A Minority pulls the sample mean: on the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect

Krzysztof Cipora[1,2]*, Jean-Philippe van Dijck[3,4], Carrie Georges[5], Nicolas Masson[6], Silke M. Göbel[7], Klaus Willmes[8], Mauro Pesenti[6], Christine Schiltz[5], & Hans-Christoph Nuerk[1,2,9]

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

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

[3] Department of Experimental Psychology, Ghent University, Belgium

[4] Department of Applied Psychology: Thomas More University College, Belgium

[5] Institute of Cognitive Science and Assessment, ECCS Research Unit, University of Luxembourg, Luxembourg

[6] Institut de Recherche en Sciences Psychologiques and Institute of Neuroscience, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[7] Department of Psychology, University of York, York, UK

[8] Department of Neurology, University Clinic, RWTH Aachen University, Aachen Germany

[9] Leibnitz-Institut für Wissenmedien, Tuebingen Germany

* Corresponding author: Department of Psychology, University of Tuebingen, Schleichstrasse 4, 72076 Tuebingen, Germany. E-mail: krzysztof.cipora@uni-tuebingen.de

ABSTRACT

The aim of cognitive psychology is to obtain insights into human cognition in general. For this purpose, group-studies are typically conducted on representative samples so that the results can be generalized to the population. Using this approach, individual differences in such group-level cognitive phenomena are typically neglected and not much is known about their prevalence at the individual level. Such information is nevertheless important for claims about the universality of phenomena, as in theory, significant effects at the group-level can in principle be driven by a minority of participants. Here we used a uniform analysis of 18 existing data sets revealing a well-replicated phenomenon in numerical cognition: the SNARC (Spatial-Numerical Association of Response Codes) effect, in order to investigate the prevalence of the effect at the individual level. Three methods of analyzing the presence of the effect at the individual level were utilized: one psychometric and two bootstrapping methods. The results show that the group-level SNARC effect is driven by a minority of individuals (≤ 45%) who reveal the effect. This finding demonstrate an important theoretical issue: whether group-level effects really reflect general principles of cognition. We discuss advantages and drawbacks of the present methods and their usefulness for investigating the prevalence of other cognitive phenomena. We posit that testing the presence of robust group-level cognitive effects at the individual level as well as ensuring their reliable measurement is an important step towards integrating two traditionally separate approaches of scientific psychology proposed back in 50' by Cronbach: experimental and correlational.

250 words (Limit 250)

Keywords: individual differences in cognitive phenomena; meta-analytic techniques; confidence intervals for cognitive phenomena; prevalence of cognitive phenomena; bootstrap confidence intervals for cognitive phenomena; reliability of cognitive phenomena; SNARC effect

INTRODUCTION

Cognitive psychology: experimental vs. correlational approach

In his presidential address to the *American Psychological Association*, Cronbach (1957) distinguished two disciplines of scientific psychology: experimental and correlational. This important distinction shaped both experimental / cognitive and differential psychology in the following decades, in particular the research approaches and questions being asked. The aim of cognitive (experimental) psychology is to identify and understand processes and mental representations that underlie human behavior and reasoning. Much of the cognitive research is employing group studies. This so-called “experimental approach” aims to precisely characterize cognitive mechanisms based on the typical or average response to a manipulation of environmental variables. In case statistically robust effects are found, and especially when these effects are replicated multiple times, conclusions are drawn for the general population.

Human behavior is characterized by many inter-individual differences. In experimental psychology, these inconsistencies are treated as noise or measurement error when averaged out in the search for the mean group effect. However, by collapsing data across participants, the experimental approach basically ignores, among other, the use of different cognitive strategies for the same task, differences in learning or subjective judgment as well as the inherent normal variance in ability and capacity. Consequently, a wealth of scientifically and clinically relevant information gets lost in experimental research. To fully characterize the cognitive processes underlying human behavior, one therefore needs to treat between-subject variance as data rather than noise (Sauce & Matzel, 2013; Thompson-Schill, Braver, & Jonides, 2005; Vogel & Awh, 2008). Focusing on individual rather than group data allows researchers not only to characterize the actual prevalence of cognitive effects within the population, but also to relate cognitive abilities to brain structures, chemistry and functions. This so-called “correlational approach” is therefore increasingly employed in the domain of cognitive psychology (see Goodhew & Edwards, 2019 for a recent discussion). In other words, bridging the gap between experimental and differential psychology is getting more and more important.

Individual differences research and the problem of low reliability

Quantifying cognitive effects at the individual level to subsequently assess their prevalence in a population is, however, not straightforward. Psychometrically speaking, an individual’s score on a task used to capture a certain cognitive effect is just an estimate of the true effect together with a measurement error – at least, this is how the score is treated in traditional analyses. The reliability of a measurement is thus crucial for quantifying participants as revealing an effect or not. Apart from this, reliability also plays an important role in correlational studies. Namely, the upper bound of a correlation that can be obtained between two measures (i.e., between observed scores; Mead, 2005) is set by their reliabilities. This upper bound is equal to $sqrt(rel_1 * rel_2)$. As such, if the correlation between two constructs equals 1, then the highest correlation between the measures of these constructs is equal to the square root of the product of their reliabilities. Therefore, by using measurements with poor reliability, one may overlook true correlations among those measures even when they are strong and robust.

Despite its importance, the issue of reliability has been largely neglected in cognitive psychology measures (Maloney, Risko, Preston, Ansari, & Fugelsang, 2010). This may be partly due to the reason discussed recently by Hedge et al. (2018). Accordingly, the inherent goal of experimental psychology is to trace robust and reproducible cognitive phenomena. Reaching this goal seems more straightforward if a given phenomenon is apparent across all individuals and there is no major variation between participants within a sample. Such homogeneity across participants is, however, not ideal when it comes to estimating reliability. Reliability is usually measured by means of correlations (or sometimes intra-class correlations), where it depends not only on measurement error but also on between-subject variance. In other terms, reliability is influenced by both task and sample characteristics. It decreases not only for larger error variance, whilst holding inter-individual variance constant, but also for smaller inter-individual variance, whilst holding error variance constant. Reliability is thus likely lower in homogeneous samples usually required in experimental psychology for the assessment of robust cognitive effects.

This problem was discussed recently by Luck (2019), who differentiated between measurement reliability, defined as above, and precision. The latter refers to how stable the measurement is in an absolute sense (e.g., how stable the difference between reaction times in milliseconds between conditions is, depending on which trials are taken for averaging). Importantly, the precision does not depend on the sample being tested but can be estimated for each participant.

Quantifying cognitive effects at the individual level

Considering that the reliability of robust and reproducible cognitive phenomena is usually far from perfect, this inaccuracy needs to be taken into account when deciding whether an individual reveals a certain cognitive effect or not. One possibility normally adhered to in classical test theory - but so far rarely in experimental psychology - is to deal with the issue of low reliability and calculate a confidence interval (CI) around the true score. The idea is that each measurement is comprised of an underlying true score and a measurement error randomly distributed around it. Usually, these errors are assumed to be normally distributed. Calculation of the CI can be conducted using different approaches that we will describe below.

Psychometric approach. In the psychometric approach, CIs are calculated by means of accounting for measurement error. More concretely, the reliability of the measurement as well as the standard deviation (SD) of scores in the sample are considered to estimate the standard error of measurement (SEM) as follows: *SEM = SD(X) * [sqrt (1 – reliability(X))]*. Assuming a normal distribution for *X*, the SEM is then taken to provide a *(1- α)* CI around the individual observed score: 95% CI = X ± (1.96 * SEM), e.g., for *α* = 0.05. Unfortunately, this approach suffers from some limitations. First, it can only be used if the sample meets a fundamental homogeneity assumption. Namely, the SEM is assumed to be constant across all members of the population from which the sample is drawn to compute the reliability estimate (cf. e.g., Willmes, 2010). Second, although the psychometric approach accounts for measurement error (i.e., task characteristics), it does not solve the problem of dependence on sample characteristics (i.e., the CI is assumed to be the same for each participant, irrespective of whether his or her response pattern is characterized by small or large variation), a problem frequently criticized from the perspective of probabilistic test theory. A possible solution to this problem is to estimate reliability and SD from a larger representative sample of healthy individuals or any other population of interest to get relatively precise and stable estimates, but this is not ideal as CIs should be determined based on the stability of the behavioral response of an individual, without taking recourse to the performance of the other participants of the experiment. This is possible within the bootstrapping approaches discussed below.

Bootstrapping approaches. In the bootstrapping approaches, a random sample of trials responded to by a given participant is selected to calculate the effect of interest.

This procedure of random sampling (with replacement) is repeated numerous times (usually a few thousand) to obtain a distribution of estimates for the effect (see Rousselet, Pernet, & Wilcox, 2019 for a comprehensive introduction to bootstrap techniques and its usefulness in psychology). The range in which a certain proportion of these estimates *(1- α)* is located is subsequently determined by removing the lowest *α/2* and highest *α/2* parts of the distribution, and can be taken to be the participant's CI. This method will be referred to as *H1 bootstrapping*, where H1 stands for "alternative hypothesis", because it tests for the robustness of an effect depending on the trials being selected for averaging[1]. As an alternative, random sampling with replacement can be used to allocate trials from different experimental conditions to two sets. This mathematical procedure reflects the null hypothesis model perfectly (i.e., samples for both "conditions" are drawn from the same pool). Subsequently, it is tested whether the empirically observed difference between conditions is likely to be observed under the null hypothesis model. If the empirical difference is located outside the "middle" *1-α* proportion of the distribution of resampled differences, one might conclude with *1- α* confidence that the empirically observed difference is unlikely under H0. This approach can therefore be referred to as *H0 bootstrapping*, where H0 stands for "null hypothesis". Crucially, with these bootstrapping methods, the breadth of the CI may be different for all participants. The CI depends only on the stability of the behavioral response within the participant, and it is not influenced by the performance of other participants like in the psychometric approach. For this reason, the bootstrap CI refer to precision rather than reliability of the measurement (Luck, 2019). This feature makes them more suited for estimating prevalence.

[1] Bootstrap confidence intervals have already been utilized to quantify group-level effects (see Crollen & Noël, 2015 for an example in numerical cognition).

Current approach - uniform analysis of multiple existing data sets: strengths and weaknesses

In the current study, we are combining raw data from 18 studies conducted in several labs in order to gain new insights into prevalence of a cognitive phenomenon at an individual level. Such an approach is situated between a typical meta-analytic approach and multi-lab initiatives, and thus shares some of the advantages (and drawbacks) of both approaches.

On the one hand, the typical meta-analytic approach is based on extracting and comparing effect sizes from several (un)published studies. A big advantage of this approach is that it is possible to take into account data coming from an extraordinary high number of participants without requiring costly and effortful (novel) data collection. However, in most cases, researchers need to deal with aggregated data and only have the statistics reported by the authors of the original studies at their disposal, which does not allow them to control for differences in experimental designs, methods of data preprocessing, outlier exclusion, and analyses. These aspects as well as the statistics actually reported in the primary studies might differ considerably between studies, which makes direct comparisons less straightforward. Thorough control for these parameters requires them to be orthogonal between studies, which is often not the case. For instance, some labs routinely use certain task parameters and data preprocessing routines, whereas other labs consistently use other configurations. Therefore, it can be difficult to determine the source of potential discrepancies between results (see Silberzahn et al., 2018 for an extreme example). On the other hand, multi-lab initiatives (e.g., LeBel, 2015) are based on data collection using exactly the same task, data preprocessing and analyses, but this approach requires new data collection. It is also theoretically possible to systematically test for different parameter configurations to see their influence on observed effects. Because new data needs to be collected, sample sizes (and number of experiments considered) are usually smaller than in case of typical meta-analyses.

The approach utilized here is intermediate between these two and, as such, offers some advantages. In line with the meta-analytic approach, it is based on considerable sample sizes without requiring new data collection and, as for multi-lab initiatives, it allows analysis of the raw data sets, thereby avoiding differences in data preprocessing and statistical reporting. At the practical level of this particular study, to address questions we ask, one needs access to data at the single trial level, which are rarely available in meta-analyses. The present approach also allows for testing new predictions and making optimal use of data collected so far to provide firm novel conclusions. Such an approach can be particularly useful in case of phenomena investigated in a relatively similar manner across labs, which is the case for many cognitive effects. Compared to these advantages, the drawbacks remain quite limited (e.g., fewer studies than for typical meta-analyses and no possibility for controlling or unifying task parameters, which is possible in multi-lab studies). To sum up, the present approach seems to provide a useful methodology, especially in light of the recently discussed “replicability crisis” in psychology (see Chambers, 2017; Pashler & Wagenmakers, 2012). Although it cannot replace meta-analyses or multi-lab initiatives, in some cases it can provide interesting complementary evidence.

Taking the SNARC effect as an example

To illustrate, within the context of a uniform analysis of multiple data sets, how the psychometric and bootstrapping approaches can be used to calculate CIs around observed individual effect estimates to determine the absence/presence of cognitive effects at the individual level and subsequently their actual prevalence within the population, this study focuses on the so-called Spatial Numerical Association of Response Codes (SNARC) effect (Dehaene, Bossini, & Giraux, 1993). The SNARC effect reflects the observation that, in binary choice reaction time (RT) tasks on numbers, participants from Western countries respond faster to small numbers with left-sided response keys and faster to large numbers with right-sided response keys. The SNARC effect was first documented over 25 years ago and has since become one of the most thoroughly investigated phenomena in the domain of numerical cognition. The most typical task for measuring the SNARC effect is parity judgment where participants decide whether a given number is odd or even with a left or right key press; response mapping is reversed halfway through the experiment so that left- and right-sided responses are collected for each number in every participant. The SNARC effect is typically attributed to the spatial mental representation of numerical magnitude within a particular cultural and linguistic context (e.g., Dehaene et al., 1993; Shaki, Fischer, & Petrusic, 2009). Some nativist accounts have also been proposed (e.g., de Hevia, Veggiotti, Streri, & Bonn, 2017; Rugani, Vallortigara, Priftis, & Regolin, 2015) as well as ones based on conceptual coding (Gevers et al., 2010) or spatial coding in working memory (Fias & van Dijck, 2016). In this last view, small numbers would be mentally represented on the left part and large numbers on the right part of a visuo-spatial medium taking the form of a mental number line.

The SNARC effect has not only been replicated in many studies using parity judgment (see Cipora, Soltanlou, Reips, & Nuerk, 2019 for a large-scale online replication). It has also been reported in various other numerical tasks in both typical and atypical populations across various developmental stages (for reviews see Fischer & Shaki, 2014; Toomarian & Hubbard, 2018; Wood, Willmes, Nuerk, & Fischer, 2008 for a metaanalysis), generally highlighting robustness of the effect. Interestingly, despite its robustness at the population level, the SNARC effect is characterized by high inter-individual variability (Wood, Nuerk, & Willmes, 2006 for early investigations). Across several studies, it was consistently reported to be present in only about 70-80% of participants (see Table 1 for details and relevant references).

This prevalence estimate is based on the most common method of quantifying the effect (Fias, Brysbaert, Geypens, & D'Ydewalle, 1996; Lorch & Myers, 1990) that goes as follows: First, for each participant separately, RT differences (dRTs) are calculated for each number separately by subtracting the average (or median) RT of the right hand from the average RT of the left hand. Consequently, negative dRTs indicate an advantage of right-handed responses as compared to left-handed responses. Second, for each participant separately, these dRTs are regressed on their corresponding number magnitude employing the least squares method. The obtained unstandardized regression coefficient (or slope) is taken as a measure of the participant's SNARC effect. It is expressed in milliseconds and can be interpreted as the increase in right-handed RT advantage compared to the left-handed RT when the magnitude of the number increases by one. Accordingly, a more negative slope indicates a stronger SNARC effect; conversely, positive slopes indicate that the right-hand advantage decreases with increasing number magnitude, which reflects a reverse SNARC effect. Third, to test whether a significant SNARC effect is observed at the group level, the sample of individual slope estimates is then tested against zero by means of a one-sample *t*-test.

Table 1. Overview of previous studies on the SNARC effect considering individual differences.

Article	Experiment / Group	n	Number range	Mean slope (SD)[a]	n slopes < 0	% slopes < 0	n slopes > 0	Comments	SNARC effect reliability[b]
Hoffmann, Pigat, and Schiltz (2014)	Young adults	28	0-9	-7.78 (NR)	25	89.3	3	Data retrieved from the Fig. 2a	-
	Older adults	54	0-9	-15.48 (NR)	51	94.4	3		-
Hoffmann, Mussolin, Martin, and Schiltz (2014)	Control	38	0-9	-8.82 (NR)	34	89.5	4	-	-
	Math expert	38	0-9	-5.25 (NR)	10	26.3	28	-	-
	Math difficulty	19	0-9	-13.23 (NR)	19	100.0	0	-	-
Fischer (2008)	Exp. 2	100	1-8 or 1,2,8,9	-6.86 (12.31)	76	76.0	24	Approximated data based on the Fig. 2; Number range differed between participants	-
Shaki and Fischer (2008)	Exp. 1	18	1-4; 6-9	-9.90 (NR)	14	77.8	4	Data retrieved from the Fig. 1, the left-to-right reading condition	-
Shaki, Fischer, and Petrusic (2009)	Canadian participants	12	1-4; 6-9	-10.81 (12.12)	10	83.3	2	-	-
Fias, Lauwereyns, and Lammertyn (2001)	Exp. 1	24	0-9	-2.03 (3.4)	17	70.8	7	Task referred to orientation of triangle superimposed on number	-
	Exp. 4	23	0-9	-3.74 (5.0)	17	73.9	6	Task referred to orientation of a line placed next to number	-
Viarouge, Hubbard, and McCandliss (2014)	Average	35	1-4; 6-9	NR (NR)	28	80.0	7	Data retrieved from the Fig. 4	.71[c]
	Session 1	35	1-4; 6-9	NR (NR)	27	77.1	8	Data retrieved from the Fig. 3	-
	Session 2	35	1-4; 6-9	NR (NR)	26	74.3	9		-
Fattorini, Pinto, Rotondaro, and Doricchi (2015)	Exp. 1	60	1-4; 6-9	-9.4 (9.1)	53	88.3	7	Data retrieved from the Fig. 2 panel C	.75
Pinhas, Shaki, and Fischer (2014)	Exp. 1	54	1,2,8,9	-8.72 (10.73)	44	81.5	10	Data retrieved from the Fig. 3	-

Schwarz & Mueller (2006)	Exp. 2	23	0-9	-4.36 (4.94)	20	87.0	3	Bimanual condition. Data retrieved from the Fig. 3	-
van Dijck and Fias (2011)	Exp. 2.	36	1-6	-9.05 (NR)	27	75.0	9	-	-
Jonas, Spiller, Jansari, and Ward (2014)	Exp. 1, non-synaesthetes only	27	1-9	-1.83 (7.50) [ns]	16	59.3	11	Data retrieved from supplementary Table 1	-
Yang et al. (2014)	Adults	42	1-9	-6.07 (9.27)	35	83.3	7	Data retrieved from Table 4	-
Ninaus et al. (2017)	Young adults	25	1,2,8,9	-10 (7)	21	84.0	4	Reliabilities calculated based on standardized reaction times	.93
	Middle-aged adults	27	1,2,8,9	-10 (10)	23	85.2	4		.95
	Elderly	24	1,2,8,9	-16 (12)	22	91.7	2		.96
Cipora, Soltanlou, Reips, and Nuerk (2019)	-	1056	1-4; 6-9	-8.49 (10.28)	877	83	179	-	.43
OVERALL		**1833**	**-**	**-**	**1492**	**81.37**	**341**	**Weighted average, considering sample size. Unweighted average = 79.61%**	**-**

[a] Number of decimals reported as in the original studies (NR = not reported)

[b] Split-half, Spearman-Brown corrected (only if reported)

[c] Calculated based on .545 raw correlation between half-based slopes from both sessions collapsed, without excluding outliers, as reported by the authors, adjusted with Spearman-Brown correction.

[d] Data from supplementary Table 1 diverges from results reported in the main text as regards sample size and significance of the SNARC effect at the group level

Even though this method is commonly employed to quantify the SNARC effect, the use of unstandardized slopes has been criticized. First, they do not consider the fit of the regression model (Pinhas, Tzelgov, & Ganor-Stern, 2012). The SNARC effect is actually one of the few instances where a raw slope measure is deemed more important in many papers (including some of our own) than explained variance or goodness of fit of the regression model (for a similar argument see also Tzelgov, Zohar-Shai, & Nuerk, 2013). Moreover, unstandardized slopes are very sensitive to a participant's mean RT and to intra-individual variability in RT. Namely, there is a very high correlation between mean RT and SD(RT), usually above .8 (e.g., Cipora et al., 2016; Cipora & Nuerk, 2013). This means that values of all differential measures calculated in absolute units (milliseconds) become larger due to higher variance. Such an influence of larger general variability can likely be avoided by using standardized slopes. They can be easily obtained from the same individual regression analysis as the unstandardized ones. In case of a statistical model with one single regressor, the standardized slope is equal to the Pearson correlation between number magnitude and dRT. To approximate a normal distribution, standardized slopes are typically Fisher z-transformed. Using standardized measures instead of unstandardized ones may thus provide a more valid measure of the SNARC effect (Lyons, Nuerk, & Ansari, 2015).

To determine whether an individual participant demonstrates a SNARC effect, one usually simply checks whether the unstandardized or standardized slope is larger or smaller than 0. About 70 to 80% of participants exhibit slopes below 0 and are therefore considered as revealing the effect (cf. Table 1). However, as already discussed, this approach is too simplistic as it does not consider that a participant's slope is just an estimate of the true slope parameter together with measurement error. Classifying participants based on whether their SNARC slope is larger or smaller than 0 would only be valid if the slope is perfectly estimated without measurement error. In the psychometric model of classical test theory, this would be true only if the reliability of the SNARC effect equals 1. Studies estimating the reliability of the SNARC effect, however, indicate that this is clearly not the case: rather, reliability varies from low to very high (Tables 1 and 3; see Appendix A for an algorithm for calculating the reliability of the SNARC effect). This inaccuracy thus necessitates the calculation of CIs around the observed slope to quantify the SNARC effect at the individual level. Note that one can think of two different types of CI: one around an observed score of a given participant, and one around the sample mean. Here we consider the first perspective. Using the psychometric approach and assuming a normal distribution, the CI around the estimated slope is computed as follows: 95% CI = x ± (1.96 * SEM), where SEM is given by *SD(x) * [sqrt (1 – reliability(x))]*. If the CI does not contain 0 and the slope is negative, one may claim with predefined confidence that the participant reveals a reliable SNARC effect. Alternatively, if the CI does not contain 0 and the slope is positive, one may claim that the participant reveals a reliable reversed SNARC effect. Finally, if the CI contains 0, the participant should be classified as not revealing a reliable SNARC effect.

In the bootstrapping H1 approach, a participant is considered to reveal a reliable SNARC effect if a negative slope is observed irrespective of the random sample of RTs used to estimate the slope. More precisely, for each participant within each experimental cell (number × response hand), a random sample of trials can be selected (with replacement) and used to calculate the SNARC slope across the random samples from all numbers. This procedure is repeated a few thousand times to obtain the participant's CI. The presence of a reliable SNARC effect at a certain (1–α) confidence level is then again checked by examining whether the CI contains 0.

In the bootstrapping H0 approach, for each participant, two sets of responses are sampled with replacement from all (right-handed and left-handed) RTs to a given number with one set being labeled left-handed and the other right-handed responses. Subsequently, the SNARC slope is calculated. After repeating this procedure multiple times, the distribution of the resulting slopes is compared with the slope observed empirically. If the latter is located outside the *(1-α)* CI, one concludes with *1-α* confidence that the empirically observed slope is unlikely under H0.

Aim of the present study

The present study aims at demonstrating the potential of using a uniform analysis of multiple existing data sets for investigating the nature of cognitive phenomena under scrutiny. To illustrate how different methodological approaches can be used to quantify robust cognitive effects at the individual level and thereby to determine their actual prevalence within the population, we took the SNARC effect as a typical example. We collected several published and unpublished raw data sets for the typical parity judgment task. For each study separately, we calculated standardized and unstandardized slopes as quantifications of the SNARC effect, determined CIs based on both the psychometric and bootstrapping methods, and calculated the proportion of participants revealing a reliable SNARC effect according to both methods. The resulting proportions were then compared to the prevalence of the SNARC effect previously reported in the literature, when only single SNARC slope estimates had been considered (see Table 1).

Alongside this main goal, we also assessed which sample and task parameters might influence the proportion of participants revealing a reliable SNARC effect. Finally, we conducted additional analyses to investigate whether participants' age and gender as well as block order in the parity task affect the observed SNARC effect (see Appendices B & C).

METHODS

Included data sets

Data from 18 studies conducted in eight different labs in five different countries were included in the present analysis[2], for a total of 1016 participants (635 females and 361 males; for 20 participants, gender data was not available)[3]. Detailed information about each data set and respective sample characteristics are presented in Table 2. Only studies utilizing bimanual parity judgment of single-digit numbers administered to healthy adults (age range 17-81) were considered. Thirteen studies used numbers 1 to 4 and 6 to 9, while five studies used all single-digit numbers from 0 to 9; for the sake of comparability between data sets, numbers 0 and 5 were excluded *post-hoc* from the latter before any other operations were performed on the data. Both blocks of hand-to-parity assignment were administered immediately one after the other, except in (Nuerk, Wood, & Willmes, 2005), where parity judgments on numbers presented in different formats (e.g., auditory, dice patterns, etc.) were intermixed. In 14 studies, the order of blocks was counterbalanced between participants; in the remaining four studies, it was fixed. Response-to-key assignment changed only once in the middle of the experiment[4]. Additionally, in van Dijck et al. (unpublished a), 50% of the trials were presented with upright digits and 50% in italics for the sake of consistency with other tasks.

[2] Raw data sets are available at https://osf.io/n7szg/

[3] In Appendix B we present the analyses of gender and age differences in the SNARC effect.

[4] Block order did not influence the SNARC effect. See Appendix C for details.

Table 2. Overview of analyzed data sets.

Data set	Number range	Block order [a]	Blocks immediately one after another	Sample description					
				N	Age (SD)	Gender (f/m)	Mother tongue	Background	Comments
Cipora et al. (2009)	0-9	c	yes	43	22.0 (1.8)	33/10	Polish	Students from Jagiellonian University, Cracow	no restrictions
Cipora & Göbel (2013)	0-9	c	yes	51	23.2 (4.9)	30/21	English	Students from the University of York	no restrictions
Cipora & Nuerk (2013)	1-4; 6-9	c	yes	71	21.8 (1.9)	49/22	Polish	Humanities and math-related subjects students	no restrictions
Cipora et al. (2016)	1-4; 6-9	c	yes	44	27.9 (1.1)	6/38	Polish	Advanced doctoral students: math, engineering, humanities and social sciences	no physics students
Cipora (2014)	1-4; 6-9	c	yes	55	23.1 (3.7)	39/16	Polish	General population	no psychology students
Georges et al. (2016)	1-4; 6-9	f	yes	61	23.3 (3.2)	27/34	heterogeneous	Students from University of Luxembourg	no psychology students
Georges et al. (2017)	1-4; 6-9	f	yes	81	23.4 (3.2)	40/41	heterogeneous	Students from University of Luxembourg	no psychology students
Georges et al. (2018)	1-4; 6-9	f	yes	26	26.9 (3.3)	15/11	heterogeneous	Students from University of Luxembourg	no psychology students
Georges et al. (unpubl.)	1-4; 6-9	f	yes	55	24.8 (3.4)	35/20	heterogeneous	Students from University of Luxembourg	no restrictions

Ginsburg et al. (unpubl.)	1-4; 6-9	c	yes	192	19.4 (3.7)	161/31	French	Students from Université Libre de Bruxelles	no restrictions
Göbel (unpubl.)	0-9	c	yes	38	23.4 (4.8)	25/13	English	Students from the University of York	no restrictions
Göbel et al. (2015)	0-9	c	yes	55	20.1 (2.3)	45/10	English	Students from the University of York	no restrictions
Masson & Pesenti (unpubl.)	1-4; 6-9	c	yes	28	21.7 (1.8)	17/11	French	Students from Université catholique de Louvain	no restrictions
Nuerk et al. (2005)	0-9	c	no[b]	32	26.2 (4.9)	16/16	German	Students and research staff of the University Hospital Aachen	no restrictions
van Dijck et al. (2009)	1-4; 6-9	c	yes	40	19.2 (1.7)[d]	15/5/20?[d]	Dutch	Students from Ghent University	no restrictions
van Dijck et al. (unpubl., a)	1-4; 6-9[c]	c	yes	63	19.7 (1.9)	39/24	Dutch	Students from Thomas More University College Antwerp	no restrictions
van Dijck et al. (unpubl., b)	1-4; 6-9	c	yes	41	19.9 (2.1)[e]	23/18	Dutch	Students from Thomas More University College Antwerp	only smokers
van Dijck et al. (unpubl., c)	1-4; 6-9	c	yes	40	34.3 (18.3)	20/20	Dutch	General population	no restrictions

[a] c = counterbalanced between participants; f = fixed

[b] blocks were intermixed with other binary decision tasks, e.g., parity judgment of numbers presented in another modality

[c] numbers presented upright or in italics for consistency with other blocks

[d] demographic data of 20 participants was not available any more

[e] age data from 2 participants were missing

Analyses

Individual trial data from each data set were processed and the SNARC effect slopes computed using the same protocol[5] as follows. First, practice sessions and incorrect responses were discarded. Second, outlier RTs were removed. In the first step, correct responses less than 200 ms were treated as anticipations and discarded. Subsequently, a sequential trimming method was applied. Mean RT and SD(RT) were calculated for each participant separately, and correct RTs outside ±3 SD from an individual's mean RT were discarded. This procedure was repeated until means and SDs no longer changed. Correct RTs, which were retained after trimming, were used for further analysis. The proportion of RTs included in the analysis for each data set is indicated in Table 3. Third, unstandardized and standardized SNARC effect slopes were computed by using the individual regression method (unstandardized, Fias et al., 1996)[6] and by calculating Pearson correlations between dRTs and digit magnitude respectively (standardized). Fourth, the standardized regression slopes were Fisher z-transformed to approximate a normal distribution.

For each approach, the next step was then to calculate the 80% 90%, 95%, and 99% CIs from these data. Because the 90% CI seems to present a relatively good balance between width of the interval and the level of confidence, this CI will be discussed in more detail[7].

Psychometric approach. Reliabilities were calculated using the split-half method by splitting the valid trials into two parts based on order of appearance with the odd-

[5] Therefore, the descriptive statistics (e.g., mean RTs and SNARC effect) reported here might differ slightly from values presented in the corresponding papers already published.

[6] See Appendix A for more details regarding the data analysis algorithm used and R scripts for running the analyses.

[7] Interested readers can conduct the analyses for any other confidence level using the data and analysis scripts shared along with this paper (see below).

even method. For each part and each participant, the unstandardized and standardized SNARC effect regression slopes as calculated above were correlated with each other to obtain a reliability index. Then, Spearman-Brown correction was applied to adjust for (double) task length. Finally, SNARC effect slopes, SDs and reliability estimates were used to determine 80%, 90%, 95% and 99% CIs around the individual unstandardized and standardized slopes by calculating the SEM. This procedure was applied to each data set separately. For each, the proportions of participants revealing reliable negative slopes (i.e., CIs not containing zero and negative), or reliable positive slopes (i.e., CIs not containing zero and positive) were determined; the remaining proportion had neither a reliable positive or negative slope.

H1 Bootstrapping approach. H1 Bootstrap CIs were calculated by custom-built *R* scripts.[8] Within each participant × number × hand configuration, a sample of *n* trials was randomly selected from valid trials within this cell. The *n* parameter was equal to the number of times a given number was repeated in each response-to-hand assignment (the number of observations sampled in each run of bootstrapping procedures should not exceed the number of data points available; e.g., Rousselet et al., 2019)[9]. Based on this set of trials, unstandardized and standardized slopes were calculated. This procedure was repeated 5000 times. To determine the 80%, 90%, 95% and 99% CIs, the range containing the mid 80%, 90%, 95% and 99% respectively of these slopes was then calculated.

H0 Bootstrapping approach. H0 Bootstrap CIs were calculated by custom-built R scripts. Within each participant × number configuration, two samples of *n* trials were

[8] All R scripts are available at https://osf.io/n7szg/

[9] We acknowledge that in some cells the number of bootstrap samples was higher than the actual number of data points because some trials were excluded during data preprocessing, and the bootstrap script does not account for it.

randomly selected from valid trials within this cell. The *n* parameter was equal to the number of times a given number was repeated in each response-to-hand assignment. These samples were treated as “left hand” and “right hand” responses and were used to calculate unstandardized and standardized slopes. This procedure was repeated 5000 times. Subsequently, 80%, 90%, 95% and 99% CIs were built around 0 based on the range containing respectively the mid 80%, 90%, 95% and 99% of these slopes. Finally, it was checked whether a given participant’s SNARC effect slope was *outside* that CI[10].

Consistency between psychometric and bootstrapping approaches. To explore consistency between the three approaches, the proportions of participants who were classified as belonging to the same category for each pair of methods as well as for all three methods together were calculated both for unstandardized and standardized slopes.

Factors influencing the proportion of participants revealing a (reliable) SNARC effect. By use of study-level calculations (*n* = 18 data sets), we checked which variables influenced the proportion of participants revealing SNARC slopes smaller than zero as well as the proportions of participants revealing reliable SNARC effects.

Heterogeneity of effects. In order to check whether between-study differences were not solely due to random error, we tested for heterogeneity of effects using the MAJOR package of the JAMOVI software (The jamovi project, 2019)[11], JAMOVI files are shared in the Supplementary material. Specifically, we used the meta-analytic function for

[10] Importantly, the bootstrapping method used for calculating the H0 CI allows also to check how many of H0 bootstrap slopes were equal to or smaller than the empirical slope for a given participant. This proportion meets the traditional definition of the p value, being the probability of obtaining the difference as empirically observed or larger if the null hypothesis is true.

[11] JAMOVI files are shared at https://osf.io/n7szg/

testing correlations (for reliability estimates and SNARC slopes) and tests for proportions. We report the I^2 values, which correspond to the percentage of variation between studies due to heterogeneity rather than chance. According to COCHRANE guidelines (The Cochrane Collaboration, 2011), I^2 values of 30%-60% correspond to moderate heterogeneity, and values of 50%-90% represent substantial heterogeneity. Additionally, *p* values are reported. Since the null hypothesis of homogeneity is tested, significant effects indicate heterogeneity.

RESULTS

Overview

Table 3 summarizes the results of all analyses. The proportions of trials that were retained for SNARC effect calculation were very similar across data sets. In all data sets, robust SNARC effects were observed at the sample level as confirmed with one-sample *t*-tests (all *p*s at least ≤ .047; in 15 cases $p < .001$ for both unstandardized and standardized slopes). Boxplots illustrating the distribution of slopes across all experiments are presented in Appendix D. Standardized slopes[12] (before applying Fisher z-transformation) were tested for heterogeneity using meta-analytical approaches. The analysis did not reveal significant differences between studies ($I^2 <$ .001%, $p = .903$; see supplementary JAMOVI file).

[12] Calculating heterogeneity requires estimating effect sizes. Because standardized slopes themselves can be considered as effect sizes, running similar tests for unstandardized slopes (which would require transforming them to effect size units) would be redundant in that case.

Table 3. Summary of results. The full size and editable table can be accessed at https://osf.io/n7szg/

Study			Cipora et al., 2009		Cipora & Goebel, 2013		Cipora & Nuerk, 2013		Cipora et al., 2016		Cipora, 2014		Georges et al., 2016		Georges et al., 2017		Georges et al., 2018		Georges et al., unpublished	
Task parameters		Repetitions / condition	12		9		20		30		30		9		9		18		9	
		Fixation point [ms]	300		1500		300		300		300		300		300		300		300	
		Response deadline [ms]	1500		2000		2000		2000		2000		1300		1300		1300		1300	
		Intertrial interval [ms]	500		none		500		500		500		1,300		1,300		1,300		1,300	
Performance descriptives		Included trials [%]	91.53		91.16		89.71		91.86		91.11		91.95		91.62		90.57		92.05	
		Mean RT [ms]	541		608		529		533		570		554		564		585		527	
		SD(RT) [ms]	67		123		79		78		79		79		103		72		57	
		RT: Intra-individual variability [ms]	104		123		89		97		103		115		124		139		94	
SNARC descriptives		SNARC slopes < 0 [%]	88.37		80.39		71.8		79.5		81.8		78.69		75.31		84.62		87.27	
		SNARC slopes > 0 [%]	11.63		19.61		28.2		20.5		18.2		21.31		24.69		15.38		12.73	
		Slope	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std
		Mean SNARC slope	-6.72	-0.44	-7.51	-0.36	-4.79	-0.35	-5.06	-0.36	-4.85	-0.34	-9.20	-0.42	-7.93	-0.40	-8.74	-0.41	-8.52	-0.49
		SD (SNARC slope)	6.91	0.52	11.56	0.41	7.18	0.52	7.02	0.55	6.90	0.51	11.71	0.48	12.75	0.48	10.23	0.59	7.27	0.46
		Reliability	0.32	0.52	0.28	0.30	0.75	0.58	0.78	0.64	0.78	0.65	0.69	0.42	0.61	0.41	0.51	0.62	0.44	0.56
Psychometric approach	80 %	Consistent SNARC < 0 [%]	41.86	34.88	29.41	41.18	40.85	42.25	40.91	38.64	40.00	40.00	47.54	50.82	34.57	48.15	42.31	50.00	60.00	54.55
	Confidence	No consistent SNARC [%]	58.14	62.79	68.63	56.86	52.11	50.70	56.82	59.09	52.73	52.73	45.90	44.26	61.73	48.15	57.69	42.31	36.36	41.82
	intervals	Consistent SNARC > 0 [%]	0.00	2.33	1.96	1.96	7.04	7.04	2.27	2.27	7.27	7.27	6.56	4.92	3.70	3.70	0.00	7.69	3.64	3.64
	90 %	Consistent SNARC < 0 [%]	32.56	27.91	17.65	27.45	35.21	32.39	36.36	34.09	40.00	36.36	31.15	36.07	27.16	34.57	38.46	38.46	52.73	47.27
	Confidence	No consistent SNARC [%]	67.44	72.09	80.39	70.59	61.97	63.38	61.36	63.64	56.36	60.00	65.57	62.30	71.60	65.43	61.54	53.85	47.27	49.09
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	1.96	1.96	2.82	4.23	2.27	2.27	3.64	3.64	3.28	1.64	1.23	0.00	0.00	7.69	0.00	3.64
	95 %	Consistent SNARC < 0 [%]	20.93	23.26	9.80	27.45	28.17	25.35	36.36	29.55	38.18	32.73	29.51	26.23	22.22	25.93	34.62	23.08	41.82	40.00
	Confidence	No consistent SNARC [%]	79.07	76.74	88.24	70.59	69.01	70.42	61.36	68.18	58.18	63.64	68.85	73.77	76.54	74.07	65.38	69.23	58.18	60.00
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	1.96	1.96	2.82	4.23	2.27	2.27	3.64	3.64	1.64	0.00	1.23	0.00	0.00	7.69	0.00	0.00
	99 %	Consistent SNARC < 0 [%]	9.30	18.60	3.92	15.69	18.31	19.72	25.00	18.18	30.91	20.00	22.95	16.39	11.11	11.11	11.54	15.38	21.82	27.27
	Confidence	No consistent SNARC [%]	90.70	81.40	94.12	84.31	81.69	78.87	72.73	79.55	65.45	78.18	77.05	83.61	87.65	88.89	88.46	80.77	78.18	72.73
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	1.96	0.00	0.00	1.41	2.27	2.27	3.64	1.82	0.00	0.00	1.23	0.00	0.00	3.85	0.00	0.00
Bootstrap H1 approach	80 %	Consistent SNARC < 0 [%]	48.84	48.84	49.02	49.02	45.07	45.07	47.73	47.73	45.45	45.45	60.66	62.30	59.26	58.02	46.15	46.15	61.82	61.82
	Confidence	No consistent SNARC [%]	48.84	48.84	49.02	49.02	47.89	47.89	50.00	50.00	47.27	47.27	36.07	34.43	37.04	38.27	46.15	46.15	34.55	34.55
	intervals	Consistent SNARC > 0 [%]	2.33	2.33	1.96	1.96	7.04	7.04	2.27	2.27	7.27	7.27	3.28	3.28	3.70	3.70	7.69	7.69	3.64	3.64
	90 %	Consistent SNARC < 0 [%]	39.53	39.53	39.22	39.22	39.44	39.44	38.64	38.64	41.82	41.82	47.54	47.54	44.44	44.44	46.15	46.15	50.91	52.73
	Confidence	No consistent SNARC [%]	58.14	58.14	58.82	58.82	54.93	54.93	59.09	59.09	50.91	50.91	50.82	50.82	53.09	51.85	50.00	50.00	45.45	43.64
	intervals	Consistent SNARC > 0 [%]	2.33	2.33	1.96	1.96	5.63	5.63	2.27	2.27	7.27	7.27	1.64	1.64	2.47	3.70	3.85	3.85	3.64	3.64
	95 %	Consistent SNARC < 0 [%]	30.23	30.23	29.41	29.41	38.03	38.03	36.36	36.36	36.36	36.36	39.34	37.70	39.51	39.51	34.62	34.62	41.82	43.64
	Confidence	No consistent SNARC [%]	69.77	69.77	68.63	68.63	57.75	57.75	61.36	61.36	60.00	60.00	60.66	62.30	59.26	59.26	61.54	61.54	56.36	54.55
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	1.96	1.96	4.23	4.23	2.27	2.27	3.64	3.64	0.00	0.00	1.23	1.23	3.85	3.85	1.82	1.82
	99 %	Consistent SNARC < 0 [%]	16.28	16.28	23.53	23.53	23.94	23.94	29.55	29.55	34.55	34.55	22.95	22.95	20.99	20.99	30.77	30.77	29.09	29.09
	Confidence	No consistent SNARC [%]	83.72	83.72	76.47	76.47	73.24	73.24	68.18	68.18	61.82	61.82	77.05	77.05	79.01	79.01	65.38	65.38	70.91	70.91
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	0.00	0.00	2.82	2.82	2.27	2.27	3.64	3.64	0.00	0.00	0.00	0.00	3.85	3.85	0.00	0.00
Bootstrap H0 approach	80 %	Consistent SNARC < 0 [%]	46.51	32.56	43.14	29.41	45.07	30.99	45.45	34.09	45.45	34.55	60.66	42.62	56.79	41.98	46.15	42.31	60.00	41.82
	Confidence	No consistent SNARC [%]	51.16	67.44	54.90	68.63	47.89	63.38	52.27	63.64	47.27	61.82	36.07	55.74	39.51	56.79	46.15	50.00	36.36	56.36
	intervals	Consistent SNARC > 0 [%]	2.33	0.00	1.96	1.96	7.04	5.63	2.27	2.27	7.27	3.64	3.28	1.64	3.70	1.23	7.69	7.69	3.64	1.82
	90 %	Consistent SNARC < 0 [%]	37.21	23.26	37.25	19.61	38.03	23.94	38.64	25.00	40.00	23.64	44.26	26.23	43.21	23.46	42.31	23.08	47.27	30.91
	Confidence	No consistent SNARC [%]	60.47	76.74	60.78	78.43	56.34	73.24	59.09	72.73	52.73	74.55	54.10	73.77	54.32	76.54	53.85	69.23	49.09	69.09
	intervals	Consistent SNARC > 0 [%]	2.33	0.00	1.96	1.96	5.63	2.82	2.27	2.27	7.27	1.82	1.64	0.00	2.47	0.00	3.85	7.69	3.64	0.00
	95 %	Consistent SNARC < 0 [%]	27.91	20.93	29.41	9.80	36.62	18.31	36.36	18.18	36.36	12.73	32.79	16.39	34.57	16.05	34.62	15.38	40.00	20.00
	Confidence	No consistent SNARC [%]	72.09	79.07	70.59	90.20	59.15	80.28	61.36	79.55	60.00	87.27	67.21	83.61	65.43	83.95	61.54	80.77	58.18	80.00
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	0.00	0.00	4.23	1.41	2.27	2.27	3.64	0.00	0.00	0.00	0.00	0.00	3.85	3.85	1.82	0.00
	99 %	Consistent SNARC < 0 [%]	18.60	6.98	19.61	0.00	23.94	4.23	29.55	11.36	30.91	1.82	21.31	4.92	18.52	4.94	26.92	11.54	27.27	7.27
	Confidence	No consistent SNARC [%]	81.40	93.02	80.39	100.00	73.24	95.77	68.18	86.36	65.45	98.18	78.69	95.08	81.48	95.06	69.23	88.46	72.73	92.73
	intervals	Consistent SNARC > 0 [%]	0.00	0.00	0.00	0.00	2.82	0.00	2.27	2.27	3.64	0.00	0.00	0.00	0.00	0.00	3.85	0.00	0.00	0.00
% participants classified the same way (in 90% CI) with:		Psychometric & Bootstrap H1	90.70	81.40	74.50	84.30	93.00	83.10	93.20	91.00	91.00	87.30	75.41	75.41	79.01	79.01	88.46	88.46	83.64	94.55
		Psychometric & Bootstrap H0	93.02	95.30	76.50	92.20	94.40	90.10	93.20	91.00	92.70	85.50	78.69	88.52	80.25	88.89	84.62	84.62	80.00	80.00
		Bootstrap H1 & H0	97.67	81.40	98.00	76.50	98.60	78.90	100.00	86.40	98.20	76.40	96.72	73.77	98.77	75.31	96.15	73.08	96.36	74.55
		All approaches	90.70	79.10	74.50	76.50	93.00	76.10	93.20	84.10	91.00	74.50	75.41	68.85	79.01	71.60	84.62	73.08	80.00	74.55

Study			Ginsburg et al., unpublished		Goebel, unpublished		Goebel et al., 2015		Masson & Pesenti, unpublished		Nuerk et al., 2005		van Dijck et al., 2009		van Dijck et al., unpublished, a		van Dijck et al., unpublished, b		van Dijck et al., unpublished, c	
Task parameters		Repetitions / condition	20		9		9		16		10		16		24		24		32	
		Fixation point [ms]	500		1500		1500		1000		500		500		500		500		500	
		Response deadline [ms]	3000		2000		2000		no deadline		1500		no deadline		1500		1500		1500	
		Intertrial interval [ms]	750		none		none		500		none		500		1000		1000		1000	
Performance descriptives		Included trials [%]	90.94		89.03		90.36		92.47		89.51		90.72		89.41		87.94		90.34	
		Mean RT [ms]	502		646		567		513		484		509		562		581		564	
		SD(RT) [ms]	55.67		148		70		49		57		73.5		59.58		62.47		87.33	
		RT: Intra-individual variability [ms]	91.39		145.6		104		88		78.7		98.23		95.21		104		101.08	
SNARC descriptives		SNARC slopes < 0 [%]	70.31		89.47		74.55		67.86		59.38		75		69.84		87.8		82.5	
		SNARC slopes > 0 [%]	29.69		10.53		25.45		32.14		40.62		25		30.16		12.2		17.5	
		Slope	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std	Unstd	Std
		Mean SNARC slope	-2.74	-0.23	-13.18	-0.44	-5.91	-0.32	-3.75	-0.27	-2.72	-0.21	-4.49	-0.31	-4.734	-0.34	-10.41	-0.66	-7.199	-0.475
		SD (SNARC slope)	6.13	0.51	12.88	0.37	9.19	0.41	8.00	0.57	7.44	0.47	5.81	0.47	6.5754	0.48	8.63	0.53	8.8964	0.5445
		Reliability	0.66	0.59	0.56	0.37	0.49	0.27	0.72	0.75	0.67	0.59	0.39	0.33	0.72	0.71	0.82	0.74	0.87	0.87
Psychometric approach	80 %	Consistent SNARC < 0 [%]	34.90	34.38	50.00	47.37	36.36	36.36	42.86	46.43	28.1	37.50	37.50	32.50	47.62	53.97	73.17	75.61	62.50	72.50
	Confidence	No consistent SNARC [%]	56.25	56.25	50.00	50.00	58.18	58.18	46.43	42.86	65.6	59.38	57.50	62.50	47.62	38.10	24.39	19.51	30.00	15.00
	intervals	Consistent SNARC > 0 [%]	8.85	9.38	0.00	2.63	5.45	5.45	10.71	10.71	6.3	3.13	5.00	5.00	4.76	7.94	2.44	4.88	7.50	12.50
	90 %	Consistent SNARC < 0 [%]	28.65	28.13	42.11	39.47	27.27	27.27	39.29	32.14	25.0	31.25	37.50	25.00	38.10	42.86	70.73	65.85	55.00	60.00
	Confidence	No consistent SNARC [%]	64.58	65.10	57.89	57.89	69.09	69.09	50.00	57.14	68.8	65.63	62.50	72.50	57.14	50.79	26.83	29.27	37.50	27.50
	intervals	Consistent SNARC > 0 [%]	6.77	6.77	0.00	2.63	3.64	3.64	10.71	10.71	6.3	3.13	0.00	2.50	4.76	6.35	2.44	4.88	7.50	12.50
	95 %	Consistent SNARC < 0 [%]	22.40	21.88	36.84	36.84	20.00	18.18	39.29	28.57	15.6	28.13	32.50	20.00	34.92	38.10	65.85	56.10	55.00	55.00
	Confidence	No consistent SNARC [%]	72.92	73.44	63.16	63.16	78.18	81.82	50.00	64.29	78.1	68.75	67.50	77.50	61.90	57.14	34.15	41.46	40.00	40.00
	intervals	Consistent SNARC > 0 [%]	4.69	4.69	0.00	0.00	1.82	0.00	10.71	7.14	6.3	3.13	0.00	2.50	3.17	4.76	0.00	2.44	5.00	5.00
	99 %	Consistent SNARC < 0 [%]	13.02	11.46	15.79	26.32	10.91	5.45	10.71	25.00	12.5	9.38	7.50	7.50	28.57	28.57	56.10	48.78	50.00	50.00
	Confidence	No consistent SNARC [%]	85.42	84.90	84.21	73.68	87.27	94.55	85.71	71.43	87.5	90.63	92.50	92.50	71.43	71.43	43.90	51.22	45.00	47.50
	intervals	Consistent SNARC > 0 [%]	1.56	3.65	0.00	0.00	1.82	0.00	3.57	3.57	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	2.50
Bootstrap H1 approach	80 %	Consistent SNARC < 0 [%]	37.50	37.50	47.37	47.37	43.64	43.64	50.00	50.00	31.25	31.25	40.00	40.00	49.21	49.21	78.05	78.05	65.00	65.00
	Confidence	No consistent SNARC [%]	52.60	52.60	50.00	50.00	49.09	49.09	35.71	35.71	56.25	56.25	55.00	55.00	42.86	42.86	19.51	19.51	22.50	22.50
	intervals	Consistent SNARC > 0 [%]	9.90	9.90	2.63	2.63	7.27	7.27	14.29	14.29	12.50	12.50	5.00	5.00	7.94	7.94	2.44	2.44	12.50	12.50
	90 %	Consistent SNARC < 0 [%]	30.73	30.73	42.11	42.11	34.55	34.55	42.86	42.86	28.13	28.13	32.50	32.50	46.03	46.03	73.17	73.17	62.50	62.50
	Confidence	No consistent SNARC [%]	62.50	62.50	55.26	55.26	60.00	60.00	46.43	46.43	68.75	68.75	65.00	65.00	50.79	50.79	24.39	24.39	27.50	27.50
	intervals	Consistent SNARC > 0 [%]	6.77	6.77	2.63	2.63	5.45	5.45	10.71	10.71	3.13	3.13	2.50	2.50	3.17	3.17	2.44	2.44	10.00	10.00
	95 %	Consistent SNARC < 0 [%]	24.48	24.48	34.21	34.21	23.64	23.64	39.29	39.29	21.88	21.88	25.00	25.00	42.86	42.86	68.29	68.29	57.50	57.50
	Confidence	No consistent SNARC [%]	71.88	71.88	65.79	65.79	72.73	72.73	53.57	53.57	75.00	75.00	75.00	75.00	55.56	55.56	29.27	29.27	32.50	32.50
	intervals	Consistent SNARC > 0 [%]	3.65	3.65	0.00	0.00	3.64	3.64	7.14	7.14	3.13	3.13	0.00	0.00	1.59	1.59	2.44	2.44	10.00	10.00
	99 %	Consistent SNARC < 0 [%]	16.67	16.67	23.68	23.68	14.55	14.55	32.14	32.14	15.63	15.63	20.00	20.00	31.75	31.75	51.22	51.22	50.00	50.00
	Confidence	No consistent SNARC [%]	81.77	81.77	76.32	76.32	85.45	85.45	60.71	60.71	84.38	84.38	80.00	80.00	68.25	68.25	48.78	48.78	42.50	42.50
	intervals	Consistent SNARC > 0 [%]	1.56	1.56	0.00	0.00	0.00	0.00	7.14	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.50	7.50
Bootstrap H0 approach	80 %	Consistent SNARC < 0 [%]	38.54	27.08	47.37	36.84	41.82	27.27	50.00	28.57	31.25	28.13	40.00	32.50	49.21	33.33	78.05	53.66	65.00	50.00
	Confidence	No consistent SNARC [%]	52.08	66.67	50.00	60.53	52.73	70.91	35.71	64.29	59.38	68.75	55.00	65.00	46.03	65.08	19.51	46.34	20.00	47.50
	intervals	Consistent SNARC > 0 [%]	9.38	6.25	2.63	2.63	5.45	1.82	14.29	7.14	9.38	3.13	5.00	2.50	4.76	1.59	2.44	0.00	15.00	2.50
	90 %	Consistent SNARC < 0 [%]	30.73	16.67	42.11	28.95	32.73	16.36	42.86	25.00	25.00	12.50	27.50	20.00	46.03	22.22	73.17	48.78	62.50	32.50
	Confidence	No consistent SNARC [%]	63.02	79.17	55.26	71.05	63.64	83.64	50.00	71.43	71.88	87.50	67.50	77.50	50.79	77.78	24.39	51.22	27.50	65.00
	intervals	Consistent SNARC > 0 [%]	6.25	4.17	2.63	0.00	3.64	0.00	7.14	3.57	3.13	0.00	5.00	2.50	3.17	0.00	2.44	0.00	10.00	2.50
	95 %	Consistent SNARC < 0 [%]	23.44	9.90	31.58	15.79	20.00	9.09	35.71	17.86	21.88	9.38	25.00	15.00	41.27	17.46	65.85	36.59	57.50	22.50
	Confidence	No consistent SNARC [%]	73.44	87.50	68.42	84.21	78.18	90.91	57.14	78.57	75.00	90.63	75.00	82.50	57.14	82.54	31.71	63.41	35.00	77.50
	intervals	Consistent SNARC > 0 [%]	3.13	2.60	0.00	0.00	1.82	0.00	7.14	3.57	3.13	0.00	0.00	2.50	1.59	0.00	2.44	0.00	7.50	0.00
	99 %	Consistent SNARC < 0 [%]	14.58	2.08	18.42	2.63	7.27	0.00	25.00	0.00	15.63	3.13	17.50	5.00	30.16	3.17	48.78	17.07	47.50	7.50
	Confidence	No consistent SNARC [%]	83.85	97.92	81.58	97.37	92.73	100.00	67.86	100.00	84.38	96.88	82.50	95.00	69.84	96.83	51.22	82.93	45.00	92.50
	intervals	Consistent SNARC > 0 [%]	1.56	0.00	0.00	0.00	0.00	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.50	0.00
% participants classified the same way (in 90% CI) with:		Psychometric & Bootstrap H1	90.60	85.90	81.60	86.80	90.90	87.30	96.40	89.30	87.50	84.40	87.50	82.50	87.30	84.10	92.70	90.20	90.00	90.00
		Psychometric & Bootstrap H0	91.10	85.90	81.60	86.80	87.30	85.50	92.90	85.70	90.60	78.10	85.00	95.00	87.30	73.00	92.70	78.00	90.00	62.50
		Bootstrap H1 & H0	99.50	80.20	100.00	78.90	96.40	72.70	96.40	75.00	96.90	81.30	92.50	82.50	100.00	73.00	100.00	73.20	100.00	62.50
		All approaches	90.60	76.00	81.60	76.30	87.30	72.70	92.90	75.00	87.50	71.90	82.50	80.00	87.30	65.10	92.70	70.70	90.00	57.50

Reliability

Data sets differed considerably with regards to the reliability of the SNARC effect. In 10 cases, reliabilities of unstandardized SNARC slopes were satisfactory (i.e., ≥ .65), but mostly lower than typical psychometric standards for individual diagnosis purposes (for personality assessment for adults, diagnostics textbooks often require at least r = .80). Heterogeneity analysis indicated systematic variation in reliability estimates between studies (I^2 = 75.88% p < .001 and I^2 = 71.13%, p < .001 for unstandardized and standardized slopes respectively).

There were no systematic differences between reliabilities of unstandardized and standardized slopes. In 13 data sets, reliabilities were higher for unstandardized slopes, while, in one case, they were equal for unstandardized and standardized slopes up to the second decimal (Table 3).

Prevalence of the SNARC effect

Weighted averages (considering sample sizes in each study) of proportions of participants revealing reliable SNARC, reliable reversed SNARC, or no SNARC effect for different confidence levels for all three methods are summarized in Figure 1. Exact percentages for each study are presented in Table 3.

No confidence intervals. When no confidence intervals are taken into account, the proportions of participants revealing reliable SNARC, reliable reversed SNARC, or no SNARC effect is the same for unstandardized and standardized SNARC slopes. The proportions of participants revealing positive and negative slopes differed between studies (from 59% to 89% for negative slopes). This range is in line with the proportions reported in the literature (Table 1). Systematic differences were confirmed by heterogeneity analysis (I^2 = 49.86%, p = .006). The weighted average (considering sample size) of participants revealing negative slopes was 76.77%.

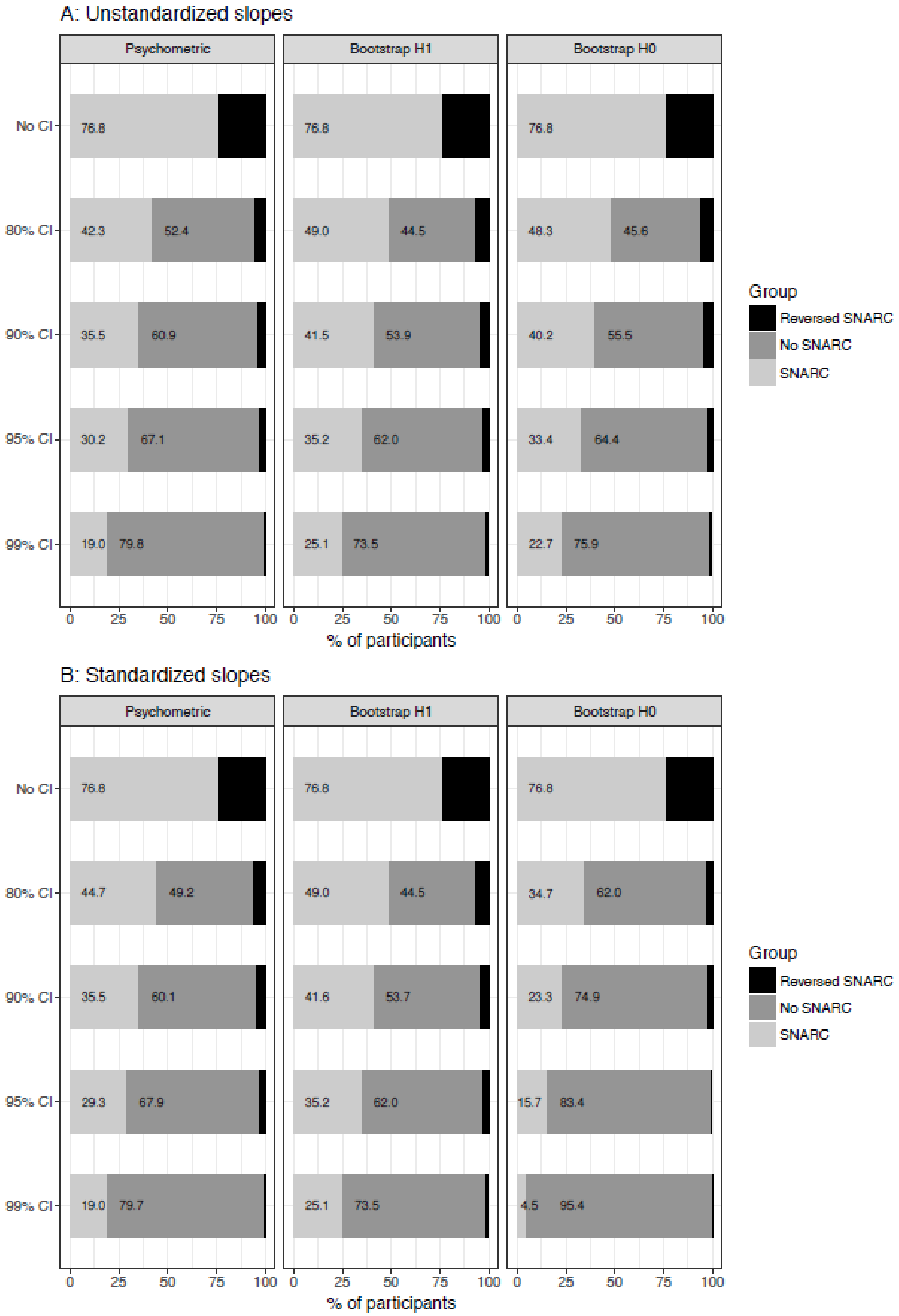

Figure 1. Weighted (by sample size) proportion of participants revealing either a SNARC effect, no SNARC effect, or a reverse SNARC effect for different confidence levels within the psychometric, bootstrapping H1, and bootstrapping H0 approaches. Panel A represents results for unstandardized SNARC slopes, and Panel B for standardized ones. Respective percentage values are presented inside

corresponding parts of bars. The top bar represents the proportion of negative and positive slopes when no confidence intervals are considered. Note that the top bars are duplicated in each column to enable direct comparison with proportions.

Psychometric approach. As expected, the proportions of participants revealing reliable SNARC, reliable reversed SNARC, or no SNARC effect vary considerably depending on the confidence level chosen. A more detailed look at the 90% CIs shows that, across studies, we can say with 90% confidence that about 20% to 70% (weighted average 35.5% for both unstandardized and standardized slopes) of participants reveal reliable SNARC effect. These proportions differed significantly between studies as indicated by heterogeneity analyses (I^2s ≥ 63.46%, *p*s < .001). There were some minor differences between estimates for unstandardized and standardized slopes, but these were not systematic between studies. Noteworthy, the proportion exceeded 50% only in the case of three studies, and was below 45% in all other 15 cases. Thus, the prevalence of a reliable SNARC effect (with 90% confidence interval) is lower than the proportion of negative slopes. On the other hand, the proportion of participants revealing a reliable reverse SNARC effect is very low (never more than 13% and typically below 5%), weighted average 3.6% and 4.4% for unstandardized and standardized slopes respectively. Interestingly, we found no evidence either for reliable SNARC or reliable reverse SNARC effects in 27% to 80% of participants.

H1 bootstrapping approach. Again, the proportions of participants revealing reliable SNARC, reliable reversed SNARC, or no SNARC effect differed considerably across the chosen confidence levels. A closer look at the 90% CI shows that the proportion of participants revealing reliable negative slopes varied between 30% and 75% (weighted average 41.5% and 41.6% for unstandardized and standardized slopes respectively). These proportions differed significantly between studies as indicated by heterogeneity analyses (I^2s ≥ 63.59%, ps < .001). Proportions did not systematically differ between unstandardized and standardized slopes. Again, the proportion of individuals revealing a reliable reverse SNARC effect never exceeded 11% (weighted average 4.6% and 4.7% for unstandardized and standardized slopes respectively). Interestingly, we found no evidence either for reliable SNARC or reliable reverse SNARC effects in more than half of the participants.

H0 bootstrapping approach. As with the other methods, the proportions of participants revealing reliable SNARC, reliable reversed SNARC, or no SNARC effect differed considerably between confidence levels. A closer look at the 90% CI shows that the proportion of participants revealing reliable negative slopes varied between 20% and 75% (weighted average 40.2% and 23.3% for unstandardized and standardized slopes, respectively). As indicated by heterogeneity analyses, the proportions differed significantly between studies for the unstandardized slopes (I^2 = 66.84%, p < .001), but not for the standardized slopes (I^2 = 31.48%, p = .072). The proportion was always lower for standardized slopes. Again, the proportion of a reliable reverse SNARC effect never exceeded 10% (weighted average 4.3% and 1.8% for unstandardized and standardized slopes respectively). Again, we found no evidence either for reliable SNARC or reliable reverse SNARC effects in more than half of the participants.

Comparison between the psychometric and bootstrapping approaches

The proportion of participants who were classified as belonging to the same category with all three methods when the 90% CIs are considered varied between 74.5% and 93.2% for unstandardized and 57.5% and 84.1% for standardized slopes. Congruencies between all pairs of methods were very similar (Table 3). Study-level calculation revealed that the proportion of participants showing a reliable SNARC effect differed significantly between methods both for unstandardized, $F(2, 34) = 9.70$, $p < .001$, $\eta_p^2 = .36$, as well as for standardized SNARC slopes, $F(2, 34) = 157.98$, $p < .001$, $\eta_p^2 = .90$. For unstandardized SNARC slopes, the average proportion was smallest for the psychometric method compared to both bootstrapping methods (*p*s ≤ .011, HSD corrected), whereas proportions did not differ between bootstrapping methods ($p = .478$, HSD corrected). For the standardized slopes, proportions differed between all methods (*p*s < .001, HSD corrected), with the lowest average proportion for the H0 bootstrapping method, and the largest one for the H1 bootstrapping method[13].

At the study level, correlations between proportions of participants revealing reliable SNARC effects were very high[14]. For unstandardized slopes, proportions estimated with the psychometric approach correlated at .84 ($p < .001$) with the proportions estimated with the H1 bootstrap and at .83 ($p < .001$) with the H0 bootstrap. The correlation between proportions estimated with the two bootstrap methods was .99 (*p* <. 001). For standardized SNARC slopes, the respective correlations were $r = .94$ ($p < .001$), $r = .86$ ($p < .001$), and $r = .93$ ($p < .001$). A similar pattern of results was observed for the reliable reverse SNARC effect (see supplementary JAMOVI files for details).

Factors related to proportions of participants revealing (reliable) SNARC slopes

Because the heterogeneity analysis yielded significant differences between studies regarding proportions of participants revealing (reliable) SNARC slopes, we analyzed which factors were related to these proportions.

[13] Analyses were conducted on proportions transformed according to the 2*ARCSIN(SQRT(proportion)) formula. However, analyses on untransformed proportions show virtually the same results.

[14] Note that the heterogeneity analysis shows that studies differed significantly in this respect, which justifies such an analysis.

No confidence intervals considered. The proportion of participants revealing negative SNARC slopes was related to slope size both for unstandardized and standardized slopes (*r* = -.79, *p* < .001; *r* = -.81, *p* < .001 respectively): in data sets showing a more pronounced SNARC effect, the proportion of participants revealing negative slopes was larger. The proportion of participants revealing negative SNARC slopes was also related to RT characteristics, such as mean RT (*r* = .65, *p* = .004) and RT variability (*r* = .60, *p* = .008). Longer and more variable RTs corresponded to larger proportions of participants revealing negative slopes. On the other hand, the proportion was not related to the SD of the unstandardized (*r* = .31, p = .210) and standardized (*r* = -.06, *p* = .824) slopes and to slope reliability (*r* = -.23, *p* = .349 for unstandardized and *r* = -.05, *p* = .839 for standardized slopes).

Confidence intervals considered. The proportion of individuals revealing negative SNARC slopes was higher in studies in which the mean slope was steeper. With the psychometric approach, it marginally relates to the mean unstandardized slope (*r* = -.40, *p* = .097), and more strongly to the mean standardized slope (*r* = -.78, *p* < .001). In case of H1 bootstrapping, the correlation with mean slopes is even more pronounced (*r* = -.58, *p* = .012 and *r* = -.88, *p* < .001 respectively for unstandardized and standardized slopes). For H0 bootstrapping, *r* equals -.55 (*p* = .018) and -.91 (*p* < .001) for unstandardized and standardized slopes respectively.

Unsurprisingly, the proportion of participants revealing reliable SNARC effects was higher in data sets in which more participants revealed negative SNARC slopes (psychometric approach: *r* = .50, *p* = .034 for unstandardized, *r* = .44, *p* = .071 for standardized slopes; H1 bootstrapping approach: *r* = .55, *p* = .017 for unstandardized, *r* = .56, *p* = .016 for standardized slopes; H0 bootstrapping: *r* = .52, *p* = .026 for unstandardized, *r* = .63, *p* = .005 for standardized slopes).

The proportion of participants revealing reliable SNARC effects was also related to task reliability. This direction of the relationship was similar for psychometric (*r* = .44, *p* = .065 and *r* = .63, *p* = .005, for unstandardized and standardized slopes respectively), and bootstrapping approaches (for H1 *r* = .41, *p* = .097 and *r* = .51, *p* = .030, for unstandardized and standardized slopes respectively; for H0 *r* = .47, *p* = .052 and *r* = .43, *p* = .076, for unstandardized and standardized slopes respectively)[15]. Higher reliability was related to a larger proportion of participants revealing a reliable SNARC effect. However, inspection of the respective scatterplots (see supplementary JAMOVI files) shows that these correlations were mostly driven by two studies (van Dijck et al., unpublished b and c), yielding relatively high reliability estimates compared to the remaining studies.

Furthermore, for the psychometric approach, the proportion of participants revealing reliable negative slopes correlated moderately with the number of repetitions (*r* = .47, *p* = .048 and *r* = .44, *p* = .066, for unstandardized and standardized slopes respectively). This was likely due to higher reliabilities obtained in these studies. This relationship did not reach significance with the bootstrapping approaches (H1: *r* = .35, *p* = .150 and *r* = .34, *p* = .172, for unstandardized and standardized slopes, respectively; H0: *r* = .40, *p* = .101 and *r* = .32, *p* = .196, for unstandardized and standardized slopes, respectively). A larger number of repetitions was thus related to a larger proportion of individuals revealing a reliable SNARC effect. This is likely to be due to the fact that with greater reliability, the CIs get narrower, and more individuals can be classified as revealing a reliable SNARC effect.

[15] Please note that in several instances, correlations do not get significant, and that we do not correct for multiple comparisons here (it is not clear for how many comparisons to correct). Therefore, these results should be interpreted with caution.

Complementary results showed that the proportion of participants revealing a reliable reverse SNARC effect increased with an increasing number of repetitions and reliability (see Supplementary JAMOVI files).

DISCUSSION

Overview

Using a uniform analysis of multiple existing data sets, we investigated the prevalence of a robust group-level phenomenon at the individual level (i.e. the SNARC effect; left/right hand advantage when responding to small/large magnitude numbers). For this, we used a psychometric approach to estimate confidence intervals around individual effect estimates based on task reliability and sample variance and two bootstrapping techniques (that allow the estimation of confidence intervals independent of sample characteristics). We show that the robust group-level SNARC effect is driven by only a relatively small proportion of participants (< 45%) who reliably show the effect at the individual level. In other words, the significant SNARC effect at the group-level seems to be produced by only a minority of subjects in the sample. By applying this analysis to an extraordinary large number of participants and multiple experiments, we ensure robustness of conclusions. These results clearly indicate that significant group-level results cannot be considered as evidence that a cognitive effect is present in the entire population (and thus for the universality of the effect). This insight is important for the development of theories and models, which are based on such group-level observations. Although we focused on the SNARC effect, there is no reason to assume that our results are limited to this effect. As such, our results raise questions about the prevalence of other well-established cognitive phenomena in and outside the domain of numerical cognition. The logic and methods we present here could serve as a starting point for similar investigations in other areas of cognitive psychology. The uniform analysis of multiple existing data sets (obtained from multiple labs) approach seems to be fruitful for reliable results.

Chasing the SNARC effect

Prevalence of the SNARC effect

In the present study, we aimed to investigate the prevalence of a reliable SNARC effect. The uniform reanalysis of 18 data sets revealed its presence at the sample/ group level in each data set, and the pattern of *p* values at the sample level (15 out of 18 *p*s < .001) indicates its robustness. The slope estimates as well as the proportion of participants revealing a regular (i.e., smaller than zero) SNARC slope were similar to those previously reported in the literature (Table 1; see also Wood et al., 2008 for a meta-analysis). However, out of the individual participants showing a negative slope, we demonstrate that a reliable non-zero effect at the individual level was present only for a minority, i.e., about 35% of all participants for the psychometric approach (Figure 1). This proportion was slightly larger with the bootstrapping approaches, but did still not exceed 45% of the participants. Interestingly, a reliable reverse SNARC effect was very rare. Irrespective of the estimation method, it was present in less than 5% of participants, and never exceeded 13%. These proportions refer only to the effect for the data from a single experimental session. Investigating whether the presence of a reliable effect is stable across multiple experimental sessions administered to the same participant can be even more challenging.

Crucially, in case of about 50 to 60% of the participants, we did not find a reliable SNARC effect. It is important to note that this refers to the absence of evidence for the SNARC effect rather than evidence for no SNARC effect in these participants. There could be many reasons why no effect is observed in these participants (see below). On the other hand, the fact that the reliability of the SNARC effect is typically within a reasonable range (see Table 3) and not correlated (if anything only marginally driven by 2 studies) with the proportion of participants showing a reliable SNARC, suggests that the large proportion of subjects who do not show SNARC is not due to an unsuited task/ instrument. One possible way to address this shortcoming more directly in the future would be to apply a Bayesian approach.

(Non)paradoxical SNARC effect reliability

High reliabilities (> .80) were observed in two data sets (van Dijck, unpublished b and especially c). In both studies, very heterogeneous samples were studied. In study b, 1st semester students of the bachelor program in applied psychology were tested. Their educational background was very heterogeneous including both professional education (e.g., bakery, tourism, social technical education) as well as general secondary education (math-modern languages, economy-modern languages, STEM, humanities). In study c, participants from heterogeneous groups and at different ages were tested. Most of them were not university students, and several participants did not have higher education degree. This clearly shows that when testing heterogeneous samples, the SNARC effect can be highly reliable. Notably, although reliability was high, the proportion of participants revealing a reliable SNARC effect did not reach 80%, even when the most liberal 80% CI criterion was taken into account. This can be an indirect argument that the lower-than-expected prevalence of the SNARC effect cannot be solely attributed to a high type II error.

Reliability strongly depends on the inter-individual variability of the property/trait assessed in the sample. Keeping other parameters constant, an increase in sample variance of true scores leads to an increase in reliability (Cooper, Gonthier, Barch, & Braver, 2017). In some cases, low reliability can still account for at least part of the low prevalence of the effect. It can affect prevalence estimates of both psychometric (due to imprecision of the measurement itself and to low sample variance leading to lower reliability estimates) and bootstrapping methods (only due to imprecision of the measurement). The imprecision of measurement refers to low consistency in the RT pattern within individuals (in absolute terms, expressed e.g., in milliseconds). Similarly, if there is no consistent pattern within an individual's RTs, slopes estimated with a bootstrap method would also strongly differ between each other. This could be the case for instance because of task characteristics (i.e., too few repetitions, suboptimal trial timings, etc.) or testing conditions. With this in mind, one might think about three recommendations for future correlational studies: (1) in studies aimed at testing individual differences in the SNARC effect in particular, the experimenters should ensure that they recruit heterogeneous samples, at least not limiting themselves to student populations, (2) reliability estimates observed in samples drawn from one population cannot be generalized to different populations and must be determined empirically for each population of interest, (3) if homogeneous samples are tested one might still increase the precision of the SNARC effect measurement (which could be then seen as relatively narrow bootstrap CIs), but such data would not be very useful for correlational analyses.

Improving SNARC effect reliability and precision

To accurately investigate the prevalence of the SNARC effect, irrespective of the method used, future studies need to ensure sufficient reliability and precision of the SNARC effect measurement. Below we list some recommendations on how to reach this goal. These recommendations seem to be valid also in case of other cognitive phenomena.

Sample heterogeneity. As discussed in the section above, testing heterogeneous samples which truly reflect the variance in the population of interest (e.g., when willing to make inferences about the general population one should not test only students) would be the first solution. Nevertheless, it must be noted that low reliabilities are observed for some SNARC studies, despite the SNARC effect being much more variable than other robust cognitive phenomena. As such, the claim of Hedge et al. (2018) that low reliability of robust phenomena can be attributed to low between-subject variability is not adequate in the case of the SNARC effect (which despite being more variable is still sometimes not very reliable, possibly due to low precision). However, there are several other means by which reliability and precision can be improved.

Task length. Increasing the number of repetitions per condition can be applied for all diagnostic assessment procedures with homogeneous items. The role of the number of trials in experimental psychology tasks has been recently raised as an important factor to be considered to ensure sufficient power (Baker et al., 2019; Brysbaert, 2019). When more trials are included for averaging, the errors associated with the single measurements tend to cancel each other out more accurately[16]. In the case of the SNARC effect, increasing the number of repetitions of each number per block up to 20 or 30 seems a reasonable solution, because the duration of the task does not change dramatically and still can be well below 20 minutes (for simulations, see Cipora & Wood, 2017). In the presented analyses, reliability is at least at an acceptable level (≥ .66) in all studies using 20 and more repetitions[17]. Importantly,

[16] However, this holds true only if the error is random. If there are systematic trends in the experiment, due to fatigue or learning for example, making the experiment longer might not solve the problems. This should be considered each time one wishes to apply this logic to cognitive phenomena.

[17] This estimate holds specifically for the SNARC effect. This number should be estimated specifically for each phenomenon.

increasing the number of trials should improve the bootstrap inferences as well. Overall increase in signal-to-noise ratio (precision) will lead to less variable bootstrap slopes and, in consequence, to more narrow CIs (in general bootstrap techniques are more accurate when more data is available, Rousselet et al., 2019).

Trial timings. Very recently, Brigadoi, Basso Moro, Falchi, Cutini, & Dell'Acqua (2018) found that increasing the length of intertrial intervals in the parity task can lead to an increase of SNARC effect reliability, perhaps because participants are more alert before each response due to longer waiting time (see e.g., Vallesi, Shallice, & Walsh, 2006). However, evidence for that comes from only one study so far, and it needs to be explored further. Unfortunately, we could not verify these findings in our data sets as timings in all studies we considered were relatively similar: they varied from 0ms to 1300ms. The increase in reliability observed in Brigadoi et al. (2018) was observed between the condition with intertrial interval 1110 -1500ms and the condition with intertrial interval 6000-10000ms.

To sum up, there are numerous means to improve reliability and measurement precision. If one wishes to use the methods described above to investigate other cognitive phenomena, it can be also worthwhile to check for optimal task parameters which influence the stability and robustness of the effect in question. As already discussed, reliability is not only crucial when investigating the prevalence of cognitive effects at the individual level, but also when studying correlations with other cognitive phenomena. Because the correlations between the SNARC effect and other cognitive variables are not the main focus of this paper, we discuss them in more detail in Appendix E.

Implications and future directions for numerical cognition research

Spatial-numerical associations: universal mechanism or one of many strategies.

In the domain of numerical cognition, it is widely accepted that numbers are mapped onto space. The SNARC effect is one of the hallmark observations that is typically used to support this point (e.g., Hubbard, Piazza, Pinel, & Dehaene, 2005). If mapping numbers onto space is a universal way that people use to represent and process numerical information, then this spatial mapping should be highly prevalent in the population and should consequently appear reliably in all participants in virtually all numerical tasks. The finding that only about 35% to 45% of participants show a reliable SNARC effect raises several fundamental questions: e.g., Is the SNARC effect a good index of this mapping, and is the link between numbers and space as universal as originally proposed? Whereas reasonable reliability is observed in the majority of our reanalyzed samples, additional empirical work is needed to solve the first question. For this purpose, studies using other number-space tasks (e.g., number interval bisection, random digit generation) should be (re-)analyzed using the same approach to see whether higher proportions of individuals with reliable effects can be found. Additionally, further insights in the usefulness of the SNARC effect as an index of the mapping from numbers to space can be obtained by investigating whether the SNARC effect is stable over time within the same individual (i.e., to determine the test-retest reliability).

Alternatively, the SNARC effect might be a good proxy of the mapping of numbers to space. In that case, the robust group-level effects are in fact driven by a spatial mapping, which is only reliably present in less than half of the participants. This would imply that this mapping is not universal. In that event, theories of numerical cognition having an oriented spatial representation as the key medium of our capacity to mentally represent numerical magnitude need to be reconsidered: at the population level, a left-to-right spatial mapping might only be one of several different possible ways to represent numerical magnitude. This idea fits nicely with recent ideas suggesting that the SNARC effect is constructed online during task execution (in working memory) as a function of the specific set of stimuli used in the task (e.g. Abrahamse, van Dijck & Fias, 2018; Fias & van Dijck, 2016). Importantly however, this account should be extended to be able to account for the high individual variability in the way the task sets are mentally constructed and organized.

New analytical directions - where to look for correlations.

Having a method to determine the presence of an effect at the individual level opens several new venues for more detailed analyses of (existing) datasets. For example, besides the overall reaction times and standard deviations, the width of an individual CI can be considered as the precision with which the effect is measured. As such this width can be considered as a way to determine whether the measurement of an effect was sufficiently precise or not (in case the width of an individual CI is very deviant from the average width inf the sample), and can thus be used as a tool to identify potentially outlying subjects. This way of data-trimming can be very interesting in the context of correlational research. After all, an imprecise measurement (which is no (bivariate) outlier in absolute size) is unlikely to have a positive influence on the correlation under investigation.

Besides being a tool for data-trimming, the method of identifying group differences is also interesting in itself. Besides comparing groups in terms of the size of an effect, it is also possible to determine whether groups differ in terms of the proportion of individuals showing the effect or not. Consequently, we could distinguish new ways for investigating individual differences in this field. For example, we can try to look at factors related to whether an individual reveals a reliable effect or not. Second, among individuals who reliably reveal it, we could investigate factors related to the strength of such an effect. As there might be qualitative differences between individuals who reveal an effect or not. Looking at whether effects are reliable at the individual level could also provide insights in populations which do not reveal an effect. After all, in theory it is possible that observed null effects are due to two subgroups showing reliable but opposing effects (which cancel each other out when averaging the data).

Chasing cognitive phenomena

Researchers in experimental psychology only rarely question the reliability of well-established group-level cognitive phenomena at the individual level. This dates back to Wilhelm Wundt, who treated individual differences as irrelevant. The presence, robustness, and especially replicability of a certain phenomenon are usually considered as sufficient to infer universal principles of the mind. Nevertheless, even in the history of this specific domain of research, examples suggest that such an approach can be misleading. As initially demonstrated by Navon (1977), there is a bias towards global processing of perceptual information. When participants are presented with stimuli such as a letter H composed of small letters S, they more efficiently recognize the meaning of the big stimuli, not the meaning of elements that are its components (i.e., they “see the forest before seeing the trees”; Happé & Frith, 2006). Despite the main effect being robust and highly replicable (see e.g., Hedge et al., 2018), further investigations show that the size (and even presence) of this general bias can differ considerably between individuals. For instance, (1) it strongly correlates with the cognitive style of field dependence (Poirel, Pineau, Jobard, & Mellet, 2008); (2) it can be related to autism or personality features related to ASD (Van der Hallen, Evers, Brewaeys, Van den Noortgate, & Wagemans, 2015 for a meta-analysis); (3) it is related to cultural differences (e.g., McKone et al., 2010); (4) some individuals are characterized by local biases (see e.g., Staudinger, Fink, Mackay, & Lux, 2011).

Our analyses present even more dramatic example – the effect seems to be driven not by the majority but rather the minority of individuals. In this context, the recent distinction between dominant and indominant psychological phenomena introduced by Rouder and Haaf (2018) seems to be useful. Phenomena characterized by dominance should be present in all individuals, and we should not expect that there are individuals showing the reverse effect. For instance, one can hardly expect that some individuals reveal a reverse Stroop effect. On the other hand, in phenomena characterized by indominance, the effect observed at the group level is driven by some individuals. Except from resolving power considerations as discussed by Rouder and Haaf (2018), the dominance-indominance distinction has a major theoretical implications for scientific psychology. We believe that the approach we proposed can help distinguishing between dominant and indominant phenomena. Importantly, the logic and methodological approach can be easily adapted to any cognitive phenomenon, which is investigated in multiple labs with a similar paradigm, and there are multiple existing datasets, which can be uniformly analyzed. Analysis scripts we share can be easily adapted to investigating most of these phenomena.

CONCLUSION

The present work shows that a uniform analysis of existing data sets can bring new insights into the nature of well-established cognitive phenomena. This was achieved by applying psychometric and bootstrapping approaches. In this study, we focused on the prevalence of the SNARC effect. More generally, our work emphasizes the idea that robust cognitive effects within the population might not always be observed in every individual. To better understand the processes underlying such effects, one therefore needs to quantify them at the individual level. This can be achieved by calculating CIs around the true score parameter using both psychometric and bootstrapping approaches.

Even in the case of a robust and highly-replicable phenomenon at the group level, the effect can be reliably present in a minority of individuals only. Such an observation leads to critical questions about the nature of the phenomenon, and whether, while being reliably present in fewer than 50% of individuals, it can be considered as evidence for a general principle of (numerical) cognition. In line with these results, as regards estimating the prevalence of the phenomena, we recommend using the bootstrap techniques to check how stable the effect is (H1 bootstrapping), and how unlikely the observed effect is given the null hypothesis is true (H0 bootstrapping).

The method of uniform analysis of multiple existing data sets can answer several questions about the phenomenon being investigated. Furthermore, a similar logic and methods for investigating the presence of group-level effects at the individual level can be applied to other cognitive phenomena, and potentially stimulate insights in other domains of cognitive psychology and bind together the two branches of psychology, which in the past few decades developed largely in parallel.

CONTEXT

On the methodological level, this paper demonstrates how psychometric and bootstrapping methods can be used to better understand supposedly well-known cognitive phenomena. It can also serve as an example of how existing data sets can be unified, re-used, and shared in a large-scale collaborative effort aimed at addressing a particular question and leading to new insights, especially at the intersection between experimental cognitive and differential psychology. Such an approach is complementary to traditional meta-analyses and multi-lab initiatives. At the content level, our approach can be treated as a first step towards broader investigations of the nature of spatial- numerical associations and their role within numerical cognition: do they reflect a universal mechanism by which humans represent numerical information or are they only one of several possible strategies of representing numbers? This paper is a follow-up on ongoing work by most co-authors aimed at investigating individual differences in the SNARC effect. Originally, the idea of this collaborative effort emerged in discussions of our endeavors in this domain during the European Workshop on Cognitive Neuropsychology (Bressanone, Italy, January 2017). During the development of these ideas, several other co-authors were invited to contribute.

ACKNOWLEDGEMENTS

KC is supported by a DFG grant [NU 265/3-1] to HCN. KC and HCN 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. MP is a Senior Research Associate at the Fonds National de la Recherche Scientifique (FRS-FNRS, Belgium). NM is supported by grant PDR-T.0047.18. INTER/FNRS/17/11781524/SNAMATH from the Fonds National de la Recherche Scientifique (FRS-FNRS, Belgium) and the Fonds National de la Recherche (FNR, Luxembourg) to MP and CS. The collection of data contributed by SG was supported by a British Academy/Leverhulme Small Research Grant [SG121544]. We would like to thank [Name to be added after actual proof-reading] for proofreading the manuscript and Wim Gevers for sharing the unpublished data-set collected by Ginsburg et al.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2019). Power contours: optimising sample size and precision in experimental psychology and human neuroscience, 1–19. Retrieved from http://arxiv.org/abs/1902.06122

Brigadoi, S., Basso Moro, S., Falchi, R., Cutini, S., & Dell'Acqua, R. (2018). On pacing trials while scanning brain hemodynamics: The case of the SNARC effect. *Psychonomic Bulletin & Review*, 1–7. https://doi.org/10.3758/s13423-017-1418-1

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.72

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Cipora, K. (2014). Between task consistency of the SNARC effect. In *Poster presented on XXXIInd European Workshop on Cognitive Neuropsychology*. Bressanone: Italy.

Cipora, K., Czernecka, K., & Szymura, B. (2009). Temperamental differences in the magnitude of the SNARC effect. In *Poster presented on International Society for the Study of Individual Differences*. Evanson, US.

Cipora, K., & Göbel, S. M. (2013). Number – Space associations: Just how reliable is the SNARC effect. In *Poster presented on XXXIst European Workshop on Cognitive Neuropsychology.* Bressanone: Italy.

Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nęcka, E. (2016). Professional mathematicians differ from controls in their spatial-numerical associations. *Psychological Research*, *80*(4).

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*(10), 1974–1991. https://doi.org/10.1080/17470218.2013.772215

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*, 1–17. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20*20 monte carlo investigation. *Frontiers in Psychology*, *8*(JUL). 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

Crollen, V., & Noël, M. P. (2015). Spatial and numerical processing in children with high and low visuospatial abilities. *Journal of Experimental Child Psychology*, *132*, 84–98. https://doi.org/10.1016/j.jecp.2014.12.006

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

de Hevia, M. D., Veggiotti, L., Streri, A., & Bonn, C. D. (2017). At Birth, Humans Associate “Few” with Left and “Many” with Right. *Current Biology*, *27*(24), 3879–3884.e2. https://doi.org/10.1016/j.cub.2017.11.024

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396.

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., Lauwereyns, J., & Lammertyn, J. (2001). Irrelevant digits affect feature-based attention depending on the overlap of neural circuits. *Cognitive Brain Research*, *12*(3), 415–423. https://doi.org/10.1016/S0926-6410(01)00078-7

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

Fischer, M. H. (2008). Finger counting habits modulate spatial-numerical associations. *Cortex*, *44*(4), 386–392. https://doi.org/10.1016/j.cortex.2007.08.004

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

Georges, C., Hoffmann, D., & Schiltz, C. (2016). How math anxiety relates to number-space associations. *Frontiers in Psychology*, *7*(SEP), 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*(2), 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., 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

Göbel, S. M., Maier, C. A., & Shaki, S. (2015). Which numbers do you have in mind? Number generation is influenced by reading direction. *Cognitive Processing*, *16*(S1), 241–244. https://doi.org/10.1007/s10339-015-0715-8

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, *69*, 14–25. https://doi.org/10.1016/J.CONCOG.2019.01.008

Happé, F., & Frith, U. (2006). The Weak Coherence Account: Detail-focused Cognitive Style in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5–25. https://doi.org/10.1007/s10803-005-0039-0

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

Hoffmann, D., Mussolin, C., Martin, R., & Schiltz, C. (2014). The impact of mathematical proficiency on the number-space association. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0085048

Hoffmann, D., Pigat, D., & Schiltz, C. (2014). The impact of inhibition capacities and age on number-space associations. *Cognitive Processing*, *15*(3), 329–342. https://doi.org/10.1007/s10339-014-0601-9

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, *6*(6), 435–448. https://doi.org/10.1038/nrn1684

Jonas, C. N., Spiller, M. J., Jansari, A., & Ward, J. (2014). Comparing Implicit and Synaesthetic Number–Space Associations: Visuospatial and Verbal Spatial–Numerical Associations of Response Codes. *Quarterly Journal of Experimental Psychology*, *67*(7), 1262–1273. https://doi.org/10.1080/17470218.2013.856928

LeBel, E. P. (2015). A New Replication Norm for Psychology. *Collabra*, *1*(1), 1–13. https://doi.org/10.1525/collabra.23

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

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

Lyons, I. M., Nuerk, H. C., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General*, *144*(5), 1021–1035. https://doi.org/10.1037/xge0000094

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*(2), 154–161. https://doi.org/10.1016/j.actpsy.2010.01.006

McKone, E., Aimola Davies, A., Fernando, D., Aalders, R., Leung, H., Wickramariyaratne, T., & Platow, M. J. (2010). Asia has the global advantage: Race and visual attention. *Vision Research*, *50*(16), 1540–1549. https://doi.org/10.1016/J.VISRES.2010.05.010

Mead, A. D. (2005). Reliability: definintions and estimation. In S. Everitt, Brian & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behaviotal Science* (pp. 1733–1735). Chichester: John Wiley & Sons.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

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., 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

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science. *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pinhas, M., Shaki, S., & Fischer, M. H. (2014). Heed the Signs: Operation Signs have Spatial Associations. *Quarterly Journal of Experimental Psychology*, *67*(8), 1527–1540. https://doi.org/10.1080/17470218.2014.892516

Pinhas, M., Tzelgov, J., & Ganor-Stern, D. (2012). Estimating linear effects in ANOVA designs: The easy way. *Behavior Research Methods*, *44*(3), 788–794. https://doi.org/10.3758/s13428-011-0172-y

Poirel, N., Pineau, A., Jobard, G., & Mellet, E. (2008). Seeing the Forest Before the Trees Depends on Individual Field-Dependency Characteristics. *Experimental Psychology*, *55*(5), 328–333. https://doi.org/10.1027/1618-3169.55.5.328

Rouder, J. N., & Haaf, J. M. (2018). Power, Dominance, and Constraint: A Note on the

Appeal of Different Design Traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rousselet, G., Pernet, D. C., & Wilcox, R. R. (2019). A practical introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations. https://doi.org/10.31234/OSF.IO/H8FT7

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

Sauce, B., & Matzel, L. D. (2013). The causes of variation in learning and behavior: why individual differences matter. *Frontiers in Psychology*, *4*, 395. https://doi.org/10.3389/fpsyg.2013.00395

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., & Fischer, M. H. (2008). Reading space into numbers - a cross-linguistic comparison of the SNARC effect. *Cognition*, *108*(2), 590–599. https://doi.org/10.1016/j.cognition.2008.04.001

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin and Review*, *16*(2), 328–331. https://doi.org/10.3758/PBR.16.2.328

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Staudinger, M. R., Fink, G. R., Mackay, C. E., & Lux, S. (2011). Gestalt perception and the decline of global precedence in older subjects. *Cortex*, *47*(7), 854–862. https://doi.org/10.1016/J.CORTEX.2010.08.001

The Cochrane Collaboration. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. (J. P. Higgins & S. Green, Eds.) (Version 5.). The Cochrane Collaboration. Retrieved from www.handbook.cochrane.org

The jamovi project. (2019). jamovi. Retrieved from https://www.jamovi.org

Thompson-Schill, S. L., Braver, T. S., & Jonides, J. (2005). INDIVIDUAL DIFFERENCES: Editorial. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(2), 115–116. https://doi.org/10.3758/CABN.5.2.115

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

Tzelgov, J., Zohar-Shai, B., & Nuerk, H.-C. (2013). On defining quantifying and measuring the SNARC effect. *Frontiers in Psychology*, *4*(MAY), 3–5. https://doi.org/10.3389/fpsyg.2013.00302

Vallesi, A., Shallice, T., & Walsh, V. (2006). Role of the Prefrontal Cortex in the Foreperiod Effect: TMS Evidence for Dual Mechanisms in Temporal Preparation. *Cerebral Cortex*, *17*(2), 466–474. https://doi.org/10.1093/cercor/bhj163

Van der Hallen, R., Evers, K., Brewaeys, K., Van den Noortgate, W., & Wagemans, J. (2015). Global processing takes time: A meta-analysis on local-global visual processing in ASD. *Psychological Bulletin*, *141*(3), 549–573. https://doi.org/10.1037/bul0000004

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., & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition*, *119*(1), 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

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

Vogel, E. K., & Awh, E. (2008). How to Exploit Diversity for Scientific Gain. *Current Directions in Psychological Science*, *17*(2), 171–176. https://doi.org/10.1111/j.1467-8721.2008.00569.x

Willmes, K. (2010). The methodological and statistical foundations of neuropsychological assessment. In J. Gurd, U. Kischka, & J. Marschall (Eds.), *The Handbook of Clinical Neuropsychology* (Second ed., pp. 28–49). New York: Oxford University Press.

Wood, G., Nuerk, H. C., & Willmes, K. (2006). Variability of the SNARC effect: Systematic interindividual differences or just random error? *Cortex*, *42*(8), 1119–1123. 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*(4), 489–525. https://doi.org/10.1027/1618-3169.52.3.187

Yang, T., Chen, C., Zhou, X., Xu, J., Dong, Q., & Chen, C. (2014). Development of spatial representation of numbers: A study of the SNARC effect in Chinese children. *Journal of Experimental Child Psychology*, *117*, 1–11. https://doi.org/10.1016/J.JECP.2013.08.011