
Article:
Bahnmueller, Julia, Cipora, Krzysztof, Göbel, Silke M. orcid.org/0000-0001-8845-6026 et al. (2 more authors) (2021) Pick the smaller number:No influence of linguistic markedness on three-digit number processing. Journal of Numerical Cognition. pp. 295-307. ISSN 2363-8761

https://doi.org/10.5964/jnc.6057

Pick the smaller number: No influence of linguistic markedness on three-digit number processing

Julia Bahnmueller[1], Krzysztof Cipora[1], Silke M. Göbel[2,3], Hans-Christoph Nuerk[4,5] & Mojtaba Soltanlou[5,6]

[1]Loughborough University, UK
[2] University of York, UK
[3] University of Oslo, Norway
[4] LEAD Graduate School & Research Network, University of Tuebingen, Germany
[5] University of Tuebingen, Germany
[6] Western University, Canada

Author Note

Julia Bahnmueller, Centre for Mathematical Cognition, Loughborough University (LE11 3TU), Loughborough, United Kingdom; Krzysztof Cipora, Centre for Mathematical Cognition, Loughborough University (LE11 3TU), Loughborough, United Kingdom; Silke M. Göbel, Department of Psychology, University of York (YO10 5DD), United Kingdom; Hans-Christoph Nuerk, Department of Psychology, University of Tuebingen (Schleichstraße 4, 72076 Tübingen), Germany; Mojtaba Soltanlou, The Brain and Mind Institute and Department of Psychology, Western University (London, ON N6A 5B7), Canada.

Correspondence concerning this article should be addressed to Julia Bahnmueller, Centre for Mathematical Cognition, Loughborough University (Schofield building (SCH.0.30), University road, Loughborough, LE11 3TU, United Kingdom). Email: j.bahnmuller@lboro.ac.uk.

Abstract

The symbolic number comparison task has been widely used to investigate the cognitive representation and underlying processes of multi-digit number processing. The standard procedure to establish numerical distance and compatibility effects in such number comparison paradigms usually entails asking participants to indicate the *larger* of two presented multi-digit Arabic numbers rather than to indicate the *smaller* number. In terms of linguistic markedness, this procedure includes the unmarked/base form in the task instruction (i.e., large). Here we evaluate distance and compatibility effects in a three-digit number comparison task observed in Bahnmueller et al. (2015) using a marked task instruction (i.e., 'pick the *smaller* number'). Moreover, we aimed at clarifying whether the markedness of task instruction influences common numerical effects and especially componential processing as indexed by compatibility effects. We instructed German- and English-speaking adults (N=52) to indicate the smaller number in a three-digit number comparison task as opposed to indicating the larger number in Bahnmueller et al. (2015). We replicated standard effects of distance and compatibility in the new pick the smaller number experiment. Moreover, when comparing our findings to Bahnmueller et al. (2015), numerical effects did not differ significantly between the two studies as indicated by both frequentist and Bayesian analysis. Taken together our data suggest that distance and compatibility effects alongside componential processing of multi-digit numbers are rather robust against variations of linguistic markedness of task instructions.

Keywords: linguistic markedness; distance effect; compatibility effects; componential processing; three-digit numbers; number comparison

Word count abstract: 229 Word count text body: 5400

The symbolic number magnitude comparison task is often used to investigate the cognitive processes of (multi-digit) number processing. In this task, participants are usually asked to indicate the larger out of two numbers. Two stable hallmark effects observed when comparing numbers are the numerical distance (Moyer & Landauer, 1967; see also e.g., Hohol et al., 2020) and the (unit-decade) compatibility effects (Nuerk, Weger, &Willmes, 2001; see Nuerk, Moeller, & Willmes, 2015 for an overview of further numerical effects in multi-digit number processing).

Numerical distance and compatibility effects

The distance effect reflects the finding that performance in number comparison tasks increases with larger distance between numbers. Thereby, the distance effect gave rise to the widely held thought that numbers are represented and processed analogically along the so-called mental number line (e.g., Dehaene & Changeux, 1993; Gallistel & Gelman, 1992; Restle, 1970). The fact that the distance effect was observed for the overall distance when comparing multi-digit numbers (Hinrichs, Yurko, & Hu, 1981) led to the assumption of a simple elongation of the mental number line from the single- into the two-digit number range (e.g., Brysbaert, 1995; Dehaene, Dupoux, & Mehler, 1990) proposing analogue, holistic processing also for multi-digit numbers. However, further research showed that next to the overall distance between numbers, distances between the single corresponding digits (i.e., hundreds, tens, units) influence numerical processing as well (e.g., Nuerk et a., 2001; Verguts & De Moor, 2005). These results favour an alternative account suggesting that rather than being processed purely holistically, the single digits of multi-digit numbers are processed componentially (e.g., hundreds, tens, units etc. are processed separately; see Huber, Nuerk, Willmes, & Moeller, 2016 for a comprehensive computational modelling approach).

Strong support for the componential processing account was further provided by the unit-decade compatibility effect (Nuerk et al., 2001). The unit-decade compatibility effect reflects performance differences between unit-decade compatible number pairs (i.e., the comparison of both tens and units leads to the same decision: 32_57, 3 < 5 and 2 < 7) and unit-decade incompatible number pairs (i.e., the comparison of tens and units leads to opposing decisions: 37_62, 3 < 6 but 7 > 2). When overall distance is held constant between compatible and incompatible number pairs, compatible number pairs are usually responded to faster and with fewer errors than incompatible ones (e.g., Nuerk et al., 2001; see also Huber et al., for a large-scale online investigation). Moreover, compatibility effects were also observed for three-digit numbers (Bahnmueller et al., 2015, 2016; Korvorst & Damian, 2008; Mann, Moeller, Pixner, Kaufmann, & Nuerk, 2012; see also Huber, Moeller, Nuerk, & Willmes, 2013 for simulated data; see also Meyerhoff, Moeller, Debus, & Nuerk, 2012 for compatibility effects in four- and six-digit numbers). Following the same logic as for two-digit numbers, hundred-decade and hundred-unit compatibility effects can be defined for three-digit numbers (e.g., the number pair 327_465 is hundred-decade compatible because 3 < 4 and 2 < 6, but it is hundred-unit incompatible because 3 < 4 but 7 > 5). In sum, compatibility effects indicate that the magnitudes of the decision-irrelevant digits (i.e., units in two-digit number comparison, tens and units in three-digit number comparison) interfere with the comparison process suggesting that the magnitudes of the single constituting digits of a number are processed componentially.

Both distance and compatibility effects were typically investigated with the magnitude comparison paradigm in which participants were asked to indicate the larger of two numbers. Following from this, one may ask whether the observed effects are generic to multi-digit number processing or whether they also, at least partly, originate from the specific task setup of selecting the larger number. The natural alternative to this task is a setup in which

participants are asked to indicate the smaller of the two numbers presented. Even though it may seem that these setups don't differ much, empirical evidence evaluating effects of task instruction (e.g., picking the larger vs picking the smaller number) is surprisingly limited. In this context, the concept of linguistic markedness might be of particular interest.

Linguistic markedness

Linguistic markedness refers to the fact that most adjective pairs have an unmarked/base form and a marked/derived form. Examples for unmarked/base adjectives are "old", "even", "right", "large" or "friendly" and their respective marked/derived counterparts are "young", "odd", "left", "small" and "unfriendly". Thereby, marked adjectives can, for instance, be constructed by adding a prefix or suffix to the unmarked form (e.g., un-friendly; formal markedness) and/or can represent the adjective form that is used less frequently (e.g., "How young are you?", "How small are you?"; distributive markedness, cf. Lyons, 1968). In this context, previous studies, for instance, indicate that marked adjectives decrease performance in sentence comprehension (e.g., Sherman, 1973, 1976). Another example can be found in the study by Hines (1990) who observed slower reactions to numbers that have to be classified as odd compared to numbers that have to be classified as even (showing a so-called "odd effect"; see also Nuerk, Iversen, & Willmes, 2004). Following from the linguistic markedness account, the default (unmarked) pick larger setup might differ from the marked pick smaller setup resulting in differences in general task performance (i.e., longer reaction times in the pick smaller setup) as well as in observed numerical effects.

Linguistic markedness and numerical effects

Up to now, only few studies investigated modulations of numerical effects resulting from manipulations of unmarked vs. marked task instructions. For instance, Verguts and De

Moor (2005) manipulated linguistic markedness of task instruction (pick the smaller vs. pick the larger number) when investigating the distance effect in a two-digit number comparison task. They found an overall distance effect for within-decade number pairs (e.g., 64_68) but not for between-decade number pairs for which decade distance was held constant (decade distance was always 1; e.g., 68_72) for both the marked and the unmarked task instructions (see Moeller, Klein, & Nuerk, 2013, for a discussion of the differential results regarding distance effects). Crucially, although there was no formal statistical comparison, descriptively overall response times in the pick smaller condition were about 60 ms slower than in the pick larger condition (see Figure 1 in Verguts & De Moor, 2005). Thus, this study seems to show an effect of linguistic markedness on overall reaction times, however, no evidence was provided indicating a modulating effect of linguistic markedness on the numerical distance effect.

Contrarily, Arend and Henik (2015) demonstrated that the linguistic markedness of the task instruction modulates the size congruity effect (SiCE). The SiCE refers to the finding that in numerical and physical comparison tasks, response times are longer when number magnitude and physical size are congruent (e.g., 2 4) than when they are incongruent (e.g., 2 4; Henik & Tzelgov, 1982). In their study, reaction times were longer in the pick smaller condition compared to the pick larger condition. Moreover, the SiCE was larger when participants were instructed to pick the larger as compared to when they were instructed to pick the smaller number in the number magnitude comparison task, but no difference was found in the physical comparison task.

Further studies show that the linguistic markedness of task instruction also affects other types of Spatial-Numerical Associations (SNAs; see e.g., Cipora, Soltanlou, Schroeder, & Nuerk, 2018). Patro and Haman (2012) found an effect of SNA congruency (i.e., faster reactions to larger numerosities on the right) only in the pick larger but not in the pick smaller

condition (i.e., reactions to smaller numerosities did not differ between left and right; c.f. Figure 2 in Patro & Haman, 2012). Type of instruction also affects comparative judgments of conceptual size of objects, but not Arabic numbers (Shaki, Petrusic, & Leth-Steensen, 2012).

To sum up, the evidence for the modulating role of linguistic markedness of task instruction on numerical effects remains inconsistent. One potential mechanism by which linguistic markedness of task instruction might affect specific numerical effects may be due to its influence on overall reaction times. For instance, the spatial-numerical association of response codes effect (SNARC effect; Dehaene, Bossini, & Giraux, 1993) was shown to increase with longer overall reaction times (Cipora, Soltanlou, Reips, & Nuerk, 2019; see Gevers, Verguts, Reynvoet, Caessens, & Fias, 2006; see Cipora et al., 2016 for a discussion of potential measurement artifacts in this context). Other cognitive effects, such as the Simon effect seem to also vary with general reaction time (Mapelli, Rusconi, Umiltà, 2003; see also Glaser & Glaser, 1982 for the Stroop effect).

With respect to the effects of interest in the present study, the distance effect was shown to be more pronounced for longer reaction times (Hohol et al. 2020). However, to the best of our knowledge, associations of overall response times and compatibility effects have not been reported yet. Nonetheless, in developmental studies overall reaction times were standardized to control for potential effects of interindividual variability in reaction times on the size of compatibility effects (Mann, Moeller, Pixner, Kaufmann, & Nuerk, 2012; Nuerk et al., 2004; Pixner, Moeller, Heřmanová, Nuerk, & Kaufmann, 2011). The reasoning behind the standardization is that prolonged processing of a stimulus might lead to increased interference of task irrelevant digits (i.e., unit digit in two-digit number pairs, unit and tens digit in three-digit number pairs) in incompatible number pairs and, thereby, to larger compatibility effects.

The present study

The current study set out to evaluate the generality of basic effects in multi-digit number processing (i.e., distance and compatibility effects) across marked and unmarked task instructions (i.e., pick the larger vs. pick the smaller number). In particular, in a conceptual replication attempt of the study by Bahnmueller et al. (2015), we employed the same three-digit number comparison paradigm with Arabic digits in a comparable sample of German- and English-speaking adults. However, instead of asking participants to indicate the *larger* of two presented three-digit numbers we asked participants to indicate the *smaller* of two three-digit numbers.

As it seems unlikely that a change in linguistic markedness of task instructions leads to major disruptions of the main underlying cognitive mechanisms of multi-digit Arabic number processing (i.e., number magnitude should still be processed, number should still be processed componentially), we predicted reliable main effects of hundred distance, hundred-decade compatibility, and hundred-unit compatibility when participants are asked to pick the smaller number.

To investigate potential modulating effects of linguistic markedness more directly, we compared overall reaction times as well as the respective numerical effects directly between the newly conducted pick smaller and the pick larger experiment in Bahnmueller et al. (2015). In line with previous reports (Arend & Henik, 2015; Verguts & De Moor, 2005), we expected prolonged reaction times when instructed to pick the smaller as compared to picking the larger number.

Regarding modulations of the numerical effects due to linguistic markedness of the task instruction, we expected to replicate the findings by Verguts and De Moor (2005) showing comparable distance effects for marked and unmarked task instructions. However, regarding the hundred-decade and the hundred-unit compatibility, we expected to find larger

compatibility effects when instructed to pick the smaller number because longer overall reaction times and, thus, prolonged processing of number pairs in the pick smaller experiment should lead to increased interference of task irrelevant digits (i.e., unit and tens digit) in incompatible number pairs and, thereby, to larger compatibility effects.

Methods

Participants

For the analyses of the *pick smaller experiment*, newly collected data of a total of 53 participants were considered (after exclusions, see below). Based on Bahnmueller et al. (2015; henceforth referring to the *pick larger experiment*), we did not expect three-digit number processing to be influenced by the number word structure (e.g., inverted vs. non-inverted number words; but see, e.g., Steiner et al., this issue, for inversion-related effects when processing multi-digit numbers in children). However, we recruited a comparable sample of German- and English-speaking participants for the *pick smaller experiment*. This allowed for optimal comparability between studies and further exploration of potential language-related modulations within the present *pick smaller experiment*.

Three participants were excluded in the *pick smaller experiment* because error rates exceeded 10% in the experimental trials. Moreover, another four participants were excluded because they consistently used the reverse response coding (i.e., they picked the larger number). Thus, the final *pick smaller* sample consisted of 30 native German speakers (24 female, all right handed, M_{age} = 22.7 years, SD = 2.8) and 23 native English speakers (16 female, all right handed, M_{age} = 19.7 years, SD =1.4).

For the re-analyses of the *pick larger experiment*, data of a total of 51 participants were considered. Two participants were excluded because error rates exceeded 10%. Thus, the final *pick larger* sample consisted of 24 native German speakers (21 female, 20 right

handed, M_{age} = 23.1 years, *SD* = 6.3) and 27 native English speakers (21 female, 25 right handed, M_{age} = 20.1 years, *SD* = 2.3).

German-speaking participants were recruited via postings at the University of Tuebingen and the Leibniz-Institut für Wissensmedien Tübingen. Participants received course credit or 5€/4£ for compensation. The study was approved by the local ethics committee of the University of York.

Power calculations. Sample size estimates for paired t-tests for the pick smaller experiment were calculated using JAMOVI (The jamovi project, 2020) and were based on the respective effect sizes observed in the *pick larger experiment*. Based on this, a sample size of 27 should be sufficient to detect a hundred-decade compatibility effect (i.e., the smallest main effect observed in the *pick larger experiment*) of an effect size of d = .59 or larger with α = .05 (one-tailed) and a power of .90. To achieve comparability between the *pick smaller* and the *pick larger experiment* and to increase sensitivity for detecting a smaller effect in the *pick smaller experiment*, we aimed at collecting a comparable number of participants (N = 51) allowing us to detect a medium sized effect of d = .46.

G*Power (Faul et al., 2009) was used for power estimates of the between-subject effect of instruction (*pick smaller* vs. *pick larger*) as well as the within-between interaction of the respective numerical effect and instruction in the 2 × 2 mixed factor ANOVAs. A total sample size of 100 is sufficient to detect a medium sized between-subject as well as interaction effect of f = .33 (η_p^2 = .1) with α = .05 and a power of .90 (see https://osf.io/27jty/ for all outputs of the power calculations).

Stimuli

The same stimulus set was used in the *pick smaller* and the *pick larger experiment*. In total, 640 three-digit number pairs were used. Of these, 320 were experimental items

manipulated orthogonally according to hundred-, decade-, and unit distance (small (1-3) vs. large (4-8)), as well as hundred-decade and hundred-unit compatibility (compatible vs. incompatible). Moreover, problem size was matched across all item categories and decade as well as unit distance was matched for the respective item categories. In addition to the 320 experimental items, 320 filler items were included in the stimulus set to avoid that participants focused only on the decision-relevant hundred-digit (160 within-hundred filler items, e.g., 672_648; 160 within-hundred-within-decade filler items, e.g., 282_284). Please refer to the supplementary material in Bahnmueller et al. (2015) for a more detailed description of the stimulus set as well as descriptive characteristics of all item categories.

Unfortunately, due to a programming error in the *pick smaller experiment*, participants were only presented with 560 of the 640 items (i.e., the last block (80 items) was not presented). The 560 items were randomly drawn from the total item set for each participant. Regarding the 320 experimental stimuli included in the analyses, an item was presented 46.4 times on average (*SD* = 2.4, range: 40-52). Because items were drawn randomly, stimulus matching was not substantially affected (see https://osf.io/27jty/ for item characteristics of the experimental items in the *pick smaller experiment* compared to item characteristics of the matched stimulus set).

Procedure

The procedure of both experiments was identical and differed only with respect to the task instruction. In particular, participants were instructed to indicate the smaller (*pick smaller experiment*) or the larger (*pick larger experiment*) of two simultaneously presented three-digit numbers as fast and as accurately as possible. Numbers were presented above each other. In the *pick smaller experiment*, participants were asked to press the upward arrow of a standard keyboard in case the upper number was smaller, and they were asked to press the

downward arrow in case the lower number was the smaller one. In contrast, in the *pick larger experiment*, participants had to indicate the location of the larger number by pressing the upward arrow in case the upper number was larger, and the downward arrow in case the lower number was larger.

The respective experiment started with 10 practice trials, followed by 8 blocks (7 blocks in the *pick smaller study*) containing 80 items each. After each block, the participant could take a short break. Stimulus order was randomized separately for each participant and across blocks. Stimuli were presented centrally in white against a black background (font: Arial, font size: 24, bold). A trial started with a fixation cross presented centrally for 500ms. Following the fixation cross, a number pair was presented and remained on the screen until a response was given. The next trial started after an inter-trial-interval of 500ms.

Results

Analyses

Analyses were performed using R (R Core Team, 2020) and RStudio (RStudio Team, 2020) as well as JASP for Bayesian analyses (JASP Team, 2020). For the interpretation of Bayes factors, we use the classification adopted in JASP (van Doorn et al., 2019) differentiating strong ($BF_{01} < 1/10$) and moderate evidence against H_0 ($1/10 < BF_{01} < 1/3$), weak/inconclusive evidence ($1/3 < BF_{01} < 3$) as well as moderate ($3 < BF_{10} < 10$) and strong evidence for H_1 ($BF_{10} > 10$). Data, analysis script and JASP output files illustrating Bayesian analyses with all the parameters used can be found at https://osf.io/27jty/.

As error rates were very low (*pick smaller experiment*: $M = 4.3\%$, $SD = 2.0\%$; *pick larger experiment*: $M = 3.7\%$, $SD = 2.1\%$) analyses focused on reaction times (RT). Practice trials and filler items were excluded from the analyses. Moreover, RTs faster than 200ms as

well as RTs deviating more than +/- 3SD from an individual participant's mean RT were excluded. This trimming procedure resulted in a loss of 1.4% of data.

Directly addressing our primary research question, we first report results of the analyses of numerical effects in the new *pick smaller experiment* using three paired *t*-tests[1] (i.e., one per numerical effect; effect sizes (Cohen's d for paired *t*-tests) along with 95% confidence intervals were estimated as implemented in JASP). Moreover, a 2 × 2 × 2 × 2 mixed design ANOVA similar to the one reported by Bahnmueller et al. (2015) discerning the within-subject factors hundred distance, hundred-decade compatibility, and hundred-unit compatibility, as well as the between-subject factor language group (German vs. English) will also be reported for the *pick smaller experiment*.

Analyses of the *pick smaller experiment* are directly followed by the re-analysis of the results of the *pick larger experiment* using the same, more focused analyses (i.e., one paired t-test per numerical effect). Afterwards, results of the direct comparison of the two experiments are reported separately for mean reaction times and each numerical effect using both frequentist as well as Bayesian measures to be able to quantify the evidence for both the null and the alternative hypothesis.

Pick smaller experiment

Results of *t*-tests indicated a regular hundred distance effect with faster RTs for number pairs with a large (M = 694ms, SD = 123ms) as compared to a small hundred distance (M = 788ms, SD = 147ms; $t(52) = 18.80$, $p < .001$; $d = 2.58$ CI[2.02; 3.14]). Moreover, both the hundred-decade ($t(52) = 6.34$, $p < .001$; $d = 0.87$ CI[0.55; 1.18]) and the hundred-unit compatibility effects were significant ($t(52) = 6.89$, $p < .001$; $d = 0.95$ CI[0.62; 1.27]). Responses were

[1] Distance effects are often investigated using a continuous measure of distance rather than a categorical one. However, because we based our analyses on Bahnmueller et al. (2015), we decided to follow the categorical approach in the orginal paper and to use a categorical variable for both the analysis focusing on the distance effect only and the more complex factorial analysis.

faster for compatible (hundred-decade: M = 731ms, SD = 135ms; hundred-unit: M = 728ms, SD = 132ms) compared to incompatible number pairs (hundred-decade: M = 749ms, SD = 134ms; hundred-unit: M = 751ms, SD = 138ms). The significance of results remains unchanged when correcting for multiple comparisons. Thus, all three numerical effects were also present when participants had to pick the smaller number.

We further ran a 2 × 2 × 2 × 2 mixed design ANOVA discerning the within-subject factors hundred distance, hundred-decade compatibility, and hundred-unit compatibility, as well as the between-subject factor language group (German vs. English) for the *pick smaller experiment*. As expected based on the results of the *t*-tests above, we observed significant main effects of hundred distance ($F(1,51) = 353.75$, $p < .001$, $\eta_p^2 = .87$), hundred-decade compatibility ($F(1,51) = 46.23$, $p < .001$, $\eta_p^2 = .48$), and hundred-unit compatibility ($F(1,51) = 51.32$, $p < .001$, $\eta_p^2 = .50$). Moreover, the interaction of hundred-distance and hundred-unit compatibility was significant ($F(1,51) = 4.66$, $p < .001$, $\eta_p^2 = .08$) indicating that the hundred-unit compatibility effect was significant for both small and large hundred distances (small: $t(52) = 6.16$, $p < .001$; large: $t(52) = 4.03$, $p < .001$) but was larger for small compared to large hundred distances ($t(52) = 2.24$, $p = .029$). Crucially, neither the main effect of language group ($F(1,51) = 1.88$, $p = .176$; $\eta_p^2 = 0.04$) nor any of the interactions with language group were significant (all $p \geq .142$). Thus, results for the *pick smaller experiment* provide no evidence for a difference in numerical effects between German and English speakers replicating observations of Bahnmueller et al. (2015) previously reported for the *pick larger experiment*. Results of a parallel Bayesian mixed design ANOVA showing a comparable pattern can be found at https://osf.io/27jty/.

Pick larger experiment

Paralleling analyses of the *pick smaller experiment* and providing a more focused analysis as presented in Bahnmueller et al. (2015), three separate paired *t*-tests were also run for the *pick larger experiment*. Comparable to the *pick smaller study*, a significant hundred distance effect was observed showing faster RTs for number pairs with a large (M = 728ms, SD = 160ms) as compared to small hundred distance (M = 815ms, SD = 176ms; $t(50)$ = 22.88, $p < .001$; d = 3.20 CI[2.52; 3.88]). In addition, the effect of hundred-decade compatibility was significant ($t(50) = 4.19$, $p < .001$; d = 0.59 CI[0.29; 0.88]; indicating that compatible number pairs (M = 764ms, SD = 172ms) were responded to faster than incompatible number pairs (M = 777ms, SD = 165ms). Finally, the effect of hundred-unit compatibility was also significant ($t(50) = 6.79$, $p < .001$; d = 0.95 CI[0.62; 1.28]) with compatible number pairs (M = 757ms, SD = 165ms) being responded to faster than incompatible number pairs (M = 784ms, SD = 172ms). Again, the significance of results remains unchanged when correcting for multiple comparisons. Refer to Bahnmueller et al. (2015) for results of the analysis of the full factorial design.

Pick smaller vs. pick larger experiment

Mean reaction time. Results of an independent *t*-tests showed no significant difference in mean RT between the pick larger (M = 770ms, SD = 168ms) and the pick smaller task instruction (M = 740ms, SD = 134ms; $t(52) = 1.03$, $p = .306$; d = 0.20 CI[-0.19; 0.59]).

Modulation of the hundred distance effect. A mixed design ANOVA with the within-subject factor hundred distance (small vs. large) and the between-subject factor instruction (*pick smaller* vs. *pick larger*) revealed a significant effect of hundred distance ($F(1,102) = 818.76$, $p < .001$, $\eta_p^2 = .89$; small: M = 801ms, SD = 162ms; large: M = 711ms, SD = 143ms). Neither the main effect of instruction ($F(1,102) = 1.05$, $p = .308$, $\eta_p^2 = .01$; *pick smaller*: M =

741ms, *SD* = 143ms; *pick larger*: *M* = 771ms, *SD* = 173ms) nor the interaction of hundred distance and instruction were significant ($F(1,102) = 1.56$, $p = .214$, $\eta_p^2 = .02$).

To quantify the evidence in case of non-significant results, we further ran a Bayesian mixed design ANOVA using default JASP prior scales. It revealed that the data were best represented by a model that included the main effect of hundred distance only. The Bayes Factor (BF_{10}) for this model was $4.33*10^{46}$, indicating strong evidence for this model over the null model. Results further showed strong evidence against the model only including the main effect of instruction ($BF_{10} = 1.29*10^{-47}$ or $BF_{01} = 7.75*10^{46}$) as the data were $7.75*10^{46}$ times more likely under the best model (i.e., the model only including the main effect of hundred distance). Moreover, results revealed weak/inconclusive evidence against the model including both main effects ($BF_{10} = 0.49$ or $BF_{01} = 2.03$) and moderate evidence against the model additionally including the interaction term ($BF_{10} = 0.18$ or $BF_{01} = 5.55$) when compared to the best model (cf. Table 1, see also https://osf.io/27jty/ for JASP output and analyses files).

Table 1

Results of the Bayesian mixed design ANOVA with the within-subject factor hundred distance and between-subject factor instruction.

Model comparison

models	P(m)	P(m|data)	BF_M	BF_{10}
HD	0.200	0.598	5.946	1.000
HD + instruction	0.200	0.294	1.670	0.493
HD + instruction + HD × instruction	0.200	0.108	0.483	0.180
null model	0.200	$1.380*10^{-47}$	$5.521*10^{-47}$	$2.309*10^{-47}$
instruction	0.200	$7.717*10^{-47}$	$3.087*10^{-47}$	$1.291*10^{-47}$

Analyses of effects

effects	P(incl)	P(incl|data)	BF_{incl}
HD	0.400	0.892	$4.146*10^{46}$
instruction	0.400	0.294	0.493
HD × instruction	0.200	0.108	0.366

Note. HD = hundred distance; m = model; incl = inclusion

Modulation of the hundred-decade compatibility effect. A mixed design ANOVA with the within-subject factor hundred-decade compatibility (compatible vs. incompatible) and the between-subject factor instruction revealed a significant effect of hundred-decade compatibility ($F(1,102) = 54.64$, $p < .001$, $\eta_p^2 = .35$; compatible: $M = 747$ms, $SD = 154$ms; incompatible: $M = 763$ms, $SD = 150$ms). The interaction of hundred-decade compatibility and instruction was not significant ($F(1,102) = 1.57$, $p = .213$, $\eta_p^2 = .02$). The paralleling Bayesian mixed design ANOVA showed that the data were best represented by a model that included the main effect of hundred-decade compatibility only. The BF_{10} for this model was $1.72*10^8$, indicating strong evidence for this model when compared to the null model. Moreover, there was strong evidence against the model only including the main effect of instruction ($BF_{10} = 2.51*10^{-9}$ or $BF_{01} = 3.98*10^8$) by indicating that the data are $3.98*10^8$ times more likely under the best model (i.e., only including the main effect of hundred-decade compatibility). Finally, results revealed weak/inconclusive evidence against the model including both main effects ($BF_{10} = 0.45$ or $BF_{01} = 2.23$) and moderate evidence against the model additionally including the interaction term ($BF_{10} = 0.17$ or $BF_{01} = 5.97$) when compared to the best model.

Table 2

Results of the Bayesian mixed design ANOVA with the within-subject factor hundred-decade compatibility and between-subject factor instruction.

Model comparison

models	P(m)	P(m|data)	BF_M	BF_{10}
HDC	0.200	0.619	6.493	1.000
HDC + instruction	0.200	0.278	1.537	0.449
HDC + instruction + HDC × instruction	0.200	0.104	0.462	0.167
null model	0.200	$3.595*10^{-9}$	$1.438*10^{-8}$	$5.810*10^{-9}$
instruction	0.200	$1.553*10^{-9}$	$6.213*10^{-9}$	$2.510*10^{-9}$

Analyses of effects

effects	P(incl)	P(incl|data)	BF_{incl}
HDC	0.400	0.896	1.741*108
instruction	0.400	0.278	0.449
HDC × instruction	0.200	0.104	0.373

Note. HDC = hundred-decade compatibility; m = model; incl = inclusion

Modulation of the hundred-unit compatibility effect. A final mixed design ANOVA with the within-subject factor hundred-unit compatibility (compatible vs. incompatible) and the between-subject factor instruction revealed a significant effect of hundred-unit compatibility ($F(1,102) = 93.43$, $p < .001$, $\eta_p^2 = .48$; compatible: $M = 742$ms, $SD = 149$ms; incompatible: $M = 767$ms, $SD = 155$ms). The interaction of hundred-decade compatibility and instruction was not significant ($F(1,102) = 0.47$, $p = .494$, $\eta_p^2 = .01$). The corresponding Bayesian mixed design ANOVA showed that the data were best represented by a model that included the main effect of hundred-decade compatibility only. The BF_{10} for this model was $1.06*10^{13}$, indicating strong evidence for this model when compared to the null model. When compared to the best model (i.e., only including the main effect of hundred-unit

compatibility), results revealed strong evidence against the model only including the main effect of instruction ($BF_{10} = 5.61*10^{-14}$ or $BF_{01} = 1.78*10^{13}$). Moreover, when compared to the best model, results revealed weak/inconclusive evidence against the model including both main effects ($BF_{10} = 0.43$ or $BF_{01} = 2.32$) and moderate evidence against the model additionally including the interaction term ($BF_{10} = 0.11$ or $BF_{01} = 9.01$; see Table 3).

Table 3
Results of the Bayesian mixed design ANOVA with the within-subject factor hundred-unit compatibility and between-subject factor instruction.

Model comparison

models	P(m)	P(m|data)	BF_M	BF_{10}
HUC	0.200	0.619	6.493	1.000
HUC + instruction	0.200	0.278	1.537	0.449
HUC + instruction + HUC × instruction	0.200	0.104	0.462	0.167
null model	0.200	$3.595*10^{-9}$	$1.438*10^{-8}$	$5.810*10^{-9}$
instruction	0.200	$1.553*10^{-9}$	$6.213*10^{-9}$	$2.510*10^{-9}$

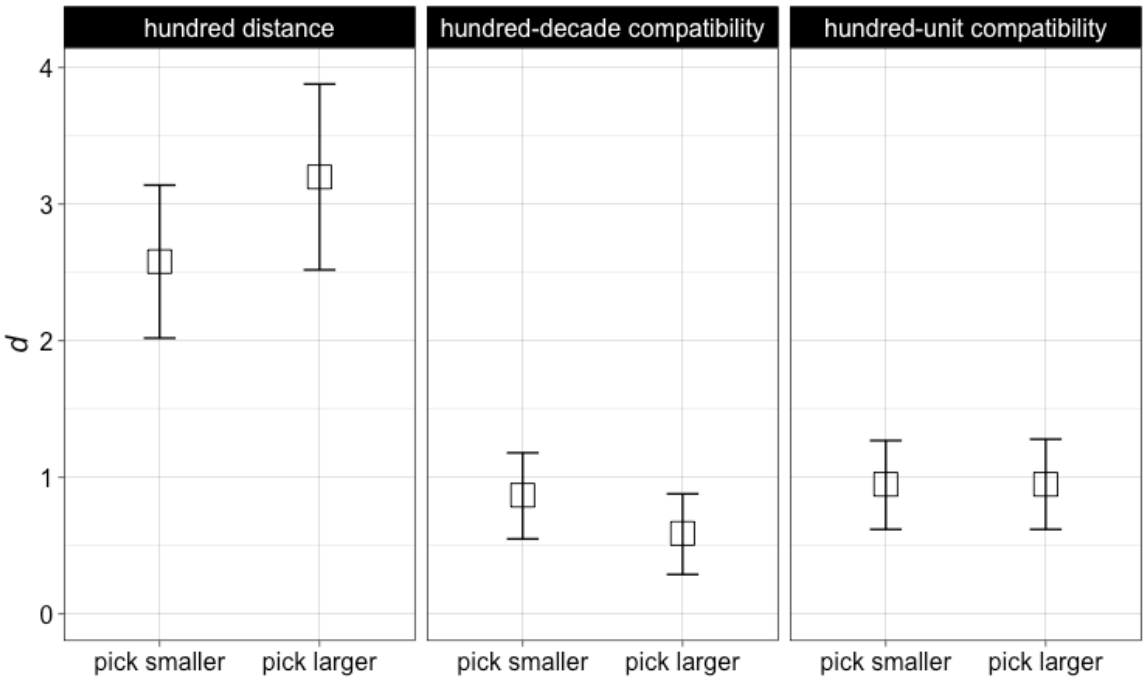
Analyses of effects

effects	P(incl)	P(incl|data)	BF_{incl}
HUC	0.400	0.928	$9.485*10^{12}$
instruction	0.400	0.279	0.431
HUC × instruction	0.200	0.072	0.258

Note. HUD = hundred-unit compatibility; m = model; incl = inclusion

Figure 1 illustrates Cohen's *d* and 95% confidence intervals around the respective effect separately for each numerical effect and instruction (pick smaller vs. pick larger). In line with Bayesian analyses, similar point estimates and largely overlapping confidence intervals do not provide evidence for a difference in numerical effect between experiments.

Figure 1. Cohen's *d* and 95% confidence intervals presented separately for each numerical effect and task instruction.



Bin analyses: To explore potential differences in the time course of the effects of interest both within and across experiments, we further ran a bin analysis dividing the RT distribution in each condition into four equal bins (i.e., from fastest to slowest RTs; cf. Arend & Henik, 2015). In contrast to Arend and Henik (2015), the results pattern did not show evidence for a systematic influence of RT bin on the numerical effects of interest (neither in the pick smaller nor in the pick larger experiment). The differential result pattern may result from differences in effects under investigation (size congruity effect versus distance and compatibility effects), and number range (single vs. multi-digit numbers). For the interested reader results of these analyses are provided in the supplementary material (https://osf.io/27jty/).

Discussion

In a conceptual replication attempt of the study by Bahnmueller et al. (2015), the present study aimed at evaluating the generalizability of basic effects in multi-digit number processing across marked and unmarked task instructions. Overall, we replicated effects of hundred distance, hundred-decade-, as well as hundred-unit compatibility that were previously reported using an unmarked task instruction (i.e., pick the larger number, cf. Bahnmueller et al., 2015) in a three-digit number comparison task using a marked task instruction (i.e., pick the smaller number). Results showed no significant difference in overall reaction times between the comparison tasks using the marked (pick smaller) and the unmarked (pick larger) task instruction. Additional Bayesian analyses provided evidence that linguistic markedness of the task instruction did not affect the numerical effects of interest. Moreover, no evidence for a difference between experiments in the size of either one of the numerical effects was observed. These results were confirmed by Bayesian analyses providing moderate evidence against the interaction of task instruction and the respective numerical effect. Taken together, our data suggest that distance and compatibility effects and with this componential processing of multi-digit numbers are largely unaffected by variations of the linguistic markedness of task instructions.

Numerical effects and task instruction

In line with previous observations regarding three-digit number comparison tasks (Bahnmueller et al., 2015, 2016; Huber et al., 2013; Korvorst & Damian, 2008; Mann et al., 2012), we replicated both the hundred-decade and the hundred-unit compatibility effect as well as the effect of hundred distance in the pick smaller experiment. Importantly, effect sizes observed in the pick smaller experiment were very similar to those observed in the pick larger

experiment, and the interaction between task instruction and the numerical effects of interest was not significant. Moreover, Bayesian analyses provided moderate evidence against an influence of linguistic markedness on the three numerical effects under investigation.

Thus, no major disruptions of the behavioural signatures of multi-digit Arabic number processing were observed when participants were confronted with a marked task instruction. Thereby, the present study provides further evidence for the robustness of the numerical effects under investigation and suggests that these numerical effects do not seem to be bound to specific experimental setups. And further, as indexed by significant compatibility effects resulting from interference due to the decision irrelevant tens/unit digit, the present study provides evidence towards the componential processing account put forward for multi-digit number processing (cf. Huber et al. 2016).

General performance and task instruction

However, in contrast to previous findings in single- and two-digit number comparison (Arend & Henik 2015; Verguts and De Moor, 2005), we did not detect reliable differences in overall response times in frequentist analyses. Although the Bayesian analysis supports the null model, the evidential value is relatively weak. Thus, it is possible that with a larger sample the direction of the evidence would change providing evidence for an effect of linguistic markedness. However, given our sample size, this scenario seems rather unlikely. What we can conclude is that an effect of linguistic markedness on general reaction times, if it exists, must be rather subtle. Furthermore, as overall reaction times were comparable between experiments, the mechanism through which we anticipated modulations of the compatibility effects (i.e., longer reaction times when confronted with the marked task instruction resulting in more elaborated processing of a stimulus and, therefore, increased

interference due to the irrelevant tens/unit digit in incompatible trails) could not be demonstrated.

Moreover, it seems that most participants in the pick smaller experiment were fairly adaptive to the marked task instruction. Interestingly, in the pick smaller experiment, four participants had to be excluded from the analyses because they consistently picked the larger number although instructed to pick the smaller one. Similar confusions did not occur in the pick larger experiment. Thereby, our results may suggest that, when comparing numbers beyond the two-digit number range, following an unmarked task instruction relies on an initial categorical internalization of the task instruction rather than on a continuous, ongoing conflict or source of interference throughout the comparison task. As this account is rather speculative, future studies might consider manipulating linguistic markedness of the task instruction in within-participant designs, for instance, using a task switching paradigm (cf. Shaki et al., 2012). In such a task switching paradigm participants would have to switch between marked and unmarked task instructions when comparing numbers on a trial by trial basis. This would allow for evaluating whether marked task instructions indeed influence multi-digit number processing on a trial by trial basis when an initial categorical internalization of the task instruction is not possible.

Conclusion

Taken together, we successfully replicated main results reported by Bahnmueller et al. (2015) showing that distance and compatibility effects in a three-digit number comparison task generalize across marked and unmarked task instructions. Crucially, however, linguistic markedness of task instructions did not seem to influence basic numerical processing as the size of numerical effects was comparable between experiments using a marked compared to an unmarked task instruction. In particular, results suggest that basic strategies in three-digit

number processing are rather robust against variations of the linguistic markedness of task instructions.

References

Arend, I., & Henik, A. (2015). Choosing the larger versus choosing the smaller: Asymmetries in the size congruity effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1821. https://doi.org/10.1037/xlm0000135

Bahnmueller, J., Huber, S., Nuerk, H. C., Göbel, S. M., & Moeller, K. (2016). Processing multi-digit numbers: A translingual eye-tracking study. *Psychological Research, 80*(3), 422-433. http://doi.org/10.1007/s00426-015-0729-y

Bahnmueller, J., Moeller, K., Mann, A., & Nuerk, H.-C. (2015). On the limits of language influences on numerical cognition – No inversion effects in three-digit number magnitude processing in adults. *Frontiers in Psychology, 6*:1216, 211–226. http://doi.org/10.3389/fpsyg.2015.01216

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

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), 710-726. https://doi.org/10.1007/s00426-015-0677-6

Cipora, K., Schroeder, P. A., Soltanlou, M., & Nuerk, H.-C. (2018). More space, better mathematics: Is space a powerful tool or a cornerstone for understanding arithmetic?. In K. S. Mix & M. T. Battista (Eds.), *Visualizing Mathematics* (pp. 77-116). Springer, Cham.

Cipora, K., Soltanlou, M., Reips, U., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects.

Behavior Research Methods, 51, 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371-396. https://doi.org/10.1037/0096-3445.122.3.371

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*(4), 390–407. http://doi.org/10.1162/jocn.1993.5.4.390

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 626–641. http://doi.org/10.1037/0096-1523.16.3.626

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*(1–2), 43–74. http://doi.org/10.1016/0010-0277(92)90050-R

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

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(6), 875-894. https://doi.org/10.1037/0096-1523.8.6.875

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition, 10*(4), 389-395. https://doi.org/10.3758/BF03202431

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

Hinrichs, J. V., Yurko, D. S., & Hu, J. M. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 890. https://doi.org/10.1037/0096-1523.7.4.890

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

Huber, S., Moeller, K., Nuerk, H.-C., & Willmes, K. (2013). A computational modeling approach on three-digit number processing. *Topics in Cognitive Science, 5*(2), 317–334. http://doi.org/10.1111/tops.12016

Huber, S., Nuerk, H. C., Reips, U. D., & Soltanlou, M. (2019). Individual differences influence two-digit number processing, but not their analog magnitude processing: a large-scale online study. *Psychological Research, 83*(7), 1444-1464. https://doi.org/10.1007/s00426-017-0964-5

Huber, S., Nuerk, H., Willmes, K., & Moeller, K. (2016). A general model framework for multisymbol number comparison. *Psychological Review, 123*(6), 667–695. http://doi.org/10.1037/rev0000040

JASP Team (2020). JASP (Version 0.13.1) [Computer software]. Retrieved from https://jasp-stats.org

Korvorst, M., & Damian, M. F. (2008). The differential influence of decades and units on multidigit number comparison. *The Quarterly Journal of Experimental Psychology, 61*(8), 1250–1264. http://doi.org/10.1080/17470210701503286

Lyons, J. (1968). *Semantics*. Cambridge: Cambridge University Press.

Mann, A., Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H.-C. (2012). On the development of Arabic three-digit number processing in primary school children. *Journal of Experimental Child Psychology, 113*(4), 594–601. http://doi.org/10.1016/j.jecp.2012.08.002

Mapelli, D., Rusconi, E., & Umiltà, C. (2003). The SNARC effect: an instance of the Simon effect?. *Cognition*, *88*(3), B1-B10. https://doi.org/10.1016/S0010-0277(03)00042-8

Meyerhoff, H. S., Moeller, K., Debus, K., & Nuerk, H.-C. (2012). Multi-digit number processing beyond the two-digit number range: A combination of sequential and parallel processes. *Acta Psychologica, 140*(1), 81–90. http://doi.org/10.1016/j.actpsy.2011.11.005

Moeller, K., Klein, E., & Nuerk, H.-C. (2013). Influences of cognitive control on numerical cognition—Adaptation by binding for implicit learning. *Topics in Cognitive Science, 5*(2), 335-353. https://doi.org/10.1111/tops.12015

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520. http://doi.org/10.1038/2151519a0

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

Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Multi-digit number processing: Overview, conceptual clarifications, and language influences. In R. C. Kadosh & A. Dowker

(Eds.), *The Oxford Handbook of Numerical Cognition* (pp. 106–139). Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199642342.013.02

Nuerk, H.-C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition, 82*(1), B25–B33. http://doi.org/10.1016/S0010-0277(01)00142-1

Patro, K., & Haman, M. (2012). The spatial–numerical congruity effect in preschoolers. *Journal of Experimental Child Psychology*, *111*(3), 534-542. https://doi.org/10.1016/j.jecp.2011.09.006

Pixner, S., Moeller, K., Heřmanová, V., Nuerk, H.-C., & Kaufmann, L. (2011). Language effects on nonverbal number processing in first grade — A trilingual study. *Journal of Experimental Child Psychology, 108*(2), 371–382. http://doi.org/10.1016/j.jecp.2010.09.002

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology, 83*(2, Pt.1), 274–278. http://doi.org/10.1037/h0028573

RStudio Team. (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. Retrieved from https://www.rstudio.com/

Shaki, S., Petrusic, W. M., & Leth-Steensen, C. (2012). SNARC effects with numerical and non-numerical symbolic comparative judgments: instructional and cultural dependencies. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 515-530.

Sherman, M. A. (1973). Bound to be easier? The negative prefix and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 12*(1), 76-84. https://doi.org/10.1016/S0022-5371(73)80062-3

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior, 15*(2), 143-157. https://doi.org/10.1016/0022-5371(76)90015-3

Steiner, A. F., Finke, S., Clayton, F. J., Banfi, C., Kemény, F., Göbel, S.M. & Landerl, K. (this issue). Language effects in early development of number writing and reading. *Journal of Numerical Cognition.*

The jamovi project (2020). jamovi (Version 1.2) [Computer Software]. Retrieved from https://www.jamovi.org

van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., … Wagenmakers, E. (2019, January 23). The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. https://doi.org/10.31234/osf.io/yqxfr

Verguts, T., & De Moor, W. (2005). Two-digit comparison - Decomposed, holistic, or hybrid? *Experimental Psychology, 52*(3), 195–200. http://doi.org/10.1027/1618-3169.52.3.195

Acknowledgements

This research was funded by a grant from the DFG (NU 265/3-1) to H-CN supporting KC and MS. H-CN is further supported by the LEAD Graduate School & Research Network (GSC1028), which was funded within the framework of the Excellence Initiative of the German federal and state governments. We thank Lia Heubner and Marie-Lene Schlenker for their help with data collection.

Supplementary Material

Bin analysis

To explore potential differences in the time course of the numerical effects of interest both within and across experiments, we ran a bin analysis dividing the RT distribution in each condition into four equal bins (cf. Arend & Henik, 2015). In particular, we split RTs for each participant, experiment and condition into four bins where bin 1 in each condition included the fastest 25% of trials and bin 4 included the slowest 25% of trials. Paired *t*-tests for each numerical effect (alongside Cohen's d and 95% confidence intervals) were then calculated for each of the four bins separately for the pick smaller and the pick larger experiment (for results of t-tests see Table S1). The results pattern (cf. Figure S1) did not show evidence for a systematic influence of RT bin on the numerical effects of interest (neither in the pick smaller nor in the pick larger experiment).

Table S1

Results of the bin analysis showing t-statistic and p-value for each bin, numerical effect, and task instruction.

		Hundred distance effect		**Hundred-decade compatibility**		**Hundred-unit compatibility**	
		t	*p*	*t*	*p*	*t*	*p*
pick smaller	**bin 1**	11.02	<.001	4.22	<.001	4.9	<.001
	bin 2	16.04	<.001	4.57	<.001	7.74	<.001
	bin 3	16.80	<.001	4.94	<.001	6.14	<.001
	bin 4	16.13	<.001	4.88	<.001	3.93	<.001
pick larger	**bin 1**	16.98	<.001	1.39	.170	8.16	<.001
	bin 2	20.01	<.001	3.72	<.001	8.27	<.001
	bin 3	23.06	<.001	7.39	<.001	6.09	<.001
	bin 4	17.37	<.001	2.56	.014	4.27	<.001

Note. df=52 pick smaller experiment, df=50 pick larger experiment

Figure S1. Cohen's *d* and 95% confidence intervals presented separately for each numerical effect, task instruction, and bin.