



The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects

Krzysztof Cipora^{1,2} · Mojtaba Soltanlou^{1,2,3} · Ulf-Dietrich Reips⁴ · Hans-Christoph Nuerk^{1,2,3}

Published online: 25 February 2019
© The Psychonomic Society, Inc. 2019

Abstract

The Spatial–Numerical Association of Response Codes (SNARC) effect (i.e., faster reactions to small/large numbers on the left-/right-hand side) is usually observed along with the linguistic Markedness of Response Codes (MARC) effect—that is, faster left-/right-hand responses to odd/even numbers. The SNARC effect is one of the most thoroughly investigated phenomena in numerical cognition. However, almost all SNARC and MARC studies to date were conducted with sample sizes smaller than 100. Here we report on a study with 1,156 participants from various linguistic and cultural backgrounds performing a typical parity judgment task. We investigated whether (1) the SNARC and MARC effects can be observed in an online setup, (2) the properties of these effects observed online are similar to those observed in laboratory setups, (3) the effects are reliable, and (4) they are valid. We found robust SNARC and MARC effects. Their magnitude and reliabilities were comparable to values previously reported in in-lab studies. Furthermore, we reproduced commonly observed validity correlations of the SNARC and MARC effects. Namely, SNARC and MARC correlated with mean reaction times and intraindividual variability in reaction times. Additionally, we found interindividual differences in the SNARC and MARC effects (e.g., finger-counting routines for the SNARC and handedness for the MARC). Large-scale testing via web-based data acquisition not only produces SNARC and MARC effects and validity correlations similar to those from small, in-lab studies, but also reveals substantial insights with regard to interindividual differences that usually cannot be revealed in the offline laboratory, due to power considerations.

Keywords SNARC effect · MARC effect · Online setting · Individual differences · Web experiment

With the present article we will demonstrate how the power of an unusually large dataset gained from an Internet-based experiment can advance research in a long-standing field related to two cognitive phenomena, the SNARC and MARC effects. For the first time, in this field of flexible cognitive effects that had exclusively been investigated using small-sample

experimental research, we were able to investigate with sufficient power hypotheses about moderating individual differences that had previously lacked replicability or otherwise remained speculative.

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01213-5>) contains supplementary material, which is available to authorized users.

✉ Krzysztof Cipora
krzysztof.cipora@uni-tuebingen.de

¹ Department of Psychology, University of Tuebingen, Tuebingen, Germany

² LEAD Graduate School & Research Network, University of Tuebingen, Tuebingen, Germany

³ Leibniz-Institut für Wissensmedien, Tuebingen, Germany

⁴ Department of Psychology, University of Konstanz, Konstanz, Germany

The SNARC effect

When making bimanual judgments on the parity of single-digit numbers, participants respond faster to small magnitude numbers with their left hand, and to large magnitude numbers with their right hand (Dehaene, Bossini, & Giraux, 1993). This phenomenon was named the *Spatial–Numerical Association of Response Codes* (SNARC) effect. Importantly, this magnitude-related effect was observed in a task in which the magnitude itself was not a relevant dimension (i.e., participants were judging the parity of numbers). Therefore, it is interpreted as evidence for automatic magnitude processing (see, e.g., Fias, Brysbaert, Geypens, & D’Ydewalle, 1996). Subsequent experiments have shown that the association is not restricted to

bimanual judgments, and can be observed in a unimanual setup (Priftis, Zorzi, Meneghello, Marenzi, & Umiltà, 2006), as well as when responses are given by means of the feet (Schwarz & Müller, 2006), eye movements (Fischer, Warlop, Hill, & Fias, 2004), or a single finger, in a pointing task (Fischer, 2003b). The SNARC effect can also be observed irrespective of the modality in which numbers are presented (visual presentation of Arabic numbers, number words, dice patterns, or numbers presented auditorily; Nuerk, Wood, & Willmes, 2005b). It can also be observed in the case of multidigit numbers (Tlauka, 2002) and negative numbers (Fischer, 2003a; Nuerk, Iversen, & Willmes, 2004), as well as with nonsymbolic notations (Nemeh, Humberstone, Yates, & Reeve, 2018).

The directionality of the SNARC effect has been linked to individuals' reading direction. In Arabic speakers the effect is reversed (i.e., small magnitude numbers are associated with the right-hand side, and large magnitude numbers are associated with the left-hand side). Not only reading, but also counting, direction plays a role for the direction of the SNARC. In Hebrew speakers, who write from right to left but count from left to right, the effect does not exist (see Shaki, Fischer, & Petrusic, 2009, for a systematic investigation of these influences). The missing SNARC in Hebrew speakers has been attributed to the counteracting influences of opposite reading and counting directions. Nevertheless, a recent study has demonstrated a regular left-to-right SNARC in Hebrew speakers (Zohar-Shai, Tzelgov, Karni, & Rubinsten, 2017). The idea that the directionality of SNARC is determined by reading direction has been challenged by studies showing associations of small-magnitude numbers with the left-hand side and large-magnitude numbers with the right-hand side in newly hatched chicks (Rugani, Vallortigara, Priftis, & Regolin, 2015) and preliterate children (Patro & Haman, 2012).¹ Without drawing trans-species inferences, which are limited in several ways (Patro & Nuerk, 2017), the existence of a directional SNARC in preliterate children can be accounted for by several mechanisms of cultural transmission during child–parent interactions (Patro, Nuerk, & Cress, 2016b). Indeed, a training study by Patro, Fischer, Nuerk, and Cress (2016a) showed that Spatial–Numerical Associations (SNAs; Cipora, Patro & Nuerk, 2015) like the SNARC effect can be induced by nonnumerical directional training (see Göbel, McCrink, Fischer, & Shaki, 2018, for cultural variations of similar trainings). For these reasons, the reading direction explanation needs at least more differentiation. Mechanisms different from pure reading must be assumed to explain culturally different SNAs even in children (Nuerk et al., 2015).

¹ As we mentioned above, the SNARC effect can be observed for both symbolic and nonsymbolic notations. In this study we focused on symbolic notation using Arabic numbers.

In the past 25 years, the SNARC effect has been subject to intense investigation (Fischer & Shaki, 2014, for a current review; Wood, Willmes, Nuerk, & Fischer, 2008, for a meta-analysis). It is considered strong evidence for spatial mapping of numerical representation in the brain, and is one of the most thoroughly investigated SNAs. However, considerable individual differences in the magnitude of the SNARC have been reported. About 70%–80% of individuals reveal the SNARC effect, indexed as negative regression slopes in which dRTs (differences in reaction times for the right hand minus the left hand) are regressed on numerical magnitude (see Fias et al., 1996, for a method description). This 70%–80% prevalence, as compared to the prevalences of other well-established cognitive phenomena, is relatively low. For instance, in the case of the Stroop or the flanker effect, the effect in the expected direction (e.g., a reaction time advantage for compatible vs. incompatible trials) can be observed in virtually all participants (see, e.g., the raw data shared by Hedge, Powell, & Sumner, 2018). Furthermore, there is a large variation in slopes within individuals who reveal the SNARC effect. What is more, SNARC effects usually do not correlate well across different notations (Wood, Nuerk, & Willmes, 2006), suggesting individual differences for different notations.

The sources of individual differences in the SNARC and MARC effects are largely unknown, yet several factors have already been postulated to play roles. These are (1) reaction time characteristics in the task measuring the SNARC effect (i.e., a stronger SNARC in individuals who respond more slowly and whose reaction times are more varied; see Cipora & Nuerk, 2013), (2) individuals' finger-counting habits (i.e., a stronger and less variable SNARC in individuals who start the finger-counting sequence from the left hand; Fischer, 2008), (3) individuals' inhibition capacities (i.e., a stronger SNARC in individuals with poorer inhibition of irrelevant semantic features of stimuli; Hoffmann, Pigat, & Schiltz, 2014b), (4) gender (i.e., males revealing stronger SNARC effects; Bull, Cleland, & Mitchell, 2013), (5) spatial capacities (i.e., a stronger SNARC in individuals with stronger mental rotation effects of two-dimensional objects; Viarouge, Hubbard, & McCandliss, 2014), (6) math expertise (i.e., a weaker SNARC in math experts; see Cipora et al., 2016; Cipora, Schroeder, & Nuerk, 2019a, for an overview), (7) math anxiety (i.e., a stronger SNARC in individuals with higher math anxiety; Georges, Hoffmann, & Schiltz, 2016), and (8) reading direction in a culture (see above). Furthermore, the SNARC effect is also prone to numerous situated influences, which can modulate, cancel, or even reverse its direction (Cipora, Patro, & Nuerk, 2018).

Although many studies on interindividual differences in the SNARC have been conducted, the studies on those differences were frequently not replicated, in line with current general concerns about reproducibility in psychology (in line with current general concerns about reproducibility in

psychology; e.g., Open Science Collaboration, 2015), or were largely inconclusive. Possible reasons include that the reported correlations or between-group differences have mostly been small to moderate (in most cases between .25 to .44; e.g., Hoffmann, Pigat, & Schiltz, 2014b); in several instances, the reliability of the SNARC was relatively low (e.g., Georges, Hoffmann, & Schiltz, 2017); and the tested samples have been relatively small (yielding insufficient power). Namely, the sample sizes in published SNARC studies typically have not exceeded 100 participants. Of course, this is particularly a problem when we investigate the moderation (by individual differences) of a rather flexible effect (Cipora, Schroeder, Soltanlou, & Nuerk, 2019b).

To provide more conclusive evidence about the individual differences influencing the SNARC effect, much greater power, and therefore, large-scale studies testing heterogeneous samples, are needed. Online experimenting is a viable method to reach such a goal (Reips, 2002, 2012).

The MARC effect

Nuerk, Iversen, and Willmes (2004) observed that in a parity judgment task, another space–number association can be observed, apart from the SNARC effect (see also Berch, Foley, Hill, & Ryan, 1999; Huber et al., 2015). Namely, responses to odd numbers are faster on the left-hand side, and responses to even numbers are faster on the right-hand side. This phenomenon is referred to as the Markedness of Response Codes (MARC) effect and is explained in terms of markedness.

The term *markedness* was developed in linguistics. It denotes the fact that most adjective pairs have an unmarked (i.e., default or basic) and a marked (i.e., derived) form. Different forms of markedness have been mentioned. The standard form is formal markedness. In such cases, marked forms are derived by adding prefixes or suffixes to unmarked forms (e.g., “happy” vs. “unhappy”).² Lyons (1968) mentions also semantic and distributive markedness. Distributive markedness is related to frequency. Usually, more frequent words represent the base form, and less frequent ones, the marked, derived form. Unmarked forms are used more often in language (e.g., usually we ask “How long does it take?” or “How old are you?”, rather than “How short does it take?” or “How young are you?”). In several cultures the right side can be considered as basic, unmarked. For instance, the seat on the right side of the host is the most prestigious one in several cultures; people shake their right hands when they greet; and so forth.

Finally, there is semantic markedness: Here, the marked word is more specific than the unmarked word. For instance, in English the word “lion” can refer both to a male and a female animal, whereas the marked term “lioness” refers specifically to a female animal.

Terms also refer to number parity as *unmarked* and *marked*. In several languages, words referring to *even* are the unmarked, basic forms, whereas words referring to *odd* are marked. This may reflect valence (the English word *odd* can also mean “weird”), or it can be seen in syntax: In several languages, the word referring to odd numbers is built by adding a negation prefix to the word for even numbers (e.g., in English, one might contrast “even” with “uneven” as a synonym of “odd”; in other languages, these are the canonical forms, such as “gerade” vs. “ungerade” in German or “parzysty” vs. “nieparzysty” in Polish).

Features determining the markedness can overlap (e.g., unmarked forms are more frequent and differ from the marked form by missing the prefix or suffix; e.g., “happy” vs. “unhappy”). For this reason, it is hard to determine which aspect is the most decisive for markedness in a particular case. So far, one can only speculate that markedness of the right versus left can reflect the fact that most of the population is right-handed, and therefore positive valence is associated with that side. However, as was shown by Huber et al. (2015), such associations can be present rather at the individual level (i.e., refer to an individual’s hand dominance), since the direction of the MARC effect is reversed in left-handers. In the case of number parity, divisibility by 2 may make even numbers less problematic, and therefore make them unmarked.

One of the most plausible explanations of the MARC effect is that responses are facilitated if the markedness of parity status of a number is the same as the markedness of the response side (i.e., odd + left and even + right) as compared to opposite response-to-key mapping (Proctor & Cho, 2006, for a general theoretical framework).

The MARC effect is more stimulus-specific than the SNARC effect. It is usually acquired when linguistic aspects are made more salient (e.g., when numbers are presented as number words; Nuerk et al., 2004). In some studies it was not replicated if the numbers were presented as Arabic digits, possibly due to a small effect size, low power, and interindividual variability (e.g., handedness; Huber et al., 2015). The MARC effect has so far received relatively less attention than the SNARC effect. Individual differences and correlates have been very rarely investigated. To the best of our knowledge, the prevalence of the MARC effect has not been discussed: The effect itself can be quantified together with the SNARC effect by means of multiple regression, in which dRTs are regressed on numerical magnitude and contrast-coded parity (Nuerk et al., 2004).

² Ironically, the terms referring to markedness are contradictory to the theory: That is, the basic forms are termed *unmarked* (rather than *marked*).

Large-scale experimentation via the Internet

A recent review of the over 20 years of Internet-based experimenting suggests that the web is a method for conducting valid psychological studies on a large scale (Krantz & Reips, 2017). For instance, several cognitive psychology phenomena observed at the level of reaction times, such as the Stroop effect, task-switching costs, the flanker effect, and the Simon effect, can successfully be reproduced in an online setup (Crump, McDonnell, & Gureckis, 2013). Moreover, in the field of numerical cognition, Huber, Nuerk, Reips, and Soltanlou (2017) replicated the robust unit-decade compatibility effect and distance effect in two-digit number comparison tasks in an online setting. This might suggest that online experimenting can also be applied to investigating the SNARC and MARC effects. To check the validity and feasibility of online experimentation methods for these phenomena and their individual differences, the following methodological and empirical questions shall be answered:

Are the SNARC and MARC effects *reproducible* with online methods? This question is important because only when this method is successful can further theoretical questions about these phenomena be examined with online methods in the future.

Are the strengths of the SNARC and MARC effects observed in an online setup *comparable to results obtained in-lab*? Principal replication of significance is not sufficient, because multiple codes (such as the order of numbers, organized along the so-called mental number line, or the order of entities currently stored in working memory) may underlie the SNARC effect (cf. Schroeder, Nuerk, & Plewnia, 2017). If some of these codes are less salient in online studies, the SNARC effect would be different, and online investigations of the SNARC effect might tackle just a subset of the underlying cognitive mechanisms. The same caveat potentially applies to the MARC.

Are the SNARC and MARC effects reliable when tested online? Any future investigations on sources of individual differences in the SNARC will require testing the *reliability* of the effect and checking whether reliability estimates are comparable to those observed in in-lab setups. Hedge et al. (2018) rightly argued that highly reproducible cognitive psychology phenomena are not necessarily suited to be individual differences measures. A group effect in cognitive psychology usually is most stable when interindividual variance is low (because it is treated as error variance). However, reliability and correlations with other variables of interindividual differences require systematic variance between individuals, because without such variance there can be no covariance. In psychometrics, reliability and validity are usually reduced when there is less interindividual variation in the sample. Therefore, a very stable cognitive effect with low interindividual variability is less suited for examining interindividual

differences. Our large-scale online study examined whether the SNARC and MARC effects are reliable and suited for examining interindividual differences. Moreover, the reliability of a given effect may differ considerably between different groups (Cooper, Gonthier, Barch, & Braver, 2017).

Are the online SNARC and MARC effects *valid*? To answer this question, it is important to check for validity features such as correlations with measures that the SNARC and MARC correlate with in in-lab setups. These measures consider reaction time characteristics (the evidence for correlations with other measures is more scarce, and conflicting results have been obtained regarding such correlations). Only if established correlations can be replicated may the online SNARC and MARC effects be viewed as valid.

Materials and method

Participants

We applied the multiple site entry technique (Reips, 2002)—that is, participants were recruited from different sources (see the section below on the specifics of the online experiment for details). The initial sample was 1,509 individuals. First, the database was screened on the basis of the following exclusion criteria: Initially, the data from 334 participants who did not finish the experiment were not included in the analysis. Subsequently, the data from 18 participants, who declared that they did not wish to participate but only wanted to see the procedure (seriousness check; e.g., Aust, Diedenhofen, Ullrich, & Musch, 2013; Bayram, 2018; Reips, 2000), were excluded. The experimental instructions clearly stated that minors were not allowed to participate; therefore, the data from eight participants who indicated their age to be below 18 years were excluded. Thereafter, the data from ten participants who reported running the experiment on devices other than a PC computer or laptop (i.e., smartphone, tablet, or e-reader) were also excluded. Finally, the data from nine participants who assessed the environment in which they performed the task to be *very noisy* or *extremely noisy* were excluded. The remaining data from 1,130 participants were further analyzed.

Average reported age was 28 years ($SD = 9.66$ years, ranging from 18 to 77 years). Our sample included 742 female, 310 male, and five transgender participants. In all, 946 participants reported being right-handed, 79 left-handed, and 29 ambidextrous, and two participants did not specify their handedness. Participants were informed about a possibility to take part in a lottery in which ten randomly chosen participants would receive an Amazon voucher for participating in the experiment. Providing a contact email address was necessary to take part in a lottery, but it was not mandatory for participation in the experiment. All experimental procedures were

approved of by the local ethics committee of the Leibniz-Institut für Wissensmedien.

Stimuli

A computerized parity judgment task was utilized. Participants were asked to decide on the parity of numbers presented on the screen using the left and right arrow keys on a standard computer keyboard. Both speed and accuracy were stressed. The task comprised two blocks with response key mappings reversed. The order of blocks was consistent for all participants.³ Where the left arrow key was assigned for odd numbers and the right arrow key for even numbers in the first block, these assignments were reversed in the second block. The participants were asked to use only their index fingers. The numbers 1, 2, 3, 4, 6, 7, 8, and 9 were used. Black stimuli (font size 96 pixels) were presented at the center of a white background. Each number was presented ten times within each block, for a total of 80 trials.

Each block was preceded by a practice session of eight trials, to familiarize participants with the task. During the practice session, the required response mapping was indicated at the top of the screen. The order of trials was pseudorandomized, with the restriction that (i) no more than three odd or even numbers could appear in a row, (ii) no more than three large (> 5) or small (< 5) numbers appeared in a row, and (iii) the same number did not appear in consecutive trials. Each number was presented until the participants responded. The next trial started immediately afterward.

Procedure

The online experiment was implemented using the Qualtrics software (Qualtrics, Provo, UT). On the first page of the experiment, participants were instructed about the experiment, and their informed consent was obtained. Participants were also given the opportunity to look at the experiment without their data being analyzed. Finally, they had to indicate whether or not the device they were using to complete the experiment had arrow keys. On the second page, we asked several questions regarding participants' demographics and our predictors of interest regarding individual differences (see Table 1; see also the [supplementary material](#) for detailed information and the Qualtrics code).

³ In most SNARC studies the order of blocks is counterbalanced between participants; however, there have been no systematic studies on how block order affects the SNARC. Our preliminary analyses (Cipora et al., in prep) showed no order effect. In the case of the MARC effect, the evidence was less consistent. Here we decided to use a fixed block order, because we mainly aimed to investigate individual differences in the SNARC and the effects in small yet theoretically important subgroups (e.g., left-handers or individuals with dyscalculia). Keeping the block order fixed across participants had also been adopted in other studies investigating individual differences in the SNARC (e.g., Georges, Hoffmann, & Schiltz, 2018).

In the process of completing this questionnaire, participants either used their keyboard (e.g., to enter their age or nationality), selected one of the options via a drop-down list (e.g., the three options for gender), or moved a slider in the range of 0–10 (e.g., to rate their latest grades in mathematics). After completing the questionnaire, participants were instructed that in the following experiment, they would be asked to indicate whether the presented number was odd or even by pressing either the left or right arrow key.

After the experiment, the participants were asked about:

- their approximate distance to the computer screen, in centimeters,
- how noisy the environment was during the experiment (*silent, very quiet, fairly quiet, fairly noisy, very noisy, or extremely noisy*),
- the arrangement of the arrow keys on the keyboard,
- technical difficulties during the experiment,
- whether they would use the data, if they were the experimenter (*yes, not all of them, or no*), and
- whether they were interested in participating in a lottery for ten €20 Amazon vouchers (and if yes, to provide a contact email address).

Analysis

Data preparation—Formal and technical criteria Reaction time (RT) data were recorded for all participants; however, because of a technical problem in the online Qualtrics survey, response accuracy data were only recorded for 604 of the participants who met the inclusion criteria. There was no significant difference in any of the dependent variables between those participants with data for response accuracy available and those without. Furthermore, the response accuracy tends to be very high in the parity judgment task and usually is not analyzed (e.g., Cipora et al., 2016). Therefore, we decided to report the analyses based on the RTs of all 1,130 participants. To provide the reader with a complete overview, however, the results of the reduced sample for which response accuracy data were available is reported in part 1 of the [supplementary material](#). These results are almost identical to the results for the whole sample.

Data preparation—RT filtering In the first step of the RT trimming, anticipated responses (< 250 ms), which were 0.9% of the data, and RTs longer than 1,500 ms, which were 3.5% of the data, were excluded. Subsequently, a sequential filtering procedure was applied within each participant. RTs beyond ± 3 SDs from the individual's mean RT were sequentially removed, leading to the exclusion of 4% of the data (see Cipora & Nuerk, 2013, for the same procedure in in-lab SNARC experiments). Subsequently, the data from 24

Table 1. Demographic questions (see also Huber et al., 2017)

Predictor	Note
Age	
Gender	male, female, or transgender
Handedness	right, left, or ambidextrous
Education	school without qualification, school with qualification/GCSEs, high school with qualification/A levels, college qualification, bachelors degree, masters degree, or PhD
Field of study	art, science, humanities and social science, medicine/veterinary/dentistry, or other
Latest grade in math	a scale of 0–10 (from <i>very poor</i> to <i>very good</i>)
Latest grade in language	a scale of 0–10 (from <i>very poor</i> to <i>very good</i>)
Numeral system	Arabic, Eastern Arabic or Indian, Chinese, or other
Percentage of numeral system	depending on whether more than one numeral system was used, how often (as a percentage) a particular system was used (only in case a specific numeral system was used more than 5% of the time)
Starting hand for counting	right or left
Learning disabilities	dyslexia, dysgraphia, dyscalculia, ADHD, or other
Psychiatric illnesses	only in case a participant suffered from one
Nationality	
Native language	
Percentage of other languages	depending on whether a particular participant was multilingual, the amount of time (in percent) she/he spoke in each language in the last 3 years (only more than 5% of the time)
Sleep quality	a scale of 0–10 (from <i>very bad</i> to <i>very good</i>)
Alcohol consumption	never, less than once a month, once a month, two to three times a month, once a week, two to three times a week, or daily
Computer games	never, less than once a month, once a month, two to three times a month, once a week, two to three times a week, or daily
Device used	PC, laptop, netbook, tablet, smartphone, e-reader, or other

participants for whom more than 30% of the RT data were discarded during the filtering procedure were excluded from further analysis. Eventually, the data from 1,106 participants met all criteria. After the trimming procedure, 92.7% of their RT data were retained for analysis.

Among these participants, 1,056 reported not having learning disabilities. Five of the participants did not respond to the question, whereas 45 reported having ADHD or learning disabilities. Only participants who explicitly reported not having such difficulties were included in the analysis. Participants who reported learning disabilities or ADHD are compared to those not having learning disabilities in part 3 of the [supplementary material](#).

Data analysis The analysis was guided by the approach proposed by Lorch and Myers (1990) and was based on repeated measures regression calculated for each participant separately. This approach was adapted to SNARC research by Fias et al. (1996). It is based on the testing relationship between dRT (RT for right-hand responses – RT for left-hand responses) and numerical magnitude (1, 2, 3, 4, 6, 7, 8, and 9) within

each participant separately. Traditionally, a simple regression would be used, in which dRTs were regressed on numerical magnitude. The unstandardized regression slope would be considered to index the SNARC effect. It can be interpreted as a measure of the change in dRT with an increase of 1 in number magnitude. Negative slopes reflect the typical SNARC effect, and more negative slopes correspond to a stronger SNARC effect. To check whether the SNARC effect was present at the group level, averaged slopes were compared against zero in a one-sample *t* test. We used this method because it is the most popular index of the SNARC effect. Therefore, it makes our results comparable to those from most SNARC studies conducted to date. This measure is further referred to as *SNARC* throughout the article.

Nevertheless, this method of quantifying the SNARC effect has been subject to critique (Nuerk, Bauer, Krummenacher, Heller, & Willmes, 2005a; Pinhas, Tzelgov, & Ganor-Stern, 2012; Tzelgov, Zohar-Shai, & Nuerk, 2013). It has been argued that unstandardized slopes do not take into account the fit of the data points to the model. One possible

alternative is to use the within-participant correlation between dRT and number magnitude (see, e.g., Hoffmann, Mussolin, Martin, & Schiltz, 2014a). These correlation coefficients are then Fisher Z -transformed to approximate the normal distribution. This measure is further referred to as *ST-SNARC* throughout the article.

Furthermore, the classical SNARC estimation method does not take into account the MARC effect. Therefore, using a multiple regression has been recommended (Nuerk, Bauer, et al., 2005a; Nuerk et al., 2004). Namely, dRTs are regressed on the number magnitude and the number parity (contrast-coded, where odd numbers are coded as -0.5 and even numbers as 0.5). Therefore, negative slopes for the parity predictor correspond to a stronger MARC effect. Thus, two measures of the MARC effect complementary to those of the SNARC effect were used: the unstandardized slope (referred to as *MARC*) for a parity predictor, and the Fisher Z -transformed standardized slope (referred to as *ST-MARC*). Additionally, we report the overall fit of the multiple regression model ($MR-R^2$).

Note that the two predictors, number magnitude and number parity, are orthogonal (i.e., the correlation between them equals 0). Therefore, the slope estimates (both unstandardized and standardized) for the magnitude predictor will not differ, irrespective of whether single (i.e., magnitude predictor only) or multiple (i.e., magnitude and parity contrast) regression is used.

In our analyses, we used both frequentist and Bayesian approaches. In the case of Bayesian between-group comparisons, we utilized a method proposed by Masson (2011) in order to test the posterior probabilities, whether the result reflects the null-hypothesis model [evidence for no between-group difference/no effect; $p(H_0 | D)$] or the alternative-hypothesis model [complementary probability reflects a true between-group difference; $p(H_1 | D)$]. On the basis of recommendations by Raftery (1995), Masson proposed that a probability above .99 should be considered very strong evidence, .95–.99 as strong evidence, .75–.95 as positive evidence, and .50–.75 as weak evidence in favor of the given hypothesis.⁴

In the case of correlations, apart from exact estimates, along with each correlation, we provide 95% confidence intervals. Nevertheless, null correlations observed in a sample, even despite the large power, do not provide direct evidence of a null correlation in the population (Wagenmakers, Verhagen, & Ly, 2016). This limitation can be overcome within the Bayesian framework. Therefore, along with all correlations, we present corresponding posterior probabilities in favor of the alternative-hypothesis model given the data [i.e., the existing correlation between the variables: $p(H_1 | D)$]. Calculations were conducted using the JASP software, version 0.8.6 (JASP Team, 2018). The nondirectional hypothesis was assumed, and a beta prior was set to 1. Output values of the Bayes factor BF_{10} were transformed to posterior

probabilities using the formula, $p(H_1 | D) = BF_{10}/(BF_{10} + 1)$ (Masson, 2011).⁵ In general, for a $p(H_1 | D)$ larger than .15 and smaller than .85, the results need to be interpreted with caution, since robustness checks have indicated their large dependence on the beta prior value.

Note that the posterior probability favoring the null-hypothesis model [i.e., no between-group difference/no correlation; $p(H_0 | D)$] is complementary to the posterior probability of the alternative hypothesis [i.e., it is equal to $1 - p(H_1 | D)$]. Thus, we report $p(H_1 | D)$ values only.

In the main text, we will address research questions on reproducibility, comparability, reliability, and the validity of the SNARC and MARC. Relations between the SNARC and MARC effects and other variables are reported in the [supplementary material](#), part 2. Other analyses regarding between-country/-culture comparisons are reported in the [supplementary material](#), parts 4–6.

The online experiment as compared to typical in-lab studies on the SNARC effect

The methodology of online experiments in general, and the guidelines for conducting them, are outlined in several articles (Reips, 2002; Reips, Buchanan, Krantz, & McGraw, 2015). Here we discuss specific issues related to investigating the SNARC and MARC effects with a large sample online, in comparison with our experiences conducting similar in-lab experiments.

First, the experiment was conducted in 15 different languages at the same time. We asked colleagues from different countries to translate the materials to their mother tongue. This procedure itself took several months.

Distributing the link, apart from using formal means such as Web Experiment List (<http://wexlist.net>; Reips & Lengler, 2005) and Psychological Research on the Net (<https://psych.hanover.edu/research/exponnet.html>), required asking several colleagues for help and sharing the link among their own social networks (via university emails, student mailing lists, social media, sport clubs, etc.). Such steps should be considered especially if one wishes to approach heterogeneous samples—that is, not only individuals who regularly engage in online experiments, such as Mechanical Turk users. Note that using a selected sample such as professional Mechanical Turk users may lead to sample-specific effects. For instance, we have shown previously in our lab studies that professional mathematicians show a different SNARC effect

⁴ The effects were calculated using the Microsoft Excel sheet provided by Masson (2011) along with his article.

⁵ We used posterior probabilities instead of BF values because we believe that, as suggested by Masson (2011), these values are more intuitive for readers who are not familiar with Bayesian statistics. Readers who prefer the BF values can calculate them from $p(H_1 | D)$ using the same formula. BF_{10} thresholds of .33 and 3, typically considered as evidence for the null and alternative hypotheses, respectively, correspond to $p(H_1 | D)$ values of .25 and .75, respectively.

than more representative samples. Because, in our view, heterogeneous samples are essential both for producing representative online effects and for investigating individual differences, the time required for study preparation and data collection in a heterogeneous, more representative sample can be relatively long.

Accessing a large and heterogeneous sample allowed us to assess several covariates, which is typically difficult in in-lab settings, because of their much smaller participant numbers. On the other hand, we were restricted by the total time available for the experiment, and for this reason we had to limit the number of trials in the main task. This time limitation originated from two realities. First, Internet connections vary in quality, depending on world region. Furthermore, intranet connection speed generally exceeds the Internet speed by far. Thus, an experiment that takes about 15 min at a university might take two to three times longer in some regions. Luckily, if an online experiment is designed well (see, e.g., Garaizar & Reips, 2018), connection speed does not affect the precision of reaction time measurement, but it does influence loading times for pages. Second, participant motivation in attending long online studies may be limited, and we wanted to minimize the dropout (especially keeping in mind that our participants were not only professional Mechanical Turk users). According to our experience, dropout is not much of a problem in the case of in-lab studies, where extending the experiment length by 5–10 min has no obvious detrimental effects. For this reason, an online study needs to be carefully planned when it is adapted from an in-lab paradigm, and sometimes difficult decisions, such as limiting the number of trials, need to be made as a trade-off for the advantages of Internet-based experimenting (Reips, 2000, 2002; Reips & Krantz, 2010).

Our data collection depended on a third-party infrastructure (Qualtrics); therefore, it was not fully under our control. As it happened, an update in the platform caused the JavaScript code that collected the accuracy data to stop working properly halfway through data collection. As a consequence, we missed accuracy data for about half of the participants. Such incidents typically do not occur when conducting in-lab experiments with established routines, and are difficult to predict.

As compared to in-lab studies, data collected online often are more noisy and require more treatment and trimming before the analysis starts. The criteria based on which we excluded data sets are described in the Participants section, and our treatment of the RT data is presented the Analysis section. As compared to our in-lab experiences, we had to consider many more exclusion criteria at the participant level. We were also stricter in excluding outlier RTs than we usually do in case of in-lab experiments. Here, we first used fixed criteria (250 ms < RT < 1,500 ms) for inclusion. Subsequently, we applied the sequential trimming. In our in-lab studies, the sequential trimming procedure was sufficient, since—partly due to the lack of variance in the equipment—there were fewer

outlier RTs. Additionally, special care must be taken when excluding participant data. The sheer amount of data, due to the larger sample size alone, creates more work, but the large amount of data also allows for stricter inclusion criteria while providing strong power. In in-lab experiments, small samples, experimenter control, homogeneity of the equipment, and the link of student participants to the institution all keep noise at a low level.

Results

Reaction time characteristics

The mean RT was 673.0 ms ($SD = 104.1$). These values are longer than those typically reported in the literature (533 ms in Cipora et al., 2016; 528 ms in Cipora & Nuerk, 2013; between 535 and 619 ms in Hoffmann, Mussolin, et al., 2014a; between 534 and 626 ms, depending on the group, in Hoffmann, Pigat, & Schiltz, 2014b; between 650 and 880 ms in Ninaus et al., 2017; 487 ms in Nuerk, Wood, & Willmes, 2005b). The intraindividual variability in RTs was 130.9 ($SD = 46.1$). This was also larger than is typically reported (e.g., 97 ms in Cipora et al., 2016). Note that the latter measure is not routinely reported in most articles. Additional data for comparison can be found in Cipora et al (in prep). Longer RTs are not unusual when conducting Internet-based experiments (Reips, 2002; Reips & Krantz, 2010; Schmidt, 2007). For the most part, they can be explained by the transmission times between server and participant.

The SNARC effect and the MARC effect

As expected, the analysis revealed robust SNARC and MARC effects (see Table 2). The overall SNARC slope is shown in Fig. 1, and histograms presenting the distributions of all measures of interest are shown on Fig. 2.

The numerical value of the unstandardized slopes indicated a strong effect (see Fig. 3). A total of 877 participants (83% of the whole sample) revealed negative SNARC slopes. This is in line with in-lab studies (88.3% in Fattorini, Pinto,

Table 2. Overall SNARC and MARC effects

Measure	Mean (<i>SD</i>)	<i>t</i> *
SNARC	– 8.49 (10.28)	– 26.84
ST-SNARC	– 0.38 (0.45)	– 27.24
MARC	– 36.19 (136.48)	– 8.62
ST-MARC	– 0.22 (0.83)	– 8.64
MR- <i>R</i> ²	.56 (.26)	69.19

* All *t* tests were against zero; *df* = 1,055; all *ps* < .001

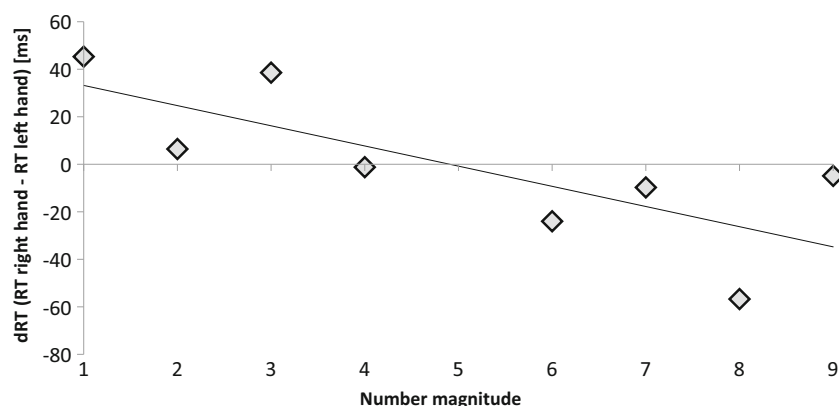


Fig. 1 Overall SNARC and MARC effects. It can be seen that larger numbers are responded to faster with the right than with the left hand (as also depicted by the regression line), which indicates the SNARC effect. Furthermore, the data points for even numbers are below the

regression line, and the data points for odd numbers are above it (i.e., right-/left-hand responses to even/odd numbers are faster), thus indicating the MARC effect

Rotondaro, & Doricchi, 2015; 70.8% and 73.9% in Exps. 1 and 4, respectively, in Fias, Lauwereyns, & Lammertyn, 2001; 87% in Schwarz & Müller, 2006, Exp. 2; 80% in Viarouge et al., 2014). The estimate of the MARC effect was also relatively high (see Fig. 4). A total of 644 participants (61% of the whole sample) showed negative MARC slopes. The data indicate that both the SNARC and MARC effects can be effectively measured in an online setup and provide a basis for our subsequent investigation of determinants of the effect.

Comparison to other studies—Effect sizes

In this step, the SNARC effect of the present online study was compared to those reported in the literature (see Fig. 3 for the SNARC measures and Fig. 4 for the MARC measures). Comparison data come from a uniform analysis of several SNARC experiments by Cipora et al. (in prep).⁶ This analysis allowed us to examine all respective numerical values and to estimate confidence intervals.

Reliability of the effects

Split-half reliability of the effects was calculated. Calculation was based on an odd-even method depending on the order of presentation. For each half, the mean RT and intraindividual variability in RTs were calculated. Then the scores were correlated. The Pearson correlations were Spearman–Brown corrected to adjust for test length (see Cipora & Nuerk, 2013). The reliability of the mean RT was .99, and that for intraindividual variability in RTs was .97. These estimates are similar to those reported in lab studies. For in-lab studies, the reliability of the mean RT was .99 (Cipora et al., 2016; Cipora & Nuerk, 2013), and the intraindividual variability was .97 and .99, respectively.

⁶ Current status of this ongoing project can be viewed on <https://osf.io/n7szg/>.

Subsequently, for each half, the within-participant multiple regressions were calculated (as in the primary analysis). The respective measures from each half were correlated, and the coefficients were Spearman–Brown corrected. In the case of the SNARC, the reliability was .43, and for the ST-SNARC it was .45. These estimates were relatively low with regard to typical psychometric criteria. In the case of in-lab studies, reliabilities have sometimes been higher (.82, .70, and .75 in the cases of Cipora et al., 2016; Cipora & Nuerk, 2013; and Fattorini et al., 2015, Exp. 1, respectively, for unstandardized slopes). However, in some in-lab studies reliabilities have been comparable to those from the online study (e.g., .55 in Georges et al., 2017).

The reliability of standardized slopes, calculated according to the method used here, is reported only in Cipora et al. (2016), and it was .75. The reliability of standardized slopes, calculated in a slightly different way, as reported by Georges et al. (2017) equaled .78.

In the case of the MARC, the reliability was .91, and for ST-MARC it was .83. We are not aware of in-lab studies in which MARC reliabilities have been reported. So this is the first such report, and as of yet it cannot be compared to those from lab studies. The reliability of $MR-R^2$ was .43.

Validity of the effects

As expected, standardized estimates of the given effects correlated highly with the unstandardized ones (.833 and .873 for SNARC and MARC, respectively). Interestingly, the SNARC and MARC effects did not correlate with each other ($r = -.014$, $p = .660$, 95% CI $[-.07, .05]$, Bayesian $p(H1 | D) = .040$); however, the correlation between ST-SNARC and ST-MARC reached significance ($r = -.104$, $p = .001$, 95% CI $[-.16, -.04]$, Bayesian $p(H1 | D) = .924$), but was very small (i.e., it corresponded to $\sim 1.1\%$ of the variance in common). SNARC and MARC estimates (both unstandardized and

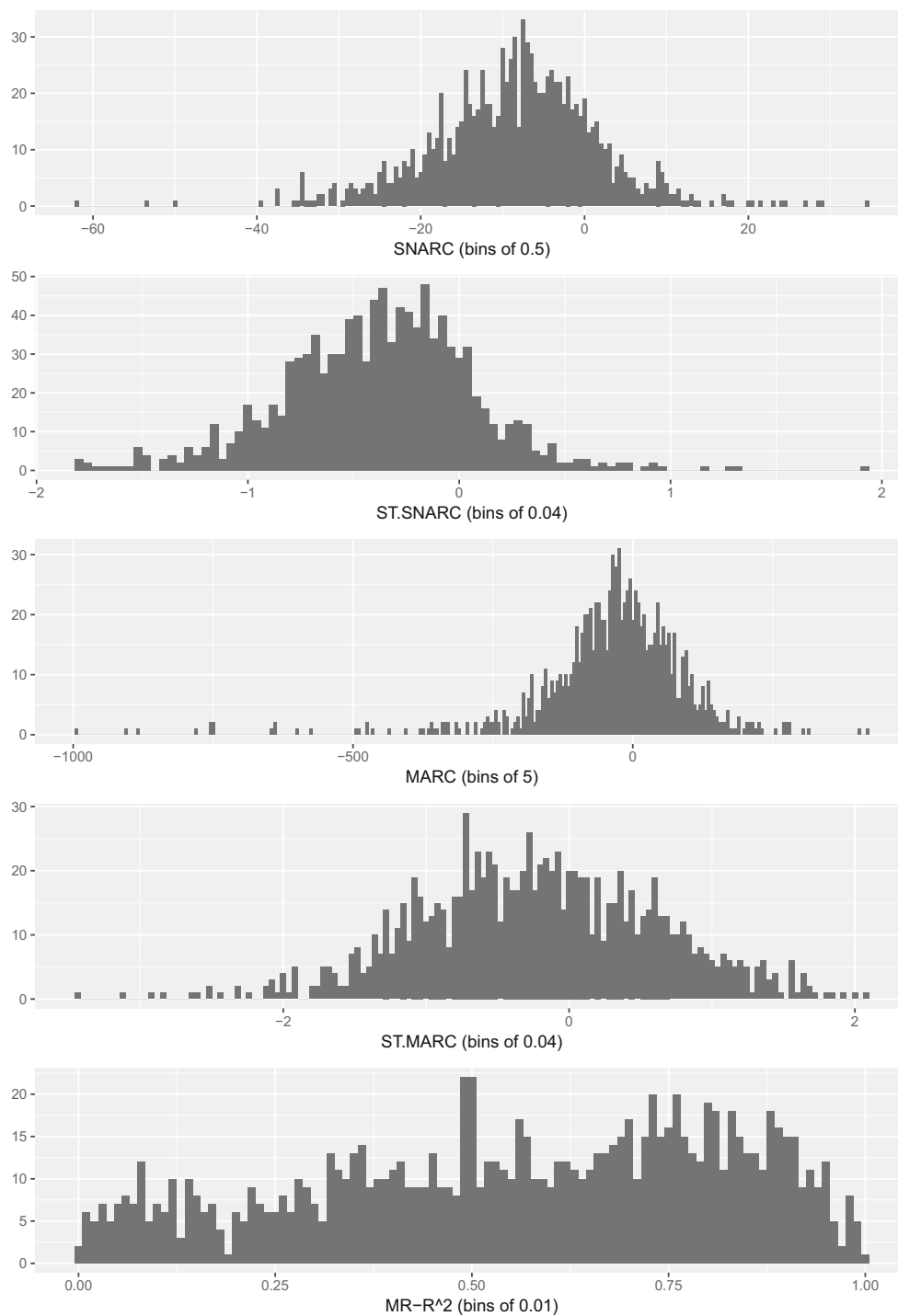


Fig. 2 Distributions of SNARC, ST-SNARC, MARC, and ST-MARC slopes and of $MR-R^2$ values

standardized) correlated negatively with $MR-R^2$, $-.283 < r < -.205$, $p < .001$. Because stronger SNARC and MARC effects correspond to more negative values and better regression fit, and $MR-R^2$ corresponds to higher positive values, this negative correlation means that participants who revealed stronger SNARC and MARC effects were also characterized by a better overall fit of the multiple regression models.

In line with earlier results (e.g., Cipora et al., 2016), as well as with the results of a meta-analysis (Wood et al., 2008), we found that the SNARC effect correlated with mean RT in the parity judgment task (see Table 3). The same was true for the correlation between the MARC effect and mean RT. However, this relationship was not observed for the ST-SNARC (evidence for a null correlation was also supported by the Bayesian analysis; see Table 3). In the case of ST-

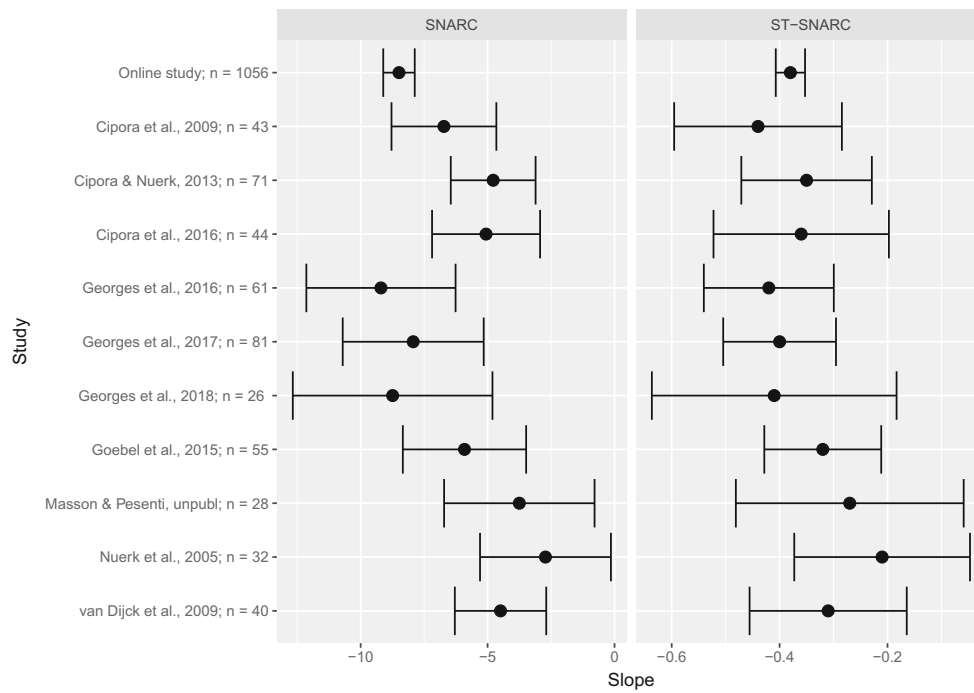


Fig. 3 Comparison of SNARC slopes (unstandardized and standardized) between the present online study and published in-lab studies. Bars represent 95% confidence intervals

MARC, despite the significant correlation with mean RT, the results of the Bayesian analysis were largely inconclusive (i.e., there was no strong evidence for either the alternative or the null hypothesis; see Table 3). On the other hand, the Bayesian analysis provided strong evidence supporting a

relationship between ST-MARC and the intraindividual variability in RTs.

We also directly compared the relationship between the dependent measures and RT characteristics in our present online study to those reported in the literature.

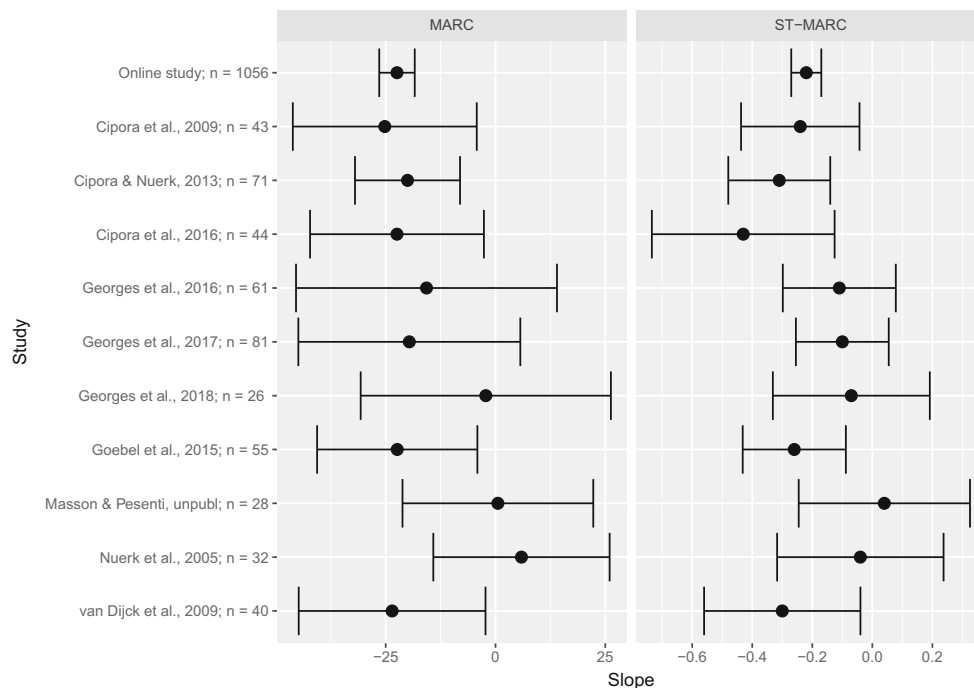


Fig. 4 Comparison of MARC slopes (unstandardized and standardized) between the present online study and published in-lab studies. Bars represent 95% confidence intervals

Table 3 Correlations between the dependent variables

RT Characteristic	Measure	SNARC	ST-SNARC	MARC	ST-MARC	MR- R^2
Mean	<i>r</i>	-.167**	.040	-.169**	-.064*	.008
	95% CI	-.225; -.108	-.020; .100	-.227; -.109	-.124; -.003	-.052; .068
	<i>p</i> (H1 D)	> .999	.084	> .999	.249	.038
SD	<i>r</i>	-.153**	.130**	-.281**	-.125**	.029
	95% CI	-.211; -.093	.071; .189	-.335; -.224	-.184; -.065	-.032; .089
	<i>p</i> (H1 D)	> .999	.997	> .999	.994	.056

N = 1,056. * *p* < .05 (two-sided); ** *p* < .001 (two-sided)

Note that this correlation is not reported in every article; therefore, we provide the estimates that could be found in the existing literature (see Fig. 5). In most cases, the confidence intervals largely overlap.

In sum, the standard SNARC and MARC slopes were observed in our large and therefore strongly powered online sample, as well as the standard correlations between SNARC slopes and both overall RT and variability, which has often previously been reported in smaller studies.

Discussion

In a web experiment, we aimed to reproduce two well-known phenomena in numerical cognition—the SNARC and MARC effects. Regarding our four initial hypotheses, we can say that:

1. SNARC and MARC effects are reproducible in an online study.
2. SNARC and MARC effects were largely comparable to those observed in lab studies.

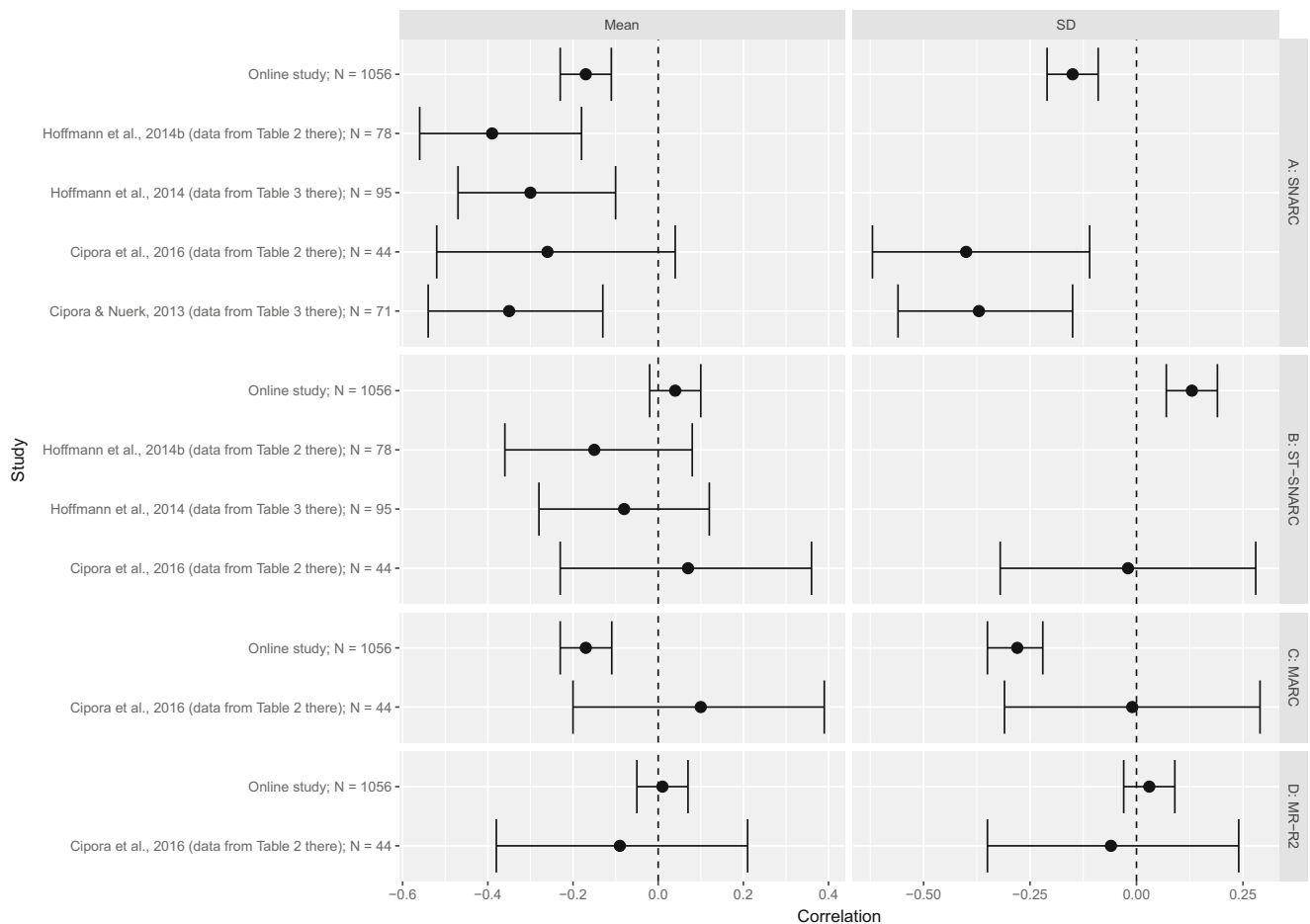


Fig. 5 Relationship between the overall reaction time characteristics (mean RT, intraindividual variability in RT) and our measures of interest across studies. Bars represent 95% confidence intervals

3. The split-half reliability for the SNARC was relatively poor, which is in line with some studies, but not with others—this may depend on the study and sample characteristics. MARC reliabilities, however, were quite high.
4. The SNARC effects were valid with respect to correlations with the mean RT and the *SD* of RTs. For the MARC effects, the picture was more mixed.

In sum, both the SNARC and MARC effects were reproduced, and several of their properties were observed, which suggests that these effects can be investigated by means of online experimentation. We now elaborate on these results in more detail.

Reaction time characteristics

The mean RTs in our online study were longer than those typically reported in bimanual parity judgments conducted with students. However, our results are similar to in-lab studies testing of heterogeneous samples. In the latter, mean RTs can be even longer than those in our online experiment (e.g., Ninaus et al., 2017). Sample heterogeneity could lead to longer RTs, even though the mechanism is unclear. On the other hand, in a similar online implementation of two-digit number comparison, the mean RTs were similar to those observed in in-lab studies testing mostly student samples (Huber et al., 2017).

In a similar vein, the intraindividual variability in RTs was also relatively large. As was reported by Cipora and Nuerk (2013), mean RTs correlate very highly (more than .80) with intraindividual variability in RTs. Importantly, intraindividual variability can arise from the sample characteristics. Participants who are not familiar with the types of tasks used in cognition experiments might respond more slowly in general, and the variance within their RTs can also be larger.

On the other hand, the relatively long RTs in Internet-based experiments can be attributed to several technical issues, such as transmission times between server and participants' computers (Reips, 2002; Reips & Krantz, 2010; Schmidt, 2007), loading times of web pages, or the software used to deliver the experimental materials and record the responses. The first issue can be accounted for by measuring RTs both server-side and client-side (Reips & Krantz, 2010), as for example is the default in the web experiment software WEXTOR (<https://wextor.eu>).

In sum, both technical issues and sample composition might have contributed to the differences in the means and *SDs* of RTs between our online and previous in-lab studies.

The SNARC and MARC effects

We reproduced SNARC and MARC effects online. The SNARC slopes we observed can be considered relatively

strong. Nevertheless, they were within the range of slopes reported in laboratory experiments. Confidence intervals largely overlap between the online experiment and in-lab studies. The same holds true for both unstandardized and standardized slopes. Note that a very recent study (Gökaydin, Brugger, & Loetscher, 2018), also replicated the SNARC in an online setup in a group of professional Mechanical Turk users (only participants, who completed more than 500 tasks before were considered). However, in this special and homogeneous group the observed SNARC effect was rather weak. Authors did not aim at providing more thorough comparisons to in-lab experiments. This finding together with findings reported here further support the claim that special groups may differ in their SNARC effect from the general population (see Cipora et al., 2016, for similar findings in in-lab setup). Furthermore, the proportion of participants revealing negative slopes is similar to proportions observed in laboratory setups. Prevalence of the SNARC effect can be further investigated by means of psychometric and bootstrapping approaches (see Cipora, van Dijck, et al., 2018b; Cipora et al., in prep). Another interesting way of investigating it is specific Bayesian models within System of Orders framework (Haaf, Klaassen, & Rouder, 2018).⁷

The group-level MARC effect (both unstandardized and standardized) was also robust. The size of the effect was also comparable to those reported in lab-based experiments, and respective confidence intervals overlapped considerably.

Reliability

The reliability of the SNARC effect was relatively low. Nevertheless, at least to some extent this was foreseeable. First of all, cognitive phenomena, including the SNARC effect, are sometimes characterized by relatively low reliability (Hedge et al., 2018). This seems to hold for other phenomena in numerical cognition as well (see Maloney, Risko, Preston, Ansari, & Fugelsang, 2010). In some studies, reliabilities of the SNARC were higher (Cipora et al., 2016; Cipora & Nuerk, 2013; Fattorini et al., 2015). However, in other studies reliabilities were comparable to those obtained in the presented study (see Cipora, van Dijck, et al., 2018, for a comparison). Recently, Cipora and Wood (2017) showed that the power to detect the SNARC as well as to find between group differences in the effect largely depends on the number of repetitions of each number in a block. Increasing the number of trials leads to higher power. Furthermore, increasing the number of repetitions increases the reliability of the effect, by reducing random error. Cipora and Wood concluded that to accurately and reliably measure the SNARC, one needs to use

⁷ Note that the data we share along with this article can be an interesting resource for testing such novel methods.

at least 20 repetitions of each number in each block. Here we used only ten repetitions. This was ultimately sufficient to observe a robust SNARC, but may have not been sufficient to ensure satisfactory reliability. Future online studies may consider using more repetitions as Cipora and Wood have recommended for in-lab studies.

Validity estimates: Correlations between measures

Replicating other findings (e.g., Gevers, Verguts, Reynvoet, Caessens, & Fias, 2006; Hoffmann, Pigat, & Schiltz, 2014b), we observed correlations between the SNARC and RTs in the parity judgment task. Participants who responded slower revealed a stronger effect. However, this effect was apparent only for the unstandardized slopes, but not for standardized ones. This is in line with previous findings from in-lab experiments (e.g., Cipora & Nuerk, 2013). Unstandardized SNARC correlated even more strongly with intraindividual variance in RTs, which also confirms the previous in-lab observations: For instance, Cipora and Nuerk did not observe such a relationship when they used *z*-standardized RTs instead of raw ones (see also Cipora et al., 2016).

In the case of the MARC effect, similar relationships were observed. Longer and more varied RTs were associated with more pronounced unstandardized MARC. However, no such effect was observed for standardized MARC. These effects have not so far been reported widely in the literature, and therefore no sufficient comparison data are available.

Interestingly, the unstandardized SNARC and MARC effects did not correlate with each other, and the correlation between the standardized slopes was very low, accounting for about 1% of the variance in common.⁸ This relationship has been reported in very few studies; however, existing reports show that these two effects do not correlate in the lab, either (Huber et al., 2015; Krajcsi, Lengyel, & Laczko, 2018).

To sum up, the results of our online study regarding correlations between the performance measures in a parity judgment task and SNARC and MARC properties seem to accurately reflect the results of in-lab experiments.

Limitations and future steps

The obvious drawback of the present study is the technical problem in Qualtrics that caused the loss of accuracy data. However, additional analyses showed that even if only correct responses were considered and some participants were excluded on the basis of an accuracy criterion, the results would not change considerably (see the [supplementary material](#)).

⁸ This correlation should be treated with extreme caution, because ST-SNARC and ST-MARC correspond to standardized slopes obtained in the same multiple regression. If two predictors are perfectly orthogonal, the sum of the squared standardized regression slopes corresponds to the proportion of explained variance (which cannot exceed 1).

Furthermore, in the context of differential studies, the reliability of the SNARC effect can be considered insufficient. Nevertheless, as has been shown, it increases with an increasing number of repetitions. Future studies should thus test setups with more repetitions. Extending the number of trials might introduce significant difficulties in online setups, however, in that experiment duration plays a very important role with regard to the number of participants, dropout, and overall quality of the data (Reips, 2002). This challenge needs to be addressed in future studies because, apart from ensuring sufficient reliability of the SNARC effect measurement, researchers need to reliably measure potential covariates.

For the findings concerning other measures—the MARC effect and $MR-R^2$ —the comparisons to other studies were not as straightforward as they were for the SNARC effect. This is because the MARC effect has not been investigated thoroughly, so far. This is clearly a gap in the literature that needs to be addressed in order to get a full picture.

$MR-R^2$ is a less popular measure; therefore, its usefulness still needs to be verified, and sufficient data need to be collected in-lab (or reanalyzed) to allow comparisons with the effects reported here.

Conclusions

Our study shows that the SNARC and MARC effects can be effectively measured by means of online experiments. Despite being small effects, on the scale of a few milliseconds, they are reproducible and comparable to those from lab studies, even in an increasingly complex browser environment (Garaizar & Reips, 2018). With ten repetitions, the SNARC effect was not highly reliable, whereas the MARC effect was. The validity of SNARC effects (i.e., their correlations with the mean and *SD* of RTs) was similar to validities found in lab studies. Therefore, we suggest that online studies provide a powerful method to conduct large-scale investigations of the SNARC effect and its modulation by interindividual differences or other variables of interest.

To conclude, our findings confirm that researchers can successfully use online experiments for investigating phenomena in cognition, especially if they wish to take a differential approach requiring testing of large samples. Nevertheless, efforts are needed to ensure sufficient measurement reliability, and this issue needs to be carefully addressed.

Author note We thank all participants. We thank our colleagues and friends who translated the experiment into different languages and helped access people. Finally, we thank Ida von Lehsten, for her help in preparing the tables in the [supplementary material](#); Urszula Mihułowicz, for her help in preparing part 3 of the supplementary material; and Julianne Skinner, for language proofreading of the article. This research was funded by the Science Campus Tuebingen, project

grant 8.4 to H.-C.N. supporting M.S. The authors K.C. and M.S. are supported by a grant from the Deutsche Forschungsgemeinschaft (NU 265/3-1) to H.-C.N. Authors K.C., M.S., and H.-C.N. are members of LEAD Graduate School & Research Network (GSC1028), which is funded within the framework of the Excellence Initiative of the German federal and state governments. The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527–535. <https://doi.org/10.3758/s13428-012-0265-2>
- Bayram, A. B. (2018). Serious subjects: A test of the seriousness technique to increase participant motivation in political science experiments. *Research & Politics*, *5*(2), 2053168018767453. <https://doi.org/10.1177/2053168018767453>
- Berch, D. B., Foley, E. J., Hill, R. J., & Ryan, P. M. D. (1999). Extracting parity and magnitude from arabic numerals: Developmental changes in number processing and mental representation. *Journal of Experimental Child Psychology*, *74*, 286–308. <https://doi.org/10.1006/jecp.1999.2518>
- Bull, R., Cleland, A. A., & Mitchell, T. (2013). Sex differences in the spatial representation of number. *Journal of Experimental Psychology: General*, *142*, 181–192. <https://doi.org/10.1037/a0028387>
- Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nečka, E. (2016). Professional mathematicians differ from controls in their spatial–numerical associations. *Psychological Research*, *80*, 710–726. <https://doi.org/10.1007/s00426-015-0677-6>
- Cipora, K., & Nuerk, H.-C. (2013). Is the SNARC effect related to the level of mathematics? No systematic relationship observed despite more power, more repetitions, and more direct assessment of arithmetic skill. *Quarterly Journal of Experimental Psychology*, *66*, 1974–1991. <https://doi.org/10.1080/17470218.2013.772215>
- Cipora, K., Patro, K., & Nuerk, H.-C. (2015). Are spatial–numerical associations a cornerstone for arithmetic learning? The lack of genuine correlations suggests no. *Mind, Brain, and Education*, *9*, 190–206. <https://doi.org/10.1111/mbe.12093>
- Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–49). Cambridge: Cambridge University Press.
- Cipora, K., Schroeder, P. A., & Nuerk, H.-C. (2019a). On the multitude of mathematics skills: Spatial–numerical associations and geometry skill? In K. S. Mix & M. Battista (Eds.), *Visualizing mathematics: The role of spatial reasoning in mathematical thought* (pp. 361–370). Cham: Springer.
- Cipora, K., Schroeder, P. A., Soltanlou, M., & Nuerk, H.-C. (2019b). More space, better mathematics: Is space a powerful tool or a cornerstone for understanding arithmetic? In K. S. Mix & M. T. Battista (Eds.), *Visualizing mathematics: The role of spatial reasoning in mathematical thought* (pp. 77–116). Cham: Springer.
- Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20*20 Monte Carlo investigation. *Frontiers in Psychology*, *8*, 1194. <https://doi.org/10.3389/fpsyg.2017.01194>
- Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*, 1482. <https://doi.org/10.3389/fpsyg.2017.01482>
- Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*, e57410. <https://doi.org/10.1371/journal.pone.0057410>
- Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*, 371–396. <https://doi.org/10.1037/0096-3445.122.3.371>
- 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*, 95–110. <https://doi.org/10.1080/135467996387552>
- Fias, W., Lauwereyns, J., & Lammertyn, J. (2001). Irrelevant digits affect feature-based attention depending on the overlap of neural circuits. *Cognitive Brain Research*, *12*, 415–423. [https://doi.org/10.1016/S0926-6410\(01\)00078-7](https://doi.org/10.1016/S0926-6410(01)00078-7)
- Fischer, M. H. (2003a). Cognitive representation of negative numbers. *Psychological Science*, *14*, 278–282. <https://doi.org/10.1111/1467-9280.03435>
- Fischer, M. H. (2003b). Spatial representations in number processing—Evidence from a pointing task. *Visual Cognition*, *10*, 493–508. <https://doi.org/10.1080/13506280244000186>
- Fischer, M. H. (2008). Finger counting habits modulate spatial–numerical associations. *Cortex*, *44*, 386–392. <https://doi.org/10.1016/j.cortex.2007.08.004>
- Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition—From single digits to arithmetic. *Quarterly Journal of Experimental Psychology*, *67*, 1461–1483. <https://doi.org/10.1080/17470218.2014.927515>
- Fischer, M. H., Warlop, N., Hill, R. L., & Fias, W. (2004). Oculomotor bias induced by number perception. *Experimental Psychology*, *51*, 91–97. <https://doi.org/10.1027/1618-3169.51.2.91>
- Garaizar, P., & Reips, U.-D. (2018). Best practices: Two Web-browser-based methods for stimulus presentation in behavioral experiments with high-resolution timing requirements. *Behavior Research Methods*, 1–13. Advance online publication. <https://doi.org/10.3758/s13428-018-1126-4>
- Georges, C., Hoffmann, D., & Schiltz, C. (2016). How math anxiety relates to number-space associations. *Frontiers in Psychology*, *7*, 1401:1–15. <https://doi.org/10.3389/fpsyg.2016.01401>
- 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*, 182–211.
- Georges, C., Hoffmann, D., & Schiltz, C. (2018). Implicit and explicit number-space associations differentially relate to interference control in young adults with ADHD. *Frontiers in Psychology*, *9*, 775. <https://doi.org/10.3389/fpsyg.2018.00775>
- 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*, 32–44. <https://doi.org/10.1037/0096-1523.32.1.32>
- Göbel, S. M., McCrink, K., Fischer, M. H., & Shaki, S. (2018). Observation of directional storybook reading influences young children's counting direction. *Journal of Experimental Child Psychology*, *166*, 49–66. <https://doi.org/10.1016/j.jecp.2017.08.001>

- Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential effects in SNARC. *Scientific Reports*, 8, 10996:1–13. <https://doi.org/10.1038/s41598-018-29337-2>
- Haaf, J. M., Klaassen, F., & Rouder, J. N. (2018). A note on using systems of orders to capture theoretical constraint in psychological science. [10.31234/osf.io/a4xu9](https://doi.org/10.31234/osf.io/a4xu9)
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50, 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>
- Hoffmann, D., Mussolin, C., Martin, R., & Schiltz, C. (2014a). The impact of mathematical proficiency on the number-space association. *PLoS ONE*, 9, 85048. <https://doi.org/10.1371/journal.pone.0085048>
- Hoffmann, D., Pigat, D., & Schiltz, C. (2014b). The impact of inhibition capacities and age on number-space associations. *Cognitive Processing*, 15, 329–342. <https://doi.org/10.1007/s10339-014-0601-9>
- Huber, S., Klein, E., Graf, M., Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Embodied markedness of parity? Examining handedness effects on parity judgments. *Psychological Research*, 79, 963–977. <https://doi.org/10.1007/s00426-014-0626-9>
- Huber, S., Nuerk, H.-C., Reips, U.-D., & Soltanlou, M. (2017). Individual differences influence two-digit number processing, but not their analog magnitude processing: A large-scale online study. *Psychological Research*, 1–21. Advance online publication <https://doi.org/10.1007/s00426-017-0964-5>
- JASP Team. (2018). Jasp (Version 0.8.6) [Computer software]. Retrieved from <https://jasp-stats.org/download/>
- Krajcsi, A., Lengyel, G., & Laczko, Á. (2018). Interference between number magnitude and parity. *Experimental Psychology*, 65, 71–83. <https://doi.org/10.1027/1618-3169/a000394>
- Krantz, J. H., & Reips, U.-D. (2017). The state of web-based research: A survey and call for inclusion in curricula. *Behavior Research Methods*, 49, 1621–1629. <https://doi.org/10.3758/s13428-017-0882-x>
- 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, 149–157. <https://doi.org/10.1037/0278-7393.16.1.149>
- Lyons, J. (1968). *Semantics*. Cambridge: Cambridge University Press.
- Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica*, 134, 154–161. <https://doi.org/10.1016/j.actpsy.2010.01.006>
- Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43, 679–690. <https://doi.org/10.3758/s13428-010-0049-5>
- Nemeh, F., Humberstone, J., Yates, M. J., & Reeve, R. A. (2018). Non-symbolic magnitudes are represented spatially: Evidence from a non-symbolic SNARC task. *PLoS ONE*, 13, e0203019. <https://doi.org/10.1371/journal.pone.0203019>
- Ninaus, M., Moeller, K., Kaufmann, L., Fischer, M. H., Nuerk, H.-C., & Wood, G. (2017). Cognitive mechanisms underlying directional and non-directional spatial-numerical associations across the lifespan. *Frontiers in Psychology*, 8, 1421. <https://doi.org/10.3389/fpsyg.2017.01421>
- Nuerk, H.-C., Bauer, F., Krummenacher, J., Heller, D., & Willmes, K. (2005a). The power of the mental number line: how the magnitude of unattended numbers affects performance in an Eriksen task. *Psychology*, 47, 34–50. Retrieved from http://www.pabst-publishers.de/psychology-science/1-2005/ps_1_2005_34-50.pdf
- Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *Quarterly Journal of Experimental Psychology*, 57A, 835–863. <https://doi.org/10.1080/02724980343000512>
- Nuerk, H.-C., Patro, K., Cress, U., Schild, U., Friedrich, C. K., & Göbel, S. M. (2015). How space-number associations may be created in preliterate children: Six distinct mechanisms. *Frontiers in Psychology*, 6, 1–6. <https://doi.org/10.3389/fpsyg.2015.00215>
- Nuerk, H.-C., Wood, G., & Willmes, K. (2005b). The universal SNARC effect: The association between number magnitude and space is amodal. *Experimental Psychology*, 52, 187–194. <https://doi.org/10.1027/1618-3169.52.3.187>
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <https://doi.org/10.1126/science.aac4716>
- Patro, K., Fischer, U., Nuerk, H.-C., & Cress, U. (2016a). How to rapidly construct a spatial-numerical representation in preliterate children (at least temporarily). *Developmental Science*, 19, 126–144. <https://doi.org/10.1111/desc.12296>
- Patro, K., & Haman, M. (2012). The spatial-numerical congruity effect in preschoolers. *Journal of Experimental Child Psychology*, 111, 534–542. <https://doi.org/10.1016/j.jecp.2011.09.006>
- Patro, K., & Nuerk, H.-C. (2017). Limitations of trans-species inferences: The case of spatial-numerical associations in chicks and humans. *Cognitive Science*, 41, 2267–2274. <https://doi.org/10.1111/cogs.12432>
- Patro, K., Nuerk, H.-C., & Cress, U. (2016b). Mental number line in the preliterate brain: The role of early directional experiences. *Child Development Perspectives*, 10, 172–177. <https://doi.org/10.1111/cdep.12179>
- Pinhas, M., Tzelgov, J., & Ganor-Stern, D. (2012). Estimating linear effects in ANOVA designs: The easy way. *Behavior Research Methods*, 44, 788–794. <https://doi.org/10.3758/s13428-011-0172-y>
- Priftis, K., Zorzi, M., Meneghello, F., Marenzi, R., & Umiltà, C. (2006). Explicit versus implicit processing of representational space in neglect: Dissociations in accessing the mental number line. *Journal of Cognitive Neuroscience*, 18, 680–688. <https://doi.org/10.1162/jocn.2006.18.4.680>
- Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, 132, 416–442. <https://doi.org/10.1037/0033-2909.132.3.416>
- Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Madsen (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwell.
- Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–118). San Diego: Academic Press. <https://doi.org/10.5167/uzh-19760>
- Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49, 243–256. <https://doi.org/10.1026/1618-3169.49.4.243>
- Reips, U.-D. (2012). Using the Internet to collect data. In H. Cooper, P. M. Camic, R. Gonzalez, D. L. Long, A. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 291–310). Washington, DC: American Psychological Association. <https://doi.org/10.1037/13620-017>
- Reips, U.-D., Buchanan, T., Krantz, J., & McGraw, K. (2015). Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica*, 15, 139–148.
- Reips, U.-D., & Krantz, J. H. (2010). Conducting true experiments on the Web. In S. Gosling & J. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 139–216). Washington DC: American Psychological Association.
- Reips, U.-D., & Lengler, R. (2005). The Web Experiment List: A web service for the recruitment of participants and archiving of Internet-based experiments. *Behavior Research Methods*, 37, 287–292. <https://doi.org/10.3758/BF03192696>

- Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, *347*, 534–536. <https://doi.org/10.1126/science.aaa1379>
- Schmidt, W. C. (2007). Technical considerations when implementing online research. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 461–472). Oxford: Oxford University Press.
- 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*. <https://doi.org/10.3389/fnins.2017.00654>
- Schwarz, W., & Müller, D. (2006). Spatial associations in number-related tasks: A comparison of manual and pedal responses. *Experimental Psychology*, *53*, 4–15. <https://doi.org/10.1027/1618-3169.53.1.4>
- Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*, 328–331. <https://doi.org/10.3758/PBR.16.2.328>
- Tlauka, M. (2002). The processing of numbers in choice-reaction tasks. *Australian Journal of Psychology*, *54*, 94–98. <https://doi.org/10.1080/00049530210001706553>
- Tzelgov, J., Zohar-Shai, B., & Nuerk, H.-C. (2013). On defining quantifying and measuring the SNARC effect. *Frontiers in Psychology*, *4*, 302:3–5. <https://doi.org/10.3389/fpsyg.2013.00302>
- Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLoS ONE*, *9*, e95756. <https://doi.org/10.1371/journal.pone.0095756>
- Wagenmakers, E. J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426. <https://doi.org/10.3758/s13428-015-0593-0>
- Wood, G., Nuerk, H.-C., & Willmes, K. (2006). Variability of the SNARC effect: Systematic interindividual differences or just random error? *Cortex*, *42*, 1119–1123. [https://doi.org/10.1016/S0010-9452\(08\)70223-5](https://doi.org/10.1016/S0010-9452(08)70223-5)
- 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 Quarterly*, *4*, 489–525. <https://doi.org/10.1027/1618-3169.52.3.187>
- Zohar-Shai, B., Tzelgov, J., Karni, A., & Rubinsten, O. (2017). It does exist! A left-to-right spatial–numerical association of response codes (SNARC) effect among native Hebrew speakers. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 719–728. <https://doi.org/10.1037/xhp0000336>