and 8 for number 8). The resulting regression slopes for *continuous magnitude* represent the advantage of right-hand responses compared to left-hand responses in ms per increase by one in continuous magnitude (i.e., traditional repeated-measures regression in the SNARC effect analysis, as first proposed by Fias et al., 1996). Second, magnitude will be contrast-coded as a categorical predictor, using -0.5 for numbers from 1 to 4 and +0.5 for numbers from 6 to 9 (e.g., -0.5 for number 2, and +0.5 for number 7). The resulting regression slopes for *categorical magnitude* represent the advantage of right-hand responses compared to left-hand responses in ms in large compared to small magnitude. Third, for the investigation of the MARC effect, contrast-coded number parity will be included as a predictor of dRTs, with -0.5 for odd and +0.5 for even numbers (as in Cipora, Soltanlou, et al., 2019). The regression slopes for *parity*

represent the advantage of right-hand responses compared to left-hand responses in ms in even compared to odd numbers. An overview of all three predictors (i.e., continuous magnitude, categorical magnitude, and parity), along with their exact coding, can be found in Table 1. For each of the predictors, a more negative coefficient estimate β points towards a stronger SNARC/MARC effect.

Table 1

Overview of dRT predictors

Continuous magnitude	1	2	3	4	6	7	8	9
Categorical magnitude	-0.5	-0.5	-0.5	-0.5	+0.5	+0.5	+0.5	+0.5
Parity	-0.5	+0.5	-0.5	+0.5	+0.5	-0.5	+0.5	-0.5

Note. This table gives an overview of the dRT predictors that will be used in the regression models summarized in Table 2. Continuous magnitude is equal to the actual presented stimulus. Categorical magnitude is contrast-coded with -0.5 for smaller and +0.5 for larger numbers. Number parity is contrast-coded with -0.5 for odd and +0.5 for even numbers.

To test the ~~Manipulation Check~~Replications 1 and 2, as well as Hypotheses 1 and 2, we will fit four regression models per participant and per task (for an overview, see Table 2). First, the presence of a SNARC effect in both tasks (~~Manipulation~~Replication Check 1, which will be used as a positive control) will be tested in a repeated-measures regression as usually done in SNARC research (see Fias et al., 1996, adapted from Lorch & Myers, 1990). For this, dRTs will be regressed on continuous magnitude in models MC-1 and PJ-1 for each participant separately. The resulting slopes will be tested against zero in a two-sided Bayesian one-sample *t*-test, with Bayesian evidence for a difference from zero indicating a continuous SNARC effect,

and negative value of the slope indicating the typical SNARC effect. In an exploratory analysis, we will also test whether the MC-SNARC is stronger than the PJ-SNARC by comparing the slopes for continuous magnitude predictors resulting from models MC-1 and PJ-1 in a two-sided Bayesian paired *t*-test.

Then, the presence of the MARC effect (~~Manipulation~~Replication Check 2) will be tested in both tasks with the same repeated-measures regression approach as the SNARC effect. For this, regression models MC-3 and PJ-3 will be computed. These models will contain a contrast-coded parity predictor and the magnitude predictor with the better fit in the previous test. Because the number-parity predictor and the number-magnitude predictor are orthogonal to each other (i.e., their correlation is zero), both can be concurrently included within one regression model without affecting the respective other parameter estimate. ~~Next~~Then, the slopes will be tested against zero in a two-sided Bayesian one-sample *t*-test, with evidence for a difference from zero indicating a MARC effect, which is only expected in PJ but not MC.

Next, we will investigate whether the MC-SNARC and the PJ-SNARC are continuous or categorical (i.e., which number magnitude predictor fits the observed dRTs better; ~~see~~ Hypotheses 1~~a~~ and ~~1b~~2). For this, besides regressing dRTs on continuous magnitude in models MC-1 and PJ-1 as previously described, they will be regressed on categorical magnitude in MC-2 and PJ-2. Then, we will logit-transform the R^2 for each model for each participant separately to approximate two normal distributions and compare the logit-transformed R^2 between the two models in a two-sided Bayesian paired *t*-test (as Koch et al., 2023, did this in a frequentist approach). A better fit of model MC-2 compared to MC-1 and of PJ-1 compared to PJ-2 as reflected by Bayesian evidence for a higher logit-transformed R^2 would indicate a categorical MC-SNARC (Hypothesis 1~~a~~) and a continuous PJ-SNARC (Hypothesis ~~1b~~2). Additionally, we will confirm these findings via a Bayesian approach: dRTs will be regressed on continuous and categorical magnitude for both PJ and MC in four separate Bayesian models, and in each task, a leave-one-out cross validation will be performed to figure out which of the

two predictors better fits our data (using the R package *brms* by Buerkner, 2017, and the R package *loo* by Vethari et al., 2017). An overview of possible SNARC effect shapes and the corresponding regression models tested in the current study can be found in Figure 2.

~~Then, we will investigate task-order effects on both the MC-SNARC and the PJ-SNARC (Hypotheses 2a and 2b). That is, we will test task-order effects by comparing SNARC slopes for both tasks (with the predictor that fits better in the respective task, according to Hypotheses 1a and 1b) in Conditions 1 and 2 (first MC, second PJ) with Conditions 3 and 4 (first PJ, second MC) for each task separately (see Figure 1 for an overview of experimental Conditions). For this, we will run two two-sided Bayesian independent-samples *t*-tests. A stronger MC-SNARC and a stronger PJ-SNARC in Conditions 1 and 2 (taken together) compared to Conditions 3 and 4 (taken together) would reflect that both SNARC effects are larger when participants start with MC instead of PJ.~~

Table 2

Overview of regression models that will be fit for each participant

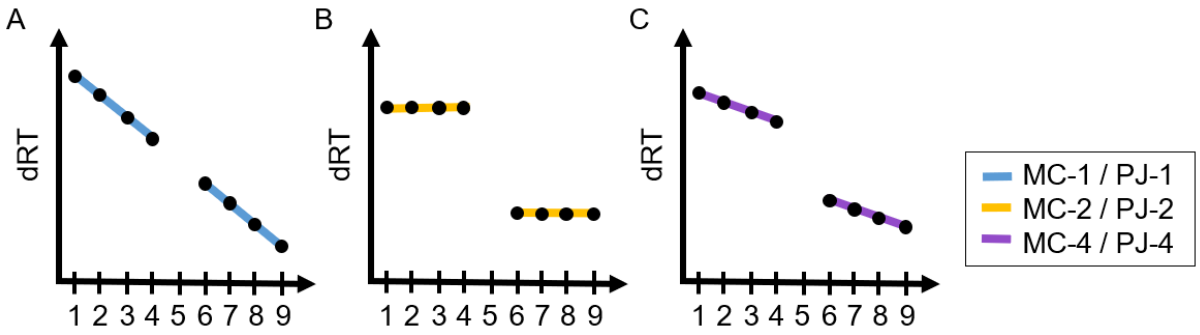
Magnitude classification	
MC-1	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous}$
MC-2	$dRT \sim \beta_0 + \beta_1 * magnitude_{categorical}$
MC-3[a]	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous/categorical} + \beta_2 * parity$
MC-4	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous} + \beta_2 * magnitude_{categorical}$
Parity judgment	
PJ-1	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous}$
PJ-2	$dRT \sim \beta_0 + \beta_1 * magnitude_{categorical}$
PJ-3[b]	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous/categorical} + \beta_2 * parity$

PJ-4	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous} + \beta_2 * magnitude_{categorical}$

Note. Four regression models will be fit for each participant separately for MC (MC-1, MC-2, MC-3, and MC-4) and PJ (PJ-1, PJ-2, PJ-3, and PJ-4). The predictors used in the models are specified in Table 1. In each model, β_1 (and β_2) are the coefficients of the respective predictors for number magnitude or number parity. β_0 denotes the model intercept. [a] For MC-3, the better magnitude predictor from MC-1 and MC-2 will be used. [b] For PJ-3, the better magnitude predictor from PJ-1 and PJ-2 will be used.

Figure 2

Different shapes of the SNARC effect



Note. In Panel A, the SNARC effect is reflected by a linear regression line with a negative slope, that is, dRTs are best predicted by continuous magnitude (models MC-1 and PJ-1). In Panel B, the SNARC effect is reflected by a step-like function, that is, dRTs are best predicted by categorical magnitude (models MC-2 and PJ-2). Panel C shows an intermediate shape of the SNARC effect, where both continuous and categorical magnitude predict dRTs (models MC-4 and PJ-4). The typically observed MC-SNARC appears as shown in Panel B (Hypothesis 1~~a~~), and the typically observed PJ-SNARC appears as shown in Panel A (Hypothesis ~~1b~~2).

For ~~Manipulation~~Replication Checks 3 and 4 as well as for Hypotheses 2a and 2b~~is 3~~, RTs will be examined in detail. First, to test whether reactions are on average shorter in MC

than in PJ (~~Manipulation~~Replication Check 3), we will compare mean RTs per participant between tasks in a two-sided Bayesian paired *t*-test. Next, the presence of a Numerical Distance Effect in MC (~~Manipulation~~Replication Check 4) and of a Numerical Size Effect in both MC and PJ (Hypotheses 3a and 3b) will be investigated with the repeated-measures regression approach (as in Hohol et al., 2020). In MC, RTs will be regressed on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9) for each participant separately. Because numerical distance and magnitude are orthogonal (i.e., their correlation is zero), both can be concurrently included within one regression model without affecting the respective other parameter estimate. In PJ, RTs will only be regressed on continuous magnitude for each participant separately. Next, resulting regression slopes will be tested against zero in a two-sided Bayesian one-sample *t*-test for each task separately. Evidence for negative slopes for the numerical distance predictor indicates faster reactions for larger numerical distance, reflecting the Numerical Distance Effect (~~Manipulation~~Replication Check 4). Evidence for positive slopes for the magnitude predictor indicates slower reactions for increasing number magnitude, reflecting the Numerical Size Effect (Hypothesis 3a). Last, we will test whether the magnitude of the Numerical Size Effect is stronger in MC than in PJ (Hypothesis 3b) by comparing resulting slopes between tasks in a two-sided Bayesian paired-samples *t*-test.

Exploratory data analysis

After the analyses for ~~manipulation~~replication checks and hypotheses, we will investigate task-order effects on both the MC-SNARC and the PJ-SNARC (Exploratory 1). That is, we will test task-order effects by comparing SNARC slopes in Conditions 1 and 2 (first MC, second PJ) with Conditions 3 and 4 (first PJ, second MC) for each task separately (see Figure 1 for an overview of experimental conditions). The predictor that fits better in the respective task will be used (according to Hypotheses 1 and 2). For this, we will run two two-sided Bayesian independent-samples t-tests.

Next, a fourth model including both continuous and categorical magnitude will be fitted for both tasks (MC-4 and PJ-4). Both resulting slopes will be tested against zero in two-sided Bayesian one-sample *t*-tests, with Bayesian evidence for both slopes being different from zero indicating a mixed shape of the SNARC effect, as illustrated in Panel C in Figure 2 (Exploratory 2). This would mean that the dRT regression slope is negative within small and within large numbers, while there is a categorical step between numbers 4 and 6 (for an empirical observation of a such pattern, see Figure 2b in Nuerk, Bauer, et al., 2005).

Further, we will exploratorily test compatibility-order effects on the MC-SNARC by comparing SNARC slopes in Conditions 1 and 3 (first SNARC-incompatible, second SNARC-compatible) with SNARC slopes in Conditions 2 and 4 (first SNARC-compatible, second SNARC-incompatible; Exploratory 3a). Note that we will use the categorical or continuous slope here depending on which of both describes the MC-SNARC better (i.e., depending on the outcome regarding Hypothesis 1a). Similarly, we will test compatibility-order effects on the MARC effect in PJ by comparing MARC slopes in Conditions 1 and 3 (first MARC-incompatible, second MARC-compatible) with MARC slopes in Conditions 2 and 4 (first MARC-compatible, second MARC-incompatible; Exploratory 3b). For this, we will run two two-sided Bayesian independent-samples *t*-tests. Evidence for a stronger SNARC/MARC effect in Conditions 2 and 4 compared to Conditions 1 and 3 would reflect larger compatibility effects when the response-key assignment is first compatible and then incompatible, and vice versa.

Moreover, we will explore whether the shape of the SNARC effect differs between earlier and later phases within each task (Exploratory 4). Importantly, it is not possible to determine the SNARC effect in the first or second block of each task separately, because both blocks are needed in order to calculate the differences between left- and right-hand responses. Therefore, we will compute the models MC-4 and PJ-4 and test the resulting slopes for both the continuous and the categorial predictors against zero in two-sided Bayesian one-sample *t*-tests,

but instead of considering all 30 repetitions per block, we will only consider the first or second halves of both blocks within each task (i.e., first or second 15 repetitions of each number in one and in the other response-to-key assignment). This way, we can investigate whether early trials in each response-to-key assignment lead to a different SNARC shape than late trials.

Lastly, we will calculate Pearson's correlation between the categorical MC-SNARC slopes and the continuous PJ-SNARC slopes (Exploratory 5). We will run a two-sided Bayesian Pearson correlation test to see whether the spatial mapping of number magnitude within participants is similar in both tasks.

Data quality and ~~manipulation check~~positive controls

To control the data quality in our study, we have implemented a seriousness check (Reips, 2009) 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 MC or PJ, we will only include correctly answered trials, trim RTs, and only include datasets with a minimum of 75% remaining valid trials (as described in the data preprocessing pipeline). Moreover, the test of the MC-SNARC and PJ-SNARC analyzed with the traditional linear regression (Replication Check 1) will serve as ~~Manipulation Check 1~~positive control. Importantly, we consider this ~~manipulation Check~~positive control as a prerequisite for all further analyses and will only proceed with testing the other hypotheses if we can find at least moderate Bayesian evidence for both the continuous MC-SNARC and the continuous PJ-SNARC at the group level. Finally, ~~we have four~~ in additional ~~manipulation~~replication checks, where we aim to replicate results from previous studies to validate our investigation.

Possible limitations and unexpected outcomes

Importantly, including both continuous and categorical magnitude within one single regression model (as in MC-4 and PJ-4) is problematic because of collinearity. These two predictors correlate highly, namely with r = .913. However, we still decided to compute one

such regression model for each task because the true shape of the SNARC effect might be determined by both continuous and categorical number magnitude simultaneously.

In the present study, we test the two most frequently used versions of MC and PJ (i.e., with symbolic single-digit numbers) in a sample in which the SNARC effect is not controversial (i.e., Western culture with left-to-right reading and writing direction). ~~However, f~~Future studies ~~must~~ will show whether our results hold true for different types of stimuli and for different samples.

Further procedure

Data collection will start after critical revisions of the current registered replication report according to peer review and is estimated to last one month. Data analysis is expected to be finished within three months after data collection.

Data and code availability

Anonymized data and analysis scripts will be available via the Open Science Framework (https://osf.io/g48s2/).

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, A. T. Overlander, U.-D. Reips, L. Roth; *Formal Analysis*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Funding Acquisition*: K. Cipora, H.-C. Nuerk, U.-D. Reips.; *Investigation*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Methodology*: K. Cipora, H.-C. Nuerk, U.-D. Reips, L. Roth; *Project Administration*: H.-C. Nuerk, U.-D. Reips, L. Roth; *Resources*: H.-C. Nuerk, U.-D. Reips; *Software*: A. T. Overlander, U.-D. Reips; *Supervision*: K. Cipora, H.-C. Nuerk, U.-D. Reips;

Validation: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Visualization*: L. Roth; *Writing - original draft*: L. Roth; *Writing - review and editing:* K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips.

Competing interests

The authors declare no conflicts of interest with the content of this article.

Acknowledgements

This research was supported by the DFG project "Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)" (NU 265/8-1 and RE 2655/3-1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips, supporting Lilly Roth and Annika Tave Overlander, with the assistance of Krzysztof Cipora as a cooperation partner. Hans-Christoph Nuerk's work on spatial-numerical associations is additionally supported by the DFG-projects 265-5/1 and 265-5/2: On the interplay of modal and amodal encodings underlying space-metric associations (SMAs). Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1). The authors would like to thank Sebastian Sandbrink for proofreading of this Registered Report.

References

Abrahamse, E., van Dijck, J.-P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology, 7*, Article 977. https://doi.org/10.3389/fpsyg.2016.00977

Bachot, J., Gevers, W., Fias, W., & Roeyers, H. (2005). Number sense in children with visuospatial disabilities: Orientation of the mental number line. *Psychology Science, 47*(1), 172–183.

Bae, G. Y., Choi, J. M., Cho, Y. S., & Proctor, R. W. (2009). Transfer of magnitude and spatial mappings to the SNARC effect for parity judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(6), 1506–1521. https://doi.org/10.1037/a0017257

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance, 2*(3), 435–447. https://doi.org/10.1037/0096-1523.2.3.435

Bargh, J. A. (1992). The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects. *The American Journal of Psychology, 105*(2), 181–199. https://doi.org/10.2307/1423027

Basso Moro, S., Dell'Acqua, R., & Cutini, S. (2018). The SNARC effect is not a unitary phenomenon. *Psychonomic Bulletin & Review, 25*(2), 688–695. https://doi.org/10.3758/s13423-017-1408-3

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

Buerkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software. 80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Bull, R., Marschark, M., & Blatto-Vallee, G. (2005). SNARC hunting: Examining number representation in deaf students. *Learning and Individual Differences*, *15*(3), 223–236. https://doi.org/10.1016/j.lindif.2005.01.004

Bulut, M., Çetinkaya, H., & Dural, S. (2024). *SNARC Effect in a Transfer Paradigm: Long-Lasting Effects of Stimulus-Response Compatibility Practices* [preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/v76ux

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (in press). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research.* Preprint at: https://doi.org/10.31234/osf.io/qcb6m~~Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2024). *Cultural aspects of the mental number line beyond reading/writing direction* [manuscript in preparation]. University of Tübingen, Germany.~~

Casasanto, D., & Pitt, B. (2019). The faulty magnitude detector: Why SNARC-like tasks cannot support a generalized magnitude system. Cognitive Science, 43(10), Article e12794. https://doi.org/10.1111/cogs.12794

Cheung, C.-N., Ayzenberg, V., Diamond, R. F. L., Yousif, S., & Lourenco, S. F. (2015). Probing the mental number line: A between-task analysis of spatial-numerical associations. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 357–362.

Cipora, K. (2014). *Between-task consistency of the SNARC effect* [poster presentation]. XXXIInd European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/nx6vq/

Cipora, K., & Göbel, S. M. (2013). *Number-space associations: Just how reliable is the SNARC effect* [poster presentation]. In 16th European Workshop on Cognitive Neuropsychology. https://osf.io/nx6vq

Cipora, K., Haman, M., Domahs, F., & Nuerk, H.-C. (2020). Editorial: On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations. *Frontiers in Psychology*, *11*, Article 182. https://doi.org/10.3389/fpsyg.2020.00182

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. L. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–59). Cambridge University Press. https://doi.org/10.1017/9781316651247.004

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. Mix & M. Batista (Eds.), *Visualizing Mathematics: The role of spatial reasoning in mathematical thought* (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*. Open Science Framework [Preprint]. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Schiltz, C., Pesenti, M., Goebel, S. M., Willmes, K., & Nuerk, H.-C. (2020). *Parity and space: On the prevalence and*

individual differences in the MARC effect [poster presentation]. XXXVIIIth European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/4qhpw

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: a 20∗ 20 Monte Carlo investigation. *Frontiers in Psychology, 8*, 243273. https://doi.org/10.3389/fpsyg.2017.01194

Cohen, J. (1988). The effect size. *Statistical power analysis for the behavioral sciences*, 77–83. https://doi.org/10.4324/9780203771587

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

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., Zhang, M., Li, Y., & Zhu, X. (2018). The association of number and space under different tasks: Insight from a process perspective. *Frontiers in Psychology*, *9*, Article 957. https://doi.org/10.3389/fpsyg.2018.00957

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

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica*, *196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

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

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

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

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitude. *Journal of Cognitive Neuroscience*, *15*(1), 47–56. https://doi.org/10.1162/089892903321107819

Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition - From single digits to arithmetic. *Quarterly Journal of Experimental Psychology (2006)*, *67*(8), 1461–1483. https://doi.org/10.1080/17470218.2014.927515

Fitousi, D., Shaki, S., & Algom, D. (2009). The role of parity, physical size, and magnitude in numerical cognition: The SNARC effect revisited. *Attention, Perception & Psychophysics*, *71*(1), 143–155. https://doi.org/10.3758/APP.71.1.143

Garaizar, P., & Reips, U.-D. (2019). Best practices: Two web browser-based methods for stimulus presentation in behavioral experiments with high resolution timing requirements. *Behavior Research Methods, 51*, 1441–1453. https://doi.org/10.3758/s13428-018-1126-4

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Georges, C., Hoffmann, D., & Schiltz, C. (2013). *The SNARC effect and its relationship to spatial abilities in women* [poster presentation]. In 18th conference of the European

Society for Cognitive Psychology, Budapest, Hungary. http://hdl.handle.net/10993/13012

Georges, C., Hoffmann, D., & Schiltz, C. (2017). How and why do number-space associations co-vary in implicit and explicit magnitude processing tasks? *Journal of Numerical Cognition*, *3*(2), 182–211. https://doi.org/10.5964/jnc.v3i2.46

Gevers, W., Ratinckx, E., Baene, W. de, & Fias, W. (2006). Further evidence that the SNARC effect is processed along a dual-route architecture: Evidence from the lateralized readiness potential. *Experimental Psychology*, *53*(1), 58-68. https://doi.org/10.1027/1618-3169.53.1.58

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

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

Gibson, L. C., & Maurer, D. (2016). Development of SNARC and distance effects and their relation to mathematical and visuospatial abilities. *Journal of Experimental Child Psychology*, *150*, 301–313. https://doi.org/10.1016/j.jecp.2016.05.009

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential Effects in SNARC. *Scientific Reports*, *8*(1), Article 10996. https://doi.org/10.1038/s41598-018-29337-2

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology General*, *121*(4), 480–506. https://doi.org/10.1037/0096-3445.121.4.480

Han, M., Mao, X., Cai, M., Jia, X., & Guo, C. (2017). The effect of positive and negative signs on the SNARC effect in the magnitude judgment task. *Acta Psychologica Sinica*, *49*(8), 995–1008. https://doi.org/10.3724/SP.J.1041.2017.00995

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. *Acta Psychologica*, *128*(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory & Cognition*, *18*(1), 40–46. https://doi.org/10.3758/BF03202644

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child Psychology*, *116*(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports*, *10*(1), Article 11531. https://doi.org/10.1038/s41598-020-68202-z

Hubbard, E. M., Ranzini, M., Piazza, M., & Dehaene, S. (2009). What information is critical to elicit interference in number-form synaesthesia? *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *45*(10), 1200–1216. https://doi.org/10.1016/j.cortex.2009.06.011

Ito, Y., & Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from the SNARC effect. *Memory & Cognition*, *32*(4), 662–673. https://doi.org/10.3758/BF03195857

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

Kiesel, A., Wendt, M., & Peters, A. (2007). Task switching: On the origin of response congruency effects. *Psychological Research*, *71*(2), 117–125. https://doi.org/10.1007/s00426-005-0004-8

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

Lohmann, J., Schroeder, P. A., Nuerk, H.-C., Plewnia, C., & Butz, M. V. (2018). How deep is your SNARC? Interactions between numerical magnitude, response hands, and reachability in peripersonal space. *Frontiers in Psychology*, *9*, Article 622. https://doi.org/10.3389/fpsyg.2018.00622

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

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

Mourad, A., & Leth-Steensen, C. (2017). Spatial reference frames and SNARC. *Journal of Cognitive Psychology*, *29*(2), 113–128. https://doi.org/10.1080/20445911.2016.1249483

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520. https://doi.org/10.1038/2151519a0

Nathan, M. B., 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

Nuerk, H.-C., Bauer, F., Krummenacher, J., Heller, D., & Willmes, K. (2005). The power of the mental number line: How the magnitude of unattended numbers affects performance in an Eriksen task. *Psychology Science*, *47*(1), 34–50. https://psycnet.apa.org/record/2005-11470-005

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. A, Human Experimental Psychology*, *57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Nuerk, H.-C., & Schroeder, P. A. (2024). *A Psychological Markedness Account of human cognition: The example of numerical cognition* [manuscript in preparation]. Department of Psychology, University of Tuebingen, Germany.

Nuerk, H.-C., Wood, G., & Willmes, K. (2005). The universal SNARC effect: The association between number magnitude and space is amodal. *Experimental Psychology*, *52*(3), 187–194. https://doi.org/10.1027/1618-3169.52.3.187

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*, Article 55976. https://doi.org/10.3389/fpsyg.2014.00419

Pfister, R., Schroeder, P. A., & Kunde, W. (2013). ~~Snarc~~ SNARC struggles: Instant control over spatial-numerical associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1953–1958. https://doi.org/10.1037/a0032991

Pinto, M., Pellegrino, M., Lasaponara, S., Scozia, G., D'Onofrio, M., Raffa, G., Nigro, S., Arnaud, C. R., Tomaiuolo, F., & Doricchi, F. (2021). Number space is made by

response space: Evidence from left spatial neglect. *Neuropsychologia*, *154*, Article 107773. https://doi.org/10.1016/j.neuropsychologia.2021.107773

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

Prpic, V., Fumarola, A., Tommaso, M. de, Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1241–1251. https://doi.org/10.1037/xhp0000217

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.0.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

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

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199561803.013.0024

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. In B. E. Rogowitz & T. N. Pappas (Eds.), *SPIE Proceedings, Human Vision and Electronic Imaging XIV*, 724008. SPIE. https://doi.org/10.1117/12.823416

Reips, U.-D. (2021). Web-based research in psychology: A review. *Zeitschrift für Psychologie, 229*(4), 198–213. https://doi.org/10.1027/2151-2604/a000475

Reips, U.-D., & Neuhaus, C. (2002). Wextor: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *34*(2), 234–240. https://doi.org/10.3758/bf03195449

Roettger, T. B., & Domahs, F. (2015). Grammatical number elicits SNARC and MARC effects as a function of task demands. *Quarterly Journal of Experimental Psychology (2006), 68*(6), 1231–1248. https://doi.org/10.1080/17470218.2014.979843

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (in press). True colors SNARCing: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Preprint at: https://doi.org/10.31234/osf.io/aeyn8~~Roth, L., Caffier, J., Reips, U. D., Cipora, K., Braun, L., & Nuerk, H.-C. (2024). *True colors SNARCing: Automaticity of the SNARC effect evidence from color judgment tasks.* PsyArXiv [Preprint]. https://doi.org/10.31234/osf.io/aeyn8~~

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (in press). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science.* Preprint at: https://osf.io/79zsy/~~Roth, L., Caffier, J., Reips, U. D., Nuerk, H.-C., & Cipora, K. (2023). One and only SNARC? A Registered Report on the SNARC effect's range dependency [in-principle acceptance of Stage 1 Registered Report by PCI in 2023]. https://doi.org/10.17605/OSF.IO/Z43PM~~

Roth, L., Huber, J., Kronenthaler, S., van Dijck, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2024). *Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space* [preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/y7hzv

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

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Saeki, E., & Saito, S. (2009). Verbal representation in task order control: An examination with transition and task cues in random task switching. *Memory & Cognition*, *37*(7), 1040–1050. https://doi.org/10.3758/MC.37.7.1040

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

Schiller, F., Eloka, O., & Franz, V. H. (2016). Using key distance to clarify a theory on the SNARC. *Perception, 45*(1-2), 196–221. https://doi.org/10.1177/0301006615616754

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

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience, 11*, Article 654. https://doi.org/10.3389/fnins.2017.00654

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

Schwarz, W., & Müller, D. (2006). Spatial associations in number-related tasks: A comparison of manual and pedal responses. *Experimental Psychology*, *53*(1), 4–15. https://doi.org/10.1027/1618-3169.53.1.4

Shaki, S., & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross-Cultural Psychology*, *42*(4), 639–650. https://doi.org/10.1177/0022022111406100

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. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition*, *6*(2-3), 441–451. https://doi.org/10.1006/ccog.1997.0303

Tzelgov, J., Ganor-Stern, D., Kallai, A. Y., & Pinhas, M. (2015). Primitives and non-primitives of numerical representations. In R. C. Kadosh & A. Dowker (Eds.), *Oxford library of psychology. The Oxford handbook of numerical cognition* (pp. 45–66). Oxford University Press.

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *18*(1), 166–179. https://doi.org/10.1037/0278-7393.18.1.166

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

Van Dijck, J.-P., & Doricchi, F. (2019). Multiple left-to-right spatial representations of number magnitudes? Evidence from left spatial neglect. *Experimental Brain Research*, *237*(4), 1031–1043. https://doi.org/10.1007/s00221-019-05483-5

Van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition*, *113*(2), 248–253. https://doi.org/10.1016/j.cognition.2009.08.005

Van Dijck, J.-P., Gevers, W., Lafosse, C., & Fias, W. (2012). The heterogeneous nature of number-space interactions. *Frontiers in Human Neuroscience*, *5*, Article 182. https://doi.org/10.3389/fnhum.2011.00182

Van Galen, M. S., & Reitsma, P. (2008). Developing access to number magnitude: A study of the SNARC effect in 7- to 9-year-olds. *Journal of Experimental Child Psychology*, *101*(2), 99–113. https://doi.org/10.1016/j.jecp.2008.05.001

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLOS ONE*, *9*(4), Article e95756. https://doi.org/10.1371/journal.pone.0095756

Weis, T., Nuerk, H.-C., & Lachmann, T. (2018). Attention allows the SNARC effect to operate on multiple number lines. *Scientific Reports*, *8*(1), Article 13778. https://doi.org/10.1038/s41598-018-32174-y

Westermann, R., & Hager, W. (1983). On severe tests of trend hypotheses in psychology. *The Psychological Record*, *33*(2), 201–211. https://doi.org/10.1007/BF03394838

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

Xiang, X., Yan, L., Fu, S., & Nan, W. (2022). *Processing stage flexibility of the SNARC effect: Task relevance or magnitude relevance?* BioRxiv [Preprint]. https://doi.org/10.1101/2022.03.07.482213

Zhang, P., Cao, B., & Li, F. (2022). The role of cognitive control in the SNARC effect: A review. *PsyCh Journal*, *11*(6), 792–803. https://doi.org/10.1002/pchj.586

Zorzi, M., Bonato, M., Treccani, B., Scalambrin, G., Marenzi, R., & Priftis, K. (2012). Neglect impairs explicit processing of the mental number line. *Frontiers in Human Neuroscience*, *6*, Article 125. https://doi.org/10.3389/fnhum.2012.00125

PCI Study Design Table

Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment

(L. Roth, K. Cipora, A. T. Overlander, H.-C. Nuerk, and U.-D. Reips)

Question	**Hypothesis**	**Sampling plan**[1]	**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 continuous SNARC effect be replicated in the magnitude-comparison (MC) and in the parity-judgment (PJ) task?	**Replication ~~Manipulation~~ Check 1:** A significant SNARC effect will be observed in both MC and PJ when determined with the standard analysis of a continuous linear regression.	The Sequential Bayes Factor with maximal n" (SBF+maxN) approach (Schönbrodt & Wagenmakers, 2018) will be applied to collect data in an efficient way. The minimal sample size will be 500 participants, and more participants will be sequentially recruited in steps of 50 until the optional stopping criterion or the maximal sample size will be reached. The maximal sample size was determined by drawing 5000 simulated datasets around the effect size of interest (Cohen's d = 0.2) to estimate the probability to obtain evidence for or against	1. Regression of dRTs on continuous number magnitude (1, 2, 3, 4, 6, 7, 8, 9; see models MC-1 and PJ-1) for each task separately per participant (as in Fias et al., 1996) 2. Two two-sided Bayesian one-sample t-tests of SNARC slopes against zero for each task separately	This replication check aims at validating the experimental manipulation and method applied in the current study. That is, finding the SNARC effect in both tasks by using the typical analysis will be a positive control in this study, and will be considered as a prerequisite for all further hypothesis tests. ~~Manipulation checks aim at replicating observations that are typically made in the used paradigm, thus serving to validate our applied methodology.~~ The sensitivity of the tests, however, depends on the final sample size determined by the	If evidence is found for the continuous SNARC slopes to differ from zero and to be negative, the SNARC effect is detectable with the standard analysis, which would be in line with previous literature and lay the groundwork for further hypothesis tests. If evidence is found against the SNARC slopes to differ from zero, no SNARC effect is observable, which is highly unlikely. If evidence is found for the continuous SNARC slopes to differ from zero and to be positive, a reversed SNARC effect is	The SNARC effect is usually detected with the standard analysis in both MC and PJ. We strongly expect to find it in this study as well, especially with our large sample size. Not finding the continuous SNARC effect would speak against its robustness in MC and/or PJ and be very surprising. Note that this ~~manipulation~~ replication check will be used as a basis for all further analyses (i.e., finding the SNARC effect with the standard analysis in both tasks is a prerequisite for testing the hypotheses in this study).

		our hypotheses in the Bayesian framework (analogously to statistical power simulations in the frequentist framework). The respective Bayesian test was conducted for the dataset in each of these 5000 simulations. Specifically, the required sample size was determined by making sure that the proportion of Bayesian tests revealing at least moderate evidence for the alternative hypothesis (BF10 > 3) or null hypothesis (BF10 < 1/3) is .90. This procedure resulted in a maximal sample size of n = 1700, see RMarkdown script at https://osf.io/4wpv6/.		SBF+maxN approach used for the hypotheses.	observable (i.e., association of small/large numbers with the right/left, respectively), which is highly unlikely.	
Can the presence of a MARC effect be replicated in PJ, and can its absence be replicated in MC?	**Replication ~~Manipulation~~ Check 2:** A MARC effect will arise in PJ, but not in MC, because the activation of parity seems not to be automatic when parity is task-irrelevant		1. Regression of dRTs on contrast-coded number parity (i.e., -0.5 for odd and +0.5 for even numbers; see models MC-3 and PJ-3) for each task separately per participant (as in Nuerk et al., 2004) 2. Two two-sided Bayesian one-sample *t*-tests of MARC slopes against zero for each task separately	~~Manipulation~~ Replication checks aim at replicating observations that are typically made in the used paradigm, instead of testing new hypotheses~~thus serving to validate our applied methodology~~. The sensitivity of the tests, however, depends on the final sample size determined by the SBF+maxN approach used for the hypotheses.	If evidence is found for the MARC slopes to differ from zero and to be negative, the MARC effect is detectable. This would be in line with previous literature for PJ. If evidence is found against the MARC slopes to differ from zero, no MARC effect is observable. This would be in line with previous literature for MC. If evidence is found for the MARC slopes to differ from zero and to be positive, a reversed MARC effect is observable (i.e., association of odd/even numbers with the right/left, respectively), which is highly unlikely.	The MARC effect is usually detected in PJ, but not in MC. A theory accounting for this is that the spatial mapping of number parity is automatic, but only when number parity is activated by the task instructions. However, number parity is not activated when being task-irrelevant, thus no spatial mapping occurs for it. We expect a replication in this study as well, and not finding the MARC effect would be rather surprising in a large Western sample. However, because a considerable proportion of Western individuals reveals a reversed MARC effect (e.g., descriptively 60% regular and 40% reversed in Cipora, Soltanlou, et al., 2019), not finding evidence for the regular MARC effect

						would not preclude further analyses.
Can responses be replicated to be faster in MC than in PJ?	**Replication ~~Manipulation~~ Check 3:** RTs are shorter in MC than in PJ.		One two-sided Bayesian paired *t*-test to compare mean RTs per participant between tasks against zero	Replication ~~Manipulation~~ checks aim at replicating observations that are typically made in the used paradigm, instead of testing new hypotheses~~thus serving to validate our applied methodology~~. The sensitivity of the tests, however, depends on the final sample size determined by the SBF+maxN approach used for the hypotheses.	Evidence for faster responses in MC than in PJ would be in line with previous literature.	The processing of number magnitude is highly automatized and a primitive in numerical cognition (Tzelgov et al., 2015). In contrast, the processing of number parity is not as highly automatized; it needs to be executed intentionally and is therefore slower. Finding no difference in RTs between tasks or even the reversed pattern would be highly surprising.
Can the ~~n~~Numerical ~~d~~Distance ~~e~~Effect (NDE) in MC be replicated?	**Replication ~~Manipulation~~ Check 4:** An NDE will arise in MC (i.e., faster reactions with increasing numerical distance between the stimulus and the reference number 5).		1. Regression of RTs on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9) for each participant separately (as in Hohol et al., 2020) 2. One two-sided Bayesian one-sample *t*-test of numerical-distance slopes against zero	Replication ~~Manipulation~~ checks aim at replicating observations that are typically made in the used paradigm, instead of testing new hypotheses~~thus serving to validate our applied methodology~~. The sensitivity of the tests, however, depends on the final sample size determined by the SBF+maxN approach used for the hypotheses.	If evidence is found for the NDE slopes to differ from zero and to be negative, the NDE is detected, which would be in line with previous literature. If evidence is found against the NDE slopes to differ from zero, no NDE effect is observable. If evidence is found for the NDE slopes to differ from zero and to be positive, a reversed NDE is observable (i.e., faster reactions with increasing	The NDE is usually detected in MC. We strongly expect to find it in this study as well, especially with our large sample size. Not finding the NDE would speak against its robustness and be very surprising.

					numerical distance between the stimulus and the reference number 5), which is highly unlikely.	
What is the shape of the SNARC effect in MC?~~Does the shape of the SNARC effect differ between the MC and PJ task?~~	**Hypothesis 1~~a~~:** The MC-SNARC will be categorical. ~~**Hypothesis 1b:**~~ ~~The PJ-SNARC will be continuous.~~		1. Regression of dRTs on continuous magnitude (model~~s~~ MC-1 ~~and PJ-1~~) and on categorical (contrast-coded with -0.5 for small and +0.5 for large numbers) magnitude (model~~s~~ MC-2 ~~and PJ-2~~) 2. Logit-transformation of R^2 for each model for each participant separately to approximate two normal distributions 3. Comparison of the logit-transformed R^2 between MC-1 and MC-2 ~~and between PJ-1 and PJ-2~~ in ~~two~~ a two-sided paired *t*-test~~s~~ (as in Koch et al., 2023) Additional purely Bayesian approach: 2. Leave-one-out cross validation to determine which of the two predictors better fits the data (using the R package brms by	The main goal of this study is to determine the shape of the SNARC effect in MC and PJ (Hypotheses 1~~a~~ and ~~1b~~2). For comparing the fit of a continuous and a categorical model against each other in MC and PJ separately, the effect size of interest (ESOI) must be chosen in a standardized unit. We determined Cohen's *d* = 0.2 as ESOI, because it is considered to be a small effect (Cohen, 1988). This ESOI was used to determine the maximal sample size for the SBF+maxN approach.	Evidence for a higher logit-transformed R^2 for the categorical (MC-2) than continuous (MC-1) MC-SNARC speaks for a stepwise shape of the MC-SNARC (in line with Hypothesis 1~~a~~). ~~Evidence for a higher logit-transformed R^2 for the continuous (PJ-1) than categorical (PJ-2) PJ-SNARC speaks for a linear shape of the PJ-SNARC (in line with Hypothesis 1b).~~	Results from many previous studies lead to the hypothesis that the SNARC effect is categorical in MC, but continuous in PJ. In this thorough investigation with a sufficiently large sample, we will investigate this difference systematically by comparing the fit of the two statistical models. ~~Finding evidence for a better fit of the continuous models (MC-1 and PJ-1) in both tasks would prove this theory wrong.~~

			Buerkner, 2017, and the R package loo by Vethari et al., 2017)			
What is the shape of the SNARC effect in PJ?	**Hypothesis 2:** The PJ-SNARC will be continuous.		1. Regression of dRTs on continuous magnitude (model PJ-1) and on categorical (contrast-coded with -0.5 for small and +0.5 for large numbers) magnitude (model PJ-2) 2. Logit-transformation of R^2 for each model for each participant separately to approximate two normal distributions 3. Comparison of the logit-transformed R^2 between PJ-1 and PJ-2 in a two-sided paired *t*-test (as in Koch et al., 2023) Additional purely Bayesian approach: 2. Leave-one-out cross validation to determine which of the two predictors better fits the data (using the R package brms by Buerkner, 2017, and the R package loo by Vethari et al., 2017)		Evidence for a higher logit-transformed R^2 for the continuous (PJ-1) than categorical (PJ-2) PJ-SNARC speaks for a linear shape of the PJ-SNARC (in line with Hypothesis 2).	

~~Does task-order influence the SNARC effects in MC and PJ?~~	~~**Hypothesis 2a:**~~ ~~The MC-SNARC is stronger when participants complete MC second (after PJ).~~ ~~**Hypothesis 2b:**~~ ~~The PJ-SNARC is stronger when participants complete PJ second (after MC).~~		~~Two two-sided independent samples *t* tests of differences between the MC-PJ task-order (Conditions 1 and 2) and the PJ-MC task-order (Conditions 3 and 4) in (a) MC-SNARC slopes and (b) PJ-SNARC slopes~~ ~~Note that according to the results for Hypotheses 1a and 1b, the predictor with the better model fit will be used for determining the SNARC effect: continuous (models MC-1 and PJ-1) or categorical (models MC-2 and PJ-2).~~	~~As explained above, we chose Cohen's *d* = 0.2 as ESOI for Hypotheses 1a and 1b. We decided to use the same ESOI for Hypotheses 2a and 2b for consistency reasons.~~	~~(a) Evidence for more negative MC-SNARC slopes in the PJ-MC task-order (Conditions 3 and 4) than in the MC-PJ task-order (Conditions 1 and 2) would speak for a stronger spatial mapping of number magnitude in MC when this has already been activated in PJ beforehand.~~ ~~(b) Evidence for more negative PJ-SNARC slopes in the MC-PJ task-order (Conditions 1 and 2) than in the PJ-MC task-order (Conditions 3 and 4) would speak for a stronger spatial mapping of number magnitude in PJ when this has already been activated in MC beforehand.~~	~~A stronger SNARC effect in the second than in the first task (no matter whether MC or PJ is the first or second task) is what one would expect because of the first task activating the automatic processing of number magnitude and its mental mapping onto space.~~ ~~Finding evidence against this task-order effect or even evidence for the reversed pattern would speak for an alleviated or even reversed SNARC effect, which could be due to habituation and decreasing attention after a certain number of trials or due to enhanced focusing on the response-relevant feature while more strongly ignoring the response-irrelevant feature.~~
Can the numerical size effect (NSE) be found in both tasks, and does it differ between tasks regarding its strength?	**Hypothesis 3a:** An NSE will arise in both tasks. **Hypothesis 3b:**		1. Regression of RTs on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9) for each participant	As explained above, we chose Cohen's *d* = 0.2 as ESOI for Hypotheses 1a and 1b. We decided to use the same ESOI for Hypotheses 3a and 3b for consistency reasons.	If evidence is found for the NSE slopes to differ from zero and to be positive, the NSE is detected, which would be in line with previous literature.	The NSE is usually detected in MC and PJ. We expect to find it in this study as well, especially with our large sample size. Not finding the NSE would speak against its robustness and be

	The NSE will be stronger in MC than in PJ.		separately (as in Hohol et al., 2020) 2. One two-sided Bayesian one-sample *t*-test of continuous-magnitude slopes against zero 3. One two-sided Bayesian paired *t*-test between continuous-magnitude slopes		If evidence is found against the NSE slopes to differ from zero, no NSE effect is observable. If evidence is found for the NSE slopes to differ from zero and to be negative, a reversed NSE is observable (i.e., faster reactions with increasing numerical magnitude), which is highly unlikely.	surprising, although some individuals seem to reveal a reversed NSE (showing that it is less consistent as the NDE; Hohol et al., 2020).

Notes. For an overview of all regression models, see Table 2 in the manuscript. BF_{10} refers to the Bayes Factor, i.e., probability of the alternative hypothesis over the null hypothesis.

Shape of SNARC: How task-dependent are Spatial-Numerical Associations?

A highly powered online experiment

Lilly Roth

Version 1: May 27th, 2024

This script provides sample size estimations for our Registered Report on the task dependency of spatial-numerical associations and more precisely of the SNARC effect (Dehaene et al., 1993, https://doi.org/10.1037/0096-3445.122.3.371). We expect the SNARC effect in a bimanual response setup with numbers from 1 to 9 (excluding 5) to differ between magnitude classification (MC; judging smaller vs. larger than 5) and parity judgment (PJ; judging odd vs. even).

We decided to calculate Bayes Factors (BFs) in our data analysis to be able to quantify evidence both in favor and against differences in the SNARC effect between MC and PJ and their the relationship. We will interpret BFs as proposed by Dienes (2021, https://doi.org/10.1037/cns0000258): A resulting BF10, which is the BF for the alternative hypothesis (H1) over the null hypothesis (H0), will be treated as moderate or strong evidence **for H1** if it is greater than 3 or 10, respectively, and as moderate or strong evidence **for H0** if it is smaller than 1/3 or 1/10, respectively.

We will make use of the "Sequential Bayes Factor with maximal n" (SBF+maxN) approach described by Schönbrodt and Wagenmakers (2018, https://doi.org/10.3758/s13423-017-1230-y) with recruitment steps of 50, and determine the maximal sample size in this script. For this, we ran simulations of the probability to obtain evidence for a true underlying effect of the size which we consider to be minimally relevant and of the probability to obtain evidence against a truly absent effect, striving for these probabilities to be as high as 0.90.

We calculated Bayes Factors with the R package *BayesFactor* by Morey et al. (2015, https://CRAN.R-project.org/package=BayesFactor). All Bayesian tests will be run two-sided. This script was created with the R packages *rmarkdown* by Allaire et al. (2023, https://cran.r-project.org/web/packages/rmarkdown/index.html) and *knitR* by Xie et al. (2023, https://cran.r-project.org/web/packages/knitr/index.html). The script can be downloaded from https://osf.io/4wpv6/.

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

Parameters for simulations

Minimal effect size of interest (ESOI) for magnitude classification (MC) and parity judgment (PJ)

The ESOI we chose for this study must be expressed in a standardized unit, as the main aim is to compare the model fits (logit-transformed R^2) for Hypothesis 1. Specifically, we chose to use a small effect size expressed as Cohen's $d = 0.2$:

```
esoi <- 0.2
```

Note that the same ESOI will be used for Hypothesis 2 (task-order effects) and Hypothesis 3 (Numerical Size Effect). Smaller effect sizes would not be practically meaningful, because $d = 0.2$ corresponds to around only 1% of explained variance (calculated according to Ruscio, 2008, using the conversion formula assuming equal-sized groups, see their Table 2).

In the following, we estimate what SNARC slopes the ESOI $d = 0.2$ corresponds to (continuous and categorical slopes in MC and continuous slopes in PJ). That is, we convert the effect size from a standardized unit to the practical unit. This can be found out by multiplying Cohen's d with a plausible standard deviation. We looked up **previously observed standard deviations** (and chose rather conservative values):

Continuous number-magnitude slope in MC

The following standard deviations are given in milliseconds:

26 in Bachot et al. (2005),
11 in Cheung et al. (2015),
13 on average in Deng et al. (2017),
11 in Fattorini et al. (2015),
13 in Georges et al. (2017),
6 in Ito & Hatta (2004),
25 on average in Mourad & Leth-Steensen (2017),
22 in healthy controls in Pinto et al. (2021),
24 in healthy controls in van Dijck et al. (2012)

The continuous MC-SNARC (i.e., increase in right- over left-hand advantage in milliseconds per increase in number magnitude of 1 unit) that corresponds to Cohen's $d = 0.2$ is approximately:

```
SD.MC.continuous <- 20
-esoi * SD.MC.continuous
```

```
## [1] -4
```

Categorical number-magnitude slope in MC

The following standard deviations are given in milliseconds:

41 in Didino et al. (2019) (SE = 7.39 for 32 participants),
51 in Hohol et al. (2020)

The categorical MC-SNARC (i.e., increase in right- over left-hand advantage in milliseconds for the switch from small to large numbers in number magnitude of 1 unit) that corresponds to Cohen's $d = 0.2$ is approximately:

```r
SD.MC.categorical <- 50
-esoi * SD.MC.categorical
```

```
## [1] -10
```

Continuous number-magnitude slope in PJ

The following standard deviations are given in milliseconds:

12 in Shaki, fischer, & Petrusic (2009),
9 in Fattorini, Pinto, Rotondaro, and Doricchi (2015),
10 in Cipora, Soltanlou, Reips, and Nuerk (2019)

In an extensive reanalysis of existing PJ datasets, Cipora, van Dijck, et al. (2019; https://doi.org/10.31234/osf.io/bwyr3) report SD for unstandardized continuous SNARC slopes from 18 previous studies between 5.81 and 12.75.

The continuous PJ-SNARC (i.e., increase in right- over left-hand advantage in milliseconds per increase in number magnitude of 1 unit) that corresponds to Cohen's $d = 0.2$ is approximately:

```r
SD.PJ.continuous <- 10
-esoi * SD.PJ.continuous
```

```
## [1] -2
```

To sum up, as ESOI, a small effect size expressed in a standardized unit was chosen, namely Cohen's $d = 0.2$. This corresponds to a continuous MC-SNARC of -4, to a categorical MC-SNARC of -10, and a continuous PJ-SNARC of -2.

Simulation loops

We also need to set a parameter for the number of samples to be drawn in the Bayes Factor simulations for each test:

```r
rep <- 5000
```

One-sample *t*-test / paired *t*-test:

We will need one-sample *t*-tests for Hypotheses 1a, 1b, and 3b. We will need a paired *t*-test for Hypothesis 3a.

We try out different sample sizes (n.onesample.H1), simulate data for these sample sizes with the ESOI (note that Cohen's d follows the standard normal distribution and hence $sd = 1$), calculate the Bayes Factor (BF.onesample.H1) for a test in each of 5000 iterations, and estimate the probability for finding at least moderate evidence for a true underlying effect by the proportion of iterations revealing at least moderate evidence (p.onesample.H1):

```r
n.onesample.H1 <- 440
BF.onesample.H1 <- replicate(rep, {
  d <- rnorm(n = n.onesample.H1, mean = esoi, sd = 1)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
(p.onesample.H1 <- format(round(mean(BF.onesample.H1 > 3), 3), nsmall = 3))
```

```
## [1] "0.910"
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 > 3) for the minimally relevant effect of $d = 0.2$, $n = 440$ datasets need to be collected.

Again, we try out different sample sizes (n.onesample.H0), simulate data for these sample sizes without any true underlying effect (mean = 0), calculate the Bayes Factor (BF.onesample.H1) for a test in each of 5000 iterations, and estimate the probability for finding at least moderate evidence against a truly absent effect by the proportion of iterations revealing at least moderate evidence (p.onesample.H0):

```
n.onesample.H0 <- 160
BF.onesample.H0 <- replicate(rep, {
  d <- rnorm(n = n.onesample.H0, mean = 0, sd = 1)
  extractBF(ttestBF(d, mu = 0, alternative = "two.sided"))$bf
})
(p.onesample.H0 <- format(round(mean(BF.onesample.H0 < 1/3), 3), nsmall = 3))
```

```
## [1] "0.900"
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 < 1/3) against a non-existent effect of $d = 0$, 160 datasets need to be collected.

Independent-samples *t*-test:

We will need independent-samples *t*-test for Hypotheses 2a and 2b.

We try out different sample sizes (n1.twosamples.H1 and n2.twosamples.H1), simulate data for these sample sizes differing by the ESOI (note that Cohen's d follows the standard normal distribution and hence $sd = 1$), calculate the Bayes Factor (BF.twosamples.H1) for a comparison between the two samples in each of 5000 iterations, and estimate the probability for finding at least moderate evidence for a true underlying difference by the proportion of iterations revealing at least moderate evidence (p.twosamples.H1):

```
n1.twosamples.H1 <- 850 # size of one subsample
n2.twosamples.H1 <- n1.twosamples.H1 # size of the other subsample

BF.twosamples.H1 <- replicate(rep, {
  subsample.1 <- rnorm(n1.twosamples.H1, mean = 0 + esoi, sd = 1)
  subsample.2 <- rnorm(n2.twosamples.H1, mean = 0, sd = 1)
  extractBF(ttestBF(x = subsample.1, y = subsample.2, mu = 0, alternative = "two.sided"))$bf
})
(p.twosamples.H1 <- format(round(mean(BF.twosamples.H1 > 3), 3), nsmall = 3))
```

```
## [1] "0.904"
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 > 3) for the minimally relevant difference between counterbalanced orders of $d = 0.2$, 850 datasets need to be collected for each order.

Again, we try out different sample sizes (n1.twosamples.H0 and n2.twosamples.H0), simulate data for these sample sizes without any true underlying difference, calculate the Bayes Factor (BF.twosamples.H0) for a

comparison between the two samples in each of 5000 iterations, and estimate the probability for finding at least moderate evidence against a truly absent difference by the proportion of iterations revealing at least moderate evidence (p.twosamples.H0):

```
n1.twosamples.H0 <- 340 # size of one subsample
n2.twosamples.H0 <- n1.twosamples.H0 # size of the other subsample

BF.twosamples.H0 <- replicate(rep, {
  subsample.1 <- rnorm(n1.twosamples.H0, mean = 0, sd = 1)
  subsample.2 <- rnorm(n2.twosamples.H0, mean = 0, sd = 1)
  extractBF(ttestBF(x = subsample.1, y = subsample.2, mu = 0, alternative = "two.sided"))$bf
})
(p.twosamples.H0 <- format(round(mean(BF.twosamples.H0 < 1/3), 3), nsmall = 3))
```

```
## [1] "0.913"
```

In order to achieve a probability of 0.90 to find at least moderate evidence (BF10 $< 1/3$) against a non-existent difference between counterbalanced orders of $d = 0$, 340 datasets need to be collected for each order.

Summary and conclusion

By simulating the probability of obtaining evidence in favor of a true underlying effect of $d = 0.2$, and against a truly absent effect of $d = 0$, we found that to achieve 0.90, we need the following sample sizes for the tests:

One-sample *t*-test / paired *t*-test:
evidence for H1: 440
evidence for H0: 160

Independent-samples *t*-test:
evidence for H1: 850
evidence for H0: 340
Note that this sample size is required per subsample, and thus needs to be doubled for the total sample size.

Largest required sample size
The largest total sample size required to test our hypotheses is 2 * 850 = 1700 for the independent-samples *t*-test. Thus, we will use this sample size as the maximal sample size for the SBF+maxN sampling approach with recruitment steps of 50.