Comparing Symbolic and Nonsymbolic Number Lines: Consistent Effects of Notation Across Output Measures

Karl K. Kopiske [1, 3] *and Volker H. Franz* [2]

[1] Istituto Italiano di Tecnologia, Center for Neuroscience and Cognitive Systems, Rovereto, TN, Italy

[2] University of Tübingen, Department of Computer Science, Experimental Cognitive Sciences, Tübingen, Germany

[3] Chemnitz University of Technology, Institute of Physics, Cognitive Systems Lab, 09126 Chemnitz, Germany

ABSTRACT

The mental number line (MNL) is a popular metaphor for magnitude representation in numerical cognition. Its shape has frequently been reported as being nonlinear, based on nonlinear response functions in magnitude estimation. We investigated whether this shape reflects a phenomenon of the mapping from stimulus to internal magnitude representation or of the mapping from internal representation to response. In five experiments, participants (total $N = 66$) viewed stimuli that represented numerical magnitude either in a symbolic notation (i.e., Arabic digits) or in a nonsymbolic notation (i.e., clouds of dots). Participants estimated these magnitudes by either adjusting the position of a mark on a ruler-like response bar (nonsymbolic response) or by typing the corresponding number on a keyboard (symbolic response). Responses to symbolic stimuli were markedly different from responses to nonsymbolic stimuli, in that they were mostly power-shaped. We investigated whether the nonlinearity could be explained by effects of previous trials, but such effects were (a) not strong enough to explain the nonlinear responses and (b) existed only between trials of the same input notation, suggesting that the nonlinearity is due to input mappings. Introducing veridical feedback improved the accuracy of responses, thereby showing a calibration based on the feedback. However, this calibration persisted only temporarily, and responses to nonsymbolic stimuli remained nonlinear. Overall, we conclude that the nonlinearity is a phenomenon of the mapping from nonsymbolic input format to internal magnitude representation and that the phenomenon is surprisingly robust to calibration.

KEYWORDS

numerical cognition, nonsymbolic magnitude, number line, calibration

INTRODUCTION

The model of the mental number line (MNL) for an internal scale of numerical magnitude has been around at least since the 1960s (Moyer & Landauer, 1967), but was postulated in its current form by Dehaene (1992). Dehaene proposed the MNL to be one element of a model of number representation that sought to explain, among other things, the ability of neurological patients to make approximate, but not exact judgements based on simple verbal input, as well as spatial-numerical stimulus-response associations (SNARC—Dehaene, Bossini, & Giraux, 1993). In this view, the MNL determines the internal mapping between numbers and other forms of magnitude. In this article, we use the MNL in a similar sense to refer to the internal representation that gives rise to an observable response function. Our goal was to investigate input-to-representation mappings and representation-to-output transformations that may give rise to particular attributes of this response function, in particular its nonlinear shape (Dehaene, 2003).

Corresponding author: Karl K. Kopiske, Cognitive Systems Lab, Institute of Physics, Chemnitz University of Technology, Reichenhainer Str. 70, 09126 Chemnitz, Germany. E-mail: karl.kopiske@physik.tu-chemnitz.de

The Shape of the Mental Number Line and its Relation to Nonsymbolic Number

Dehaene (1992) noted that the mapping of nonsymbolic magnitude to symbolic magnitude (e.g., numbers) tends to be nonveridical, showing systematic underestimation for large magnitudes; a finding that caused him to propose a MNL that was compressed and possibly logarithmic in shape: response(x)~log(x), with x being the numerical magnitude of the stimulus.

Two things should be noted with respect to the shape of the MNL. Firstly, if the MNL refers to the internal representation giving rise to the response function, it might still be linear even if the response function happens to be nonlinear. A related point is that the origin of both the shapes of the MNL and the response function has been much debated (Barth & Paladino, 2011; Cantlon, Cordes, Libertus, & Brannon, 2009; Cicchini, Anobile, & Burr, 2014; Cohen & Blanc-Goldhammer, 2011; Dehaene, 2003; Dehaene, Izard, Pica, & Spelke, 2009; Siegler & Opfer, 2003), as has the question whether testing one allows inferences about the other: This requires knowledge of the mapping between the two and of potential response biases. For example, it has been argued that in a classic task of locating numbers (or other forms of magnitude) on a horizontal line akin to a ruler, participants actually perform a proportion judgement, which, in turn, relies heavily on reference points (Barth & Paladino, 2011; but see also Opfer, Siegler, & Young, 2011). At the same time, this ruler-like task is one of the few tasks that allow a nonsymbolic output of magnitude, which bypasses a nonsymbolic-to-symbolic transformation that is required for other tasks, such as, for example, verbalising magnitude. A recently proposed potential solution to the problem of proportion judgments is allowing participants to go beyond the presented ruler, thereby effectively allowing a judgement of multiples (Cohen & Blanc-Goldhammer, 2011; Link, Huber, Nuerk, & Moeller, 2014). Another potential confound that may bias responses independently of internal representation is the known tendency towards the mean, be it of a scale or of previous responses. This has recently been brought up as criticism of the notion that a compressed response suggests a compressed representation of magnitude (Cicchini et al., 2014) and has been applied to other judgements for a long time (Haubensak, 1992; Parducci & Perret, 1971).

Secondly, the systematic underestimation in numerical estimation for large magnitudes can be explained in several ways that do not assume an exactly logarithmic transformation. A fairly similar view is that of the response being a power function of x, which is indeed what even proponents of an internal logarithmic MNL have argued (Izard & Dehaene, 2008). This will often result in very similar fits to behavioural data (indeed, the models may be virtually indistinguishable unless the number range is extended, see Opfer et al., 2011) and fit similarly well to the corresponding neural activation (Nieder & Miller, 2003). That said, the two are based on slightly different classic concepts with slightly different implications, as a logarithmic function implies an additive effect when stimulus magnitude is increased by a given factor (Fechner, 1860), while a power function implies a multiplicative effect (Stevens, 1957). Alternatively, in designs employing a bounded response, a linear internal representation may also be compatible with a compressed response function purely due to size-dependent variability (see Weber's Law, Fechner, 1860; such a relationship has also been found for transformations between symbolic and nonsymbolic magnitudes, see, e.g., Cordes, Gelman, Gallistel, & Whalen, 2001; Dehaene, 1992; Whalen, Gallistel, & Gelman, 1999). If variability increases with the response, a larger tail of the distribution would be truncated by the bound but more so for large responses than for smaller response, resulting in a systematic underestimation of large magnitudes (see, e.g., Cantlon et al., 2009).

Finally, it should also be noted that the model of the MNL as the basis of manipulation of approximate magnitudes is not universally accepted. Prominently, McCloskey (1992) and McCloskey, Caramazza, and Basili (1985) proposed a model in which a similar role is occupied by a semantic, abstract representation that sits between comprehension and production of numerical magnitudes, but that does not have a spatial aspect. However, the degree to which a semantic representation is necessary to process numerical magnitudes has been debated (see, e.g., Cipolotti & Butterworth, 1995, or, more recently, Gebuis, Gevers, & Cohen Kadosh, 2014; Leibovich, Katzin, Harel, & Henik, 2017; for reviews about the neural representation of magnitude, see Dehaene, Piazza, Pinel, & Cohen, 2003; Nieder, 2016).

Our Study

Our goals were twofold: Firstly, to investigate which step of an input-output mapping creates the nonlinear response function in magnitude estimation. We were also interested in whether the same shape would be achieved not only with symbolic-output measures, but also with different variations of a nonsymbolic number estimation task. To do this, we compared number lines obtained from symbolic-to-nonsymbolic as well as nonsymbolic-to-symbolic transformations, having participants map different types of input to the same output measure, as well as the same input to different output measures. Secondly, we wanted to find out if previous-trials effects and calibration (i.e., learning of an input-output mapping) could explain the shape of the responses (e.g., by biasing responses towards the mean of previous magnitudes and thus the mean of the scale, Cicchini et al., 2014). Further, we wanted to test if previous-trial effects existed, and if they did, whether any effects of previous trials and of calibration would persist between trials of different input notations (needing separate input-to-representation mappings) as well as between different responses (requiring different mappings from representation to output).

To test this, we employed two tasks of estimating numerosity: a nonsymbolic response, in which participants were asked to indicate the magnitude represented by the stimulus on a response bar akin to a ruler (henceforth referred to as the *ruler-based task*; this was similar to previous studies, e.g., Cicchini et al., 2014; Dehaene, Izard, Spelke, & Pica, 2008; Siegler & Opfer, 2003) as well as a task with a simple numeric (symbolic) response, in which participants typed the corresponding number on a keyboard (which we will refer to as the *typed-response task*). The ruler-based task was performed with both Arabic digits (Experiments 1, 2a-4a) and clouds of dots as nonsymbolic stimulus

magnitudes (Experiments 2b-4b, 5). Its primary purpose was to have participants estimate magnitudes (as opposed to making binary comparison judgements, which may produce somewhat different effects: see Gebuis & Reynvoet, 2012) in a way that did not rely on verbal or symbolic representation (i.e., numbers). The stimuli were drawn randomly from magnitudes between 1 and 200. For nonsymbolic stimuli, this implies that most numbers would be in a range where participants would be unable to instantly perceive the precise magnitude (*subitizing*, which participants are able to do with numbers up to 4, or perhaps even up to 7; see Mandler & Shebo, 1982; Trick, 2008; Trick & Pylyshyn, 1994) and, since we also imposed a time constraint on the task, also be unable to count dots even for relatively small arrays. This reflected the fact that the main interest of our study was in approximate estimations of magnitude—that is, the processing of magnitudes that could not be subitized.

We also varied features of the response bar between experiments to rule out alternative explanations (e.g., we employed both numerosities and numbers as endpoints). The typed-response task was performed only in Experiment 5. In this experiment, participants received both veridical as well as perturbed feedback in order to investigate the mechanisms behind a possible calibration of the MNL. For an overview of all conditions and experiments, see Table 1.

For each experiment and each type of stimulus and response, we fit separate linear, logarithmic, and power functions to assess the shape of the response function. Our predictions were as follows: We expected the response function for symbolic stimuli (Arabic digits) to be almost linear and the response function for nonsymbolic stimuli (clouds of dots) to be better fitted through a logarithmic or power-function model (Dehaene, 1992). We expected this relation to hold true for both the ruler-based task (Experiments 1-4) and the typed-response task (Experiment 5). We also expected dependencies between consecutive trials. The presence of such serial dependencies between different types of trials (i.e., a symbolic-stimulus trial followed by a nonsymbolic-stimulus trial or vice-versa) would speak for a calibration of the mapping from internal magnitude to the response, since this would mean calibration generalizing across different input types. The absence of such between-trial-type serial dependencies would indicate stimulus-specific calibration and, thus, calibration of input-to-representation mapping. These two possibilities were investigated in more detail in Experiment 5.

TABLE 1.
Tasks and Stimuli Used in Experiments

Experiment	Stimulus	Response	Ruler endpoints	Comment
Exp. 1	symbolic	ruler-like bar	symbolic	
Exp. 2a	symbolic	ruler-like bar	symbolic	
Exp. 2b	nonsymbolic		symbolic	
Exp. 3a	symbolic	ruler-like bar	symbolic	starting position random
Exp. 3b	nonsymbolic		nonsymbolic	starting position random
Exp. 4a	symbolic	ruler-like bar	symbolic	endpoint mapping & starting position random
Exp. 4b	nonsymbolic		nonsymbolic	endpoint mapping & starting position random
Exp. 5a	nonsymbolic	ruler-like bar	nonsymbolic	with feedback, otherwise like 3b
Exp. 5b	nonsymbolic	typed response	none	with feedback

Note. Details about the ruler-like response bar are described in the Apparatus section. Detail about the symbolic/nonsymbolic endpoints of the ruler given in the Results and Discussion subsection in Experiment 3. For each experiment, Conditions A and B were presented in a blocked design that included both separate and mixed blocks. For Experiment 4, endpoint mapping—that is, which point on the number line each endpoint of the ruler corresponded to—could be left-small/right-large (in half the trials) or right-large/left-small (in the other half).

EXPERIMENT 1 (PILOT): TESTING THE RESPONSE FUNCTION

In this pilot experiment, we tested the response function of the ruler-based task. Participants clicked on a response bar displayed horizontally on a computer screen. A similar method has been used in numerous experiments (e.g., Cicchini et al., 2014; Dehaene et al., 2008; Siegler & Opfer, 2003). However, since we wanted to verify whether this method and its implementation was not in itself susceptible to artefacts, we tested it in Experiment 1 in the simplest, most straightforward version we could find: mapping numbers written in Arabic digits to a horizontal ruler marked on the left and on the right by numbers in the same notation (see Figure 1). A relatively linear response function close to unity would indicate that estimating magnitudes in such a way does not per se produce distorted responses. We tested if responses would be influenced by stimuli presented in previous trials, with "oddball blocks" of deliberately imbalanced magnitudes being included to maximize the discrepancy between trials.

Methods

PARTICIPANTS

Six participants (age range: 25 to 39 years, M_{age} = 32.2, 4 females) took part in the experiment. All participants were volunteers taking part without compensation, consisting of graduate students and faculty members of the Department of Psychology at the University of Hamburg. In this and all following experiments, participants gave written informed consent and their data were protected in accordance with the 1964 Declaration of Helsinki.

APPARATUS

Sitting in front a 21 in. LCD monitor (effective screen diagonal: 52 cm), participants were presented with a centrally displayed number (written in Arabic digits, font size 60 px, approximately 2° of visual angle). All stimuli were black, presented against a white background. Below the number, a black response bar was displayed (located near the bottom of the screen, centrally on the x-axis, approximately 20° of visual angle). In the middle of the response bar was a black mark (a square of 20 px × 20 px, corresponding to approximately 6 mm × 6 mm). This mark could be moved horizontally by the participants using a standard USB mouse. Participants were asked via on-screen instructions to move the mark to the location that they perceived as the position the number belonged to and then click the mouse button to register this position (see Figure 1).

PROCEDURE

In each trial, the number was displayed until the participant performed the mouse click, after which a fixation cross appeared for on average 500 ms (this interval was defined as 400 ms + a pseudo-random value from an exponential distribution with $M = 100$ ms) until the next trial. At the start of the experiment, a response bar was presented on the screen below the instructions to let participants familiarise themselves with the bar and the adjustment mark for as long as they liked, while clicking was disabled. All experiments were implemented in a custom MATLAB program using Psychtoolbox 3 (Kleiner et al., 2007).

To be able to verify the results from our models and to enable easier detection of previous-trial effects, the experiment was conducted in a blocked design, with one block containing pseudo-random numbers (in which stimuli were randomly drawn from numbers 1-200, with no duplicates), as well as two oddball blocks. These were included to test specifically the degree to which the responses would be influenced by the trials directly preceding them by maximizing the discrepancy between magnitudes in oddball trials and the rest of the trials in such blocks: In these blocks, either were 87.5% of trials with high numbers (from the top third, 134-200) and 12.5% of trials with numbers from the opposite end of the range (1-66) or vice-versa for high and low numbers. An influence on the response by previous trials, and, indeed, the range of stimuli, is often found in repeated-measures designs (Haubensak, 1992; Parducci & Perret, 1971) and has been proposed specifically for processing of nonsymbolic numbers (Cicchini et al., 2014). All participants started with a random-number block, followed by two oddball blocks. The order of oddball-up and oddball-down blocks was counterbalanced between participants. Each block consisted of 64 trials, resulting in a total of 192 trials for the whole experiment.

DATA ANALYSIS

Our main dependent variable was the location of the click on the response bar. This was coded as the relative position on the response bar, with 0 corresponding to a click on the leftmost end of the response bar and 1 to a click on the rightmost end. Responses given after less than 500 ms were excluded. Outliers were excluded according to the following method: For each presented magnitude x we linearly interpolated the "expected" response based on responses to stimuli $[x - 10, x + 10]$ (truncated for responses near the top or bottom end of the stimulus range). We determined the *SD* of the residuals of the linear interpolation for all responses excluding the response to x. Responses with a distance of more than 3 *SD* from the expected response were considered outliers and excluded. In Experiment 1, this applied to 28 trials (2.4% of all trials).

Responses, aggregated by numeric values, were fitted to three models: a linear function ($y = a + b \times x$), a logarithmic function ($y = a + b \times \log[x]$), and a power function ($y = a \times x\text{^}b$), with x being the numeric value of the stimuli and y the response of the participants. Note that the intercept for the power function was fixed at 0, since (a) we wanted to fit the same number of parameters in all models, and (b) this form is the classic power function common in research on human perception (Stevens, 1957; Teghtsoonian, 1965).

Coefficients for the models were fit for each participant. The best-fitting model type, as indicated by the lowest Akaike information criterion (AIC; Akaike, 1974; see Burnham & Anderson, 2004, for guidelines on the interpretation), was selected. This model was then used to investigate if previous-trial effects would explain more variance in the data. To do this, we fit a linear, logarithmic, or power model to the trial-by-trial (i.e., unaggregated) data and compared the fit of this simple model to a model that included the numeric magnitude in the previous trial as an additional predictor. We also conducted a repeated-measures analysis of variance (ANOVA) with relative error—defined as: $(\text{response}[x] - x)/x$—as the dependent variable and the 3-level factor oddball (up/down/no oddball) to investigate if oddball trials displayed a systematically different error to other trials, that is, whether responses would be affected by preceding trials. Greenhouse-Geisser correction (Greenhouse & Geisser, 1959) was applied in all situations where sphericity could be violated. Bonferroni-Holm correction was applied to all t tests (Holm, 1979).

RESULTS AND DISCUSSION

We found that the data were fit best by a linear model of $y = 1.07x - 7.90$, explaining 99% of the variance. Including the magnitude presented in the previous trial created a slightly better fit ($\Delta AIC = -2.71$), with the predictor having a negative weight ($b = -0.0139$). Detailed descriptions of all models can be seen in Table 2, visualisations of the models can be seen in Figure 2.

Our ANOVA revealed a main effect of the factor oddball, $F(2, 10) = 28.01$, $p_{GG} < .001$, $e_{GG} = .83$, although post-hoc *t* tests (comparing relative error in oddball trials to relative error in all range-matched nonoddball trials; this range-matching of magnitudes was done only for the *t* tests, and not for the data included in the ANOVA) indicated that while descriptively, oddball trials where the oddball was smaller produced a smaller response than nonoddball trials and vice-versa for oddball-trials "upwards", these were not statistically significant differences, $t(5) = -1.228$, $p = .274$ and $t(5) = 1.153$, $p = .301$, respectively.

The main goal of Experiment 1 was to function as a pilot of sorts and ascertain that the mapping of numbers to our ruler-like response bar was relatively accurate. This was the case: The mapping was almost

perfectly linear, and the slope of the model was close to 1. Thus, we felt confident using this condition as a control in the following experiments. With regards to previous-trial effects, a positive coefficient in the model for previous magnitude indicates that relatively larger previous trials would lead to somewhat larger responses, while the analysis of oddball trials showed no clear pattern. These were exploratory analyses, however, and we aimed to test this in the experiments that followed.

EXPERIMENT 2, 3, AND 4: NONSYMBOLIC MAGNITUDE AND ITS RELATION TO SYMBOLIC MAGNITUDE

Experiments 2, 3, and 4 incorporated not only symbolic, but also nonsymbolic magnitudes as stimuli. That is, half of the trials consisted of participants being presented with Arabic digits and clicking on the respective position on the response bar, the other half consisted of the same task but with participants being presented with clouds of dots instead of digits. They were instructed to assess the number of dots in these clouds and select the appropriate location on the response bar in the same way as with Arabic digits.

The experiments differed only in subtle, but nevertheless important details (see Figure 1 and Table 1). Experiment 2 employed a response bar with symbolic endpoints (i.e., Arabic digits) for both symbolic and nonsymbolic stimuli. These were presented in a blocked design that included both blocks of one notation only and mixed blocks containing both notations. This gave us a first idea if our method was appropriate for nonsymbolic stimuli and whether our results would be in line with the literature. In Experiments 3 and 4, we attempted to rule out potential confounds that might have influenced our results in Experiment 2, and test whether the pattern of results was robust to small variations in the experimental design. In Experiment 3, we employed a response bar with endpoints defined by nonsymbolic numerosities for nonsymbolic stimuli (so that the mapping from nonsymbolic magnitude to output did not involve a symbolic notation). In Experiment 4, we used the design of Experiment 3, but flipped around the response bar in half of the trials, such that the upper end would now be on the left side. We also randomised the starting position of the adjustment mark on the response bar, which had previously always been in the middle, to prevent participants from learning to execute movements rather than

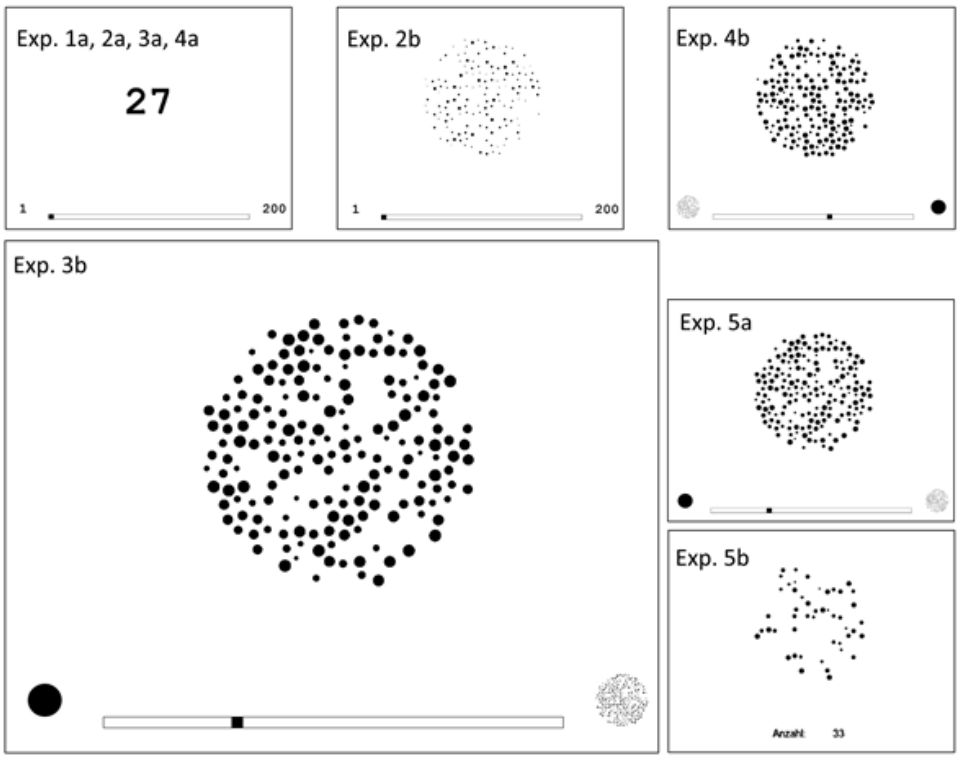

FIGURE 1.

Screenshots from each of our experiments. Top row, left = Experiments 1A-4A, ruler-based response, symbolic stimuli, symbolic endpoints; Middle = Experiment 2B, ruler-based response, nonsymbolic stimuli, symbolic endpoints; Right = Experiment 4B, ruler-based response, nonsymbolic stimuli, nonsymbolic (and sometimes flipped) endpoints; Right column, middle = Experiment 5A, ruler-based response, nonsymbolic stimuli, nonsymbolic endpoints; Bottom = Experiment 5B, typed response to nonsymbolic stimuli. The line asking for and displaying the response is magnified for visibility in this figure. The word Anzahl is German for number or numerousness; Bottom left, big panel = Experiments 3B and 5A, ruler-based response, nonsymbolic stimuli, nonsymbolic endpoints. See also Table 1 for a summary of the conditions.

performing an estimate for each trial. Each participant took part in only one of the Experiments 1-4. The purpose of these experiments was to investigate whether (a) the typical nonlinear MNL shape with underestimation for relatively large numbers could reliably be found in number-line tasks with symbolic and nonsymbolic input, even when controlling for the above-mentioned possible confounds of the ruler-based task, and (b) whether there would be previous-trial effects within, or even between trial-types.

Experiment 2: Symbolic and Nonsymbolic Magnitude on a Standard Response Bar

In Experiment 2, we introduced nonsymbolic stimuli in addition to symbolic stimuli (numbers). Thus, we were able to compare the responses in the same task – that is, responses on a ruler-like response bar – for symbolic and nonsymbolic stimuli. We expected our design with these stimuli to replicate results obtained in previous studies using a similar design that showed a markedly nonlinear response function (e.g., Dehaene et al., 2008; Siegler & Opfer, 2003), as well as effects of previous-trial magnitudes (Cicchini et al., 2014). With respect to the latter, we also wanted to investigate whether the effect would persist when nonsymbolic stimuli were preceded by symbolic stimuli (i.e., in different input-type dyads). This would be expected if previous-trial effects were driven by a calibration of the mapping from internal representation to output, but not if previous-trial effects show a calibration of the input-to-representation mapping.

METHODS

Participants. Eight participants (students of University of Hamburg, aged between 19 and 26 years, M_{age} = 22.9; 6 females) were tested. Each participant received course credit or €8/hr.

Apparatus. The same setup as in Experiment 1 was used. The main difference was that we employed not only symbolic but also nonsymbolic stimuli. These nonsymbolic stimuli were generated using a modified version of a program developed by Gebuis and Reynvoet (2011) that will generate clouds of dots and was designed to keep visual stimulus properties uninformative about the number of dots in a certain design. In our design, keeping visual dimensions completely uninformative about number would have been impossible, since nonsymbolic magnitude is ultimately defined by visual features, and all of our stimuli differed in magnitude. Thus, we settled for a compromise in which the visual features total area of the clouds of dots (r = .41), density (r = .28), surface area of the dots (r = .57) and circumference of the cloud (r = .77) were all imperfectly correlated with nonsymbolic magnitude.

Procedure. As in Experiment 1, participants indicated the position of a symbolic or nonsymbolic stimulus on a ruler-like bar presented horizontally between the numbers 1 and 200 (in all tables and figures, the symbolic condition is indicated as *Exp. 2A* and the nonsymbolic as *Exp. 2B*; the same nomenclature is used for Experiments 3 and 4) and were instructed to click on the location on the ruler/like response bar where they thought each current stimulus belonged. We now included a time limit of 3 s. When this was exceeded, an error message ("Please answer within 3 seconds!" in German) appeared on the screen. Symbolic and nonsymbolic stimuli were presented in a counterbalanced blocked design that contained single-type random magnitude blocks, single-type oddball blocks, and mixed-type oddball blocks. Blocks were randomised in the same manner as in Experiment 1, with the order of symbolic, nonsymbolic, and mixed blocks counterbalanced between participants. The random-number block was always the first for each stimulus type and the mixed blocks were always at the end of the experiment. This gave us 8 possible permutations: (oddball up first/oddball down first) × (nonsymbolic first /symbolic first) × (mixed, nonsymbolic oddballs first) × (mixed, symbolic oddballs first). In total, participants completed 11 blocks of 64 trials each, for 704 trials overall.

Data analysis. The modelling mirrored that from Experiment 1, with the addition of stimulus type as another independent variable. Because of this, we fit the same three models (linear, logarithmic, power) to responses to each of the stimulus types (symbolic and nonsymbolic). To be able to investigate previous-trial effects between stimulus types, we added an interaction term to the previous-trial model that allowed for a differential effect of "same type" or "different type" previous trials.

RESULTS AND DISCUSSION

A total of 114 trials (2%) had to be removed, as participants had not given an answer or answered too quickly (i.e., in less than 500 ms). A further 83 trials (1.5%) were excluded as outliers (see the Data Analysis section in Experiment 1). Of the remaining data, trials that contained symbolic stimuli and trials that contained nonsymbolic stimuli were each separately fitted to three models (linear, logarithmic, power model; see Figure 2). For the symbolic stimuli, the best fit was again a linear model, $y = 1.06x - 8.23$; $R^2 = .99$. The same was true for the nonsymbolic stimuli, although the model was markedly different, $y = 0.69x + 17.88$, and the fit was not as good, $R^2 = .91$. The data were fitted better when the model included the previous magnitude as a predictor with a positive weight ($\Delta AIC = -7.65$, $b = 0.0344$). Introducing an additional interaction between previous trial magnitude and previous trial type improved the fit marginally ($\Delta AIC = -2.01$) and revealed that the weight for previous trials was somewhat smaller when the trial type was different to the current trial ($b_{diff} = 0.0041$, $b_{same} = 0.0380$).

A 2 × 2 × 3 (Trial-type [symbolic, nonsymbolic] × Block-type [mixed, homogenous] × Oddball [up, down, no oddball]) factor repeated-measures ANOVA on relative error indicated only one interaction: Trial-type × Oddball ($F[2, 14] = 19.26$, $p_{GG} = .003$, $e_{GG} = .53$). All other effects were nonsignificant ($p > .12$ in each instance). Post-hoc t tests gave only tentative evidence, as only comparing upwards oddballs with range-matched regular trials gave some indication of an effect, $t(7) = -3.82$, $p = .007$, $p > .2$ in all other instances.

To summarise, our results from Experiment 2 also showed that ruler-based responses to symbolic magnitudes (numbers) were almost perfectly linear and veridical. Responses to nonsymbolic magnitudes showed the characteristic underestimation for relatively large numbers,

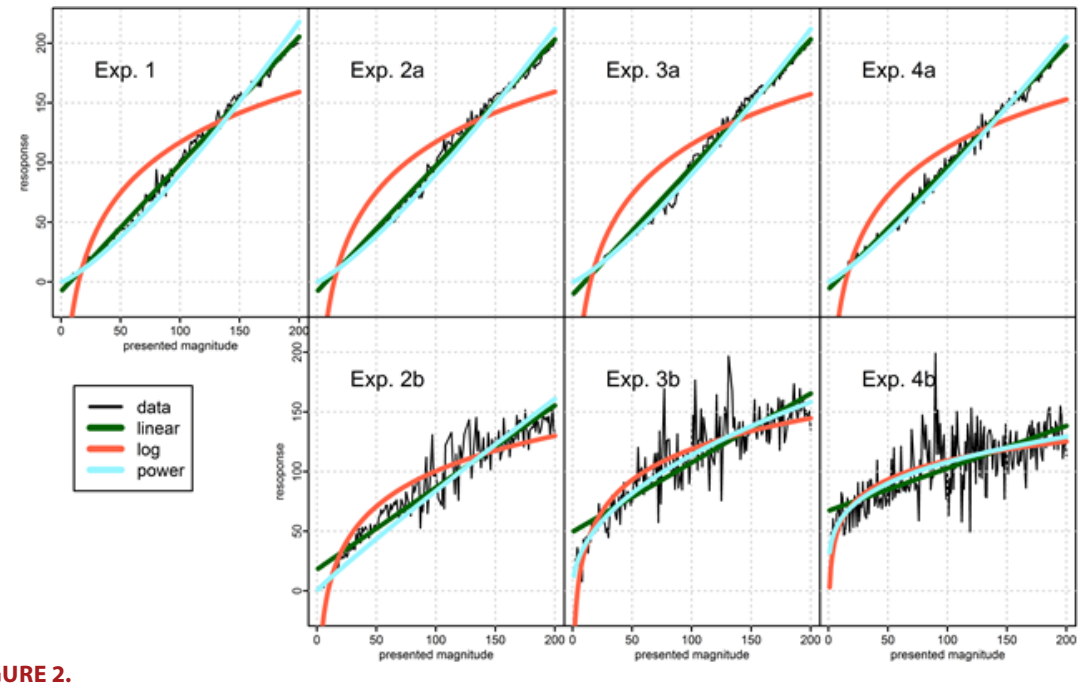

FIGURE 2.

Data from Experiments 1-4, with fitted linear, logarithmic, and power models. Left to right = Experiments 1-4; Top row = symbolic stimuli. Bottom row = nonsymbolic stimuli. For details on the experiments, see Table 1.

but, interestingly, still were fit better by a linear function than a logarithmic or power function. Additionally, larger magnitudes displayed in previous trials correlated with slightly larger responses on a given current trial.

Experiment 3: Symbolic and Nonsymbolic Magnitude on Response Bars with Symbolic and Nonsymbolic Endpoints

A potential drawback of Experiment 2 was its use of numeric endpoints, which means that one could argue that the version of the task we employed, and thus the output measure, was in fact not really nonsymbolic; that is, the stimuli would be compared to symbolic magnitudes (the endpoints) in order to find the correct location on the response bar. To remedy this, we conducted Experiment 3, in which we used nonsymbolic magnitudes as endpoints to the response bar. In all other respects, the experiment was identical to Experiment 2. Thus, we also expected the results to be largely similar to those in Experiment 2.

METHODS

Again, we recruited eight participants from the same pool as in Experiment 2 (aged between 20 and 29 years, M_{age} = 23.9; 6 females). As in Experiment 2, all participants indicated the position of a symbolic or nonsymbolic stimulus on a response bar. The only difference to Experiment 2 was that the endpoints of the response bar for the nonsymbolic stimulus now were a single dot on the left and a cloud of 200 dots on the right instead of Arabic digits (see Figure 1). These endpoints were always the same (i.e., not rendered anew for each trial) and were rendered using the same script that was used for the stimuli. Participants were again instructed to click on where they thought the stimulus belonged in the response bar, based on the number of dots in it.

RESULTS AND DISCUSSION

We removed 78 trials (1.3%) because of a lack of a valid answer or for being too quick, and 72 trials (1.3%) as outliers. The remaining data were modelled in the same fashion as in Experiment 2, indicating a linear model as the best, and once again near perfect, fit for symbolic stimuli, $y = 1.07x - 10.74$; $R^2 = .99$. Responses to nonsymbolic stimuli were fit best by a power model, $y = 12.66 \times x$^0.48, which explained 82% of the variance. The data were not fit better when previous number was included as a predictor, ΔAIC = 0.83, $b = 0.0175$, but slightly better when accounting for previous number split up by previous trial type, ΔAIC = −1.66, $b_{diff} = -0.0283$, $b_{same} = 0.0233$. The usual 2 × 2 × 3 ANOVA on relative error revealed no interactions between factors ($p > .13$ in each instance) and no main effects either ($p > .16$ in each instance). However, when oddball trials were tested against magnitude-matched nonoddball trials, this showed a significant difference between upwards oddball trials with nonsymbolic stimuli, $t(7) = -3.90$, $p = .006$, while other differences were not significant when correcting for multiple comparisons (downwards oddballs, nonsymbolic: $t[7] = 2.55$, $p = .038$; upwards symbolic: $t[7] = -2.61$, $p = .035$; downwards symbolic: $t[7] = 0.86$, $p = .421$), although they all pointed in the same direction: Oddball trials tended to err more towards the middle than other trials in the same number range, consistent with the fact that previous trials had a positive weight in the model.

Again, we found a virtually veridical response function to symbolic magnitudes, with a notable underestimation and previous-trial

dependency present for responses to nonsymbolic magnitudes. This was fit best by a power model. Responses to nonsymbolic stimuli were generally not predicted as well by the actual magnitude presented, as in Experiment 2 (see Table 2), perhaps indicating that the ruler-based task was more difficult with nonsymbolic endpoints. Despite the fact that the preferred model was a different one compared to Experiment 2B, with respect to other key features—for example, over-/underestimation for small/large numbers, respectively—the response function was quite similar to the response function in Experiment 2B (see Figure 2).

Experiment 4: Symbolic and Nonsymbolic Magnitude with Left/Right Flipped Endpoints

In Experiment 4, we wanted to preclude that what participants had been giving were stereotyped responses. To prevent them from using such a strategy, we slightly increased the difficulty of the task and introduced another input-response mapping by (a) flipping randomly the response bar in half of the trials, thus displaying the largest magnitude on its left side and the smallest magnitude on the right and (b) randomly varying the starting position of the adjustment mark in each trial. In essence, this experiment was done to preclude motor learning of responses as a confound, so we again expected a similar pattern of responses as we found in Experiments 2 and 3.

Flipping the response bar also reversed the standard right-large relationship between space and magnitude that is encountered in many everyday situations (see, e.g., Dehaene et al., 1993; Fischer & Shaki, 2015). While we would not expect these associations to bias responses in our task when they are congruent with the response bar, finding a similar response function with a flipped response bar would speak to the robustness of our findings. There may also be hemispheric differences in the representation of space: Consistently with this notion, van der Lubbe, Schölvinck, Kenemans, and Postma (2006) reported a more finely-grained spatial resolution for the left field. Such a mechanism could contribute to both the higher observed variability, and the "flatter" response function for larger numbers, if these are always responded to in the right field.

METHODS

Eight participants from the same pool as in Experiments 2 and 3 (aged between 20 and 32 years, M_{age} = 25.7; 4 females) took part in the experiment. The task was mostly the same as in Experiment 3, but in 50% of the trials (randomly distributed within each block), the response bar was flipped, such that the lower end was on the right and the higher end was on the left. Additionally, the starting position for the adjustment mark was randomized, such that the mark was equally likely to appear anywhere on the response bar at the start of each trial. Participants were given 4 s to respond.

RESULTS AND DISCUSSION

We had to remove 41 trials (0.7%) for a lacking valid answer or being outside the allowed response times, and 99 trials (1.8%) as outliers. Modelling the responses to symbolic stimuli, the best model was a linear fit of $y = 1.02x - 6.26$, explaining 99% of the variance (see Figure 2). The responses to nonsymbolic stimuli were once again fit best by a power model ($y = 32.50 \times x\text{^}0.26$; $R^2 = .54$). Including the magnitude of the previous trial did not improve the fit (ΔAIC = 0.99), although including an interaction term of Previous-trial Magnitude × Previous-trial Type did (ΔAIC model with interaction vs. simple model: −5.19), indicating that previous trials actually had a negative weight if they were of a different type ($b_{diff} = -0.0777$), and much smaller negative weight when they were of the same type ($b_{same} = -0.0061$). The standard 2 × 2 × 3 repeated-measures ANOVA on relative error revealed a main effect of oddball, $F(2, 14) = 20.20$, $p_{GG} = .003$, $e_{GG} = .51$, but no other main effect ($p > .6$ in each instance), with a statistically significant interaction of Block-type × Trial-type, $F(1, 7) = 7.23$, $p = .031$, and the three-way interaction Oddball × Block-type × Trial-type, $F(2, 14) = 7.89$, $p_{GG} = .025$, $e_{GG} = .51$. No t test comparing oddball trials to range-matched regular trials indicated any significant difference ($p > .6$ in each instance).

Once again, responses to symbolic stimuli were fit best by a linear function close to unity, while responses to nonsymbolic stimuli resembled a power function. That is, results were similar to those obtained in Experiments 2 and 3, even when the experiment prevented participants from learning mouse movements as opposed to considering the desired location of the click. The fact that these results once again show an underestimation and a relatively shallow response function for magnitudes larger than the mean (100) also excludes the possibility that the smaller increments in responses for higher numbers are confounded with differences in spatial resolution of the left and right field (van der Lubbe et al., 2006).

Discussion of Experiments 2-4

Two findings appeared robustly in all experiments: Responses to symbolic stimuli had an almost perfectly linear shape, and responses to nonsymbolic stimuli tended to overestimate low magnitudes and underestimate higher magnitudes. The former is readily explained by the proficiency of participants: Despite having limited time available, the task was overall not very difficult when the stimuli were symbolic magnitudes. Responses to nonsymbolic stimuli, on the other hand, mirror known patterns of nonlinearity (Dehaene, 2003; van Oeffelen & Vos, 1982) that are robust across different variants of the task. Notably, the responses were in most cases fit better by a power function than by a logarithmic function (see Table 2; note that the linear model outperformed both others in Experiment 2), which is corroborated by the fact that when plotted in a log-log graph, the functions appear roughly equivalent (see Figure 3). These results are also quite compatible with Dehaene's notion of a logarithmic number line (Dehaene, 1992) with an *output grid* transformation (Izard & Dehaene, 2008), in which the mapping from somewhat categorical internal representation to responses can be stretched or compressed through calibration. Since in each case, previous trials (at least of the same input type) influenced the current trial, they are also compatible with a dynamic encoding mechanism like the one of Cicchini et al. (2014), who proposed the shape of the MNL to be due to each trial representing a weighted

combination of an estimates of the current and of previously presented magnitudes. Notions such as a linear response function with scalar variability, however, can be dismissed; while the variability does reach a ceiling of sort in ruler-based tasks (see Figure 3), this model would predict underestimation for higher numbers, but not overestimation for smaller numbers.

For the explanation of dynamic encoding mechanisms based on previous stimuli to be plausible, it is a prerequisite that sequential dependencies exist, which was the second question our experiments sought to answer. Indeed, this was the case, as the fitting of previous-trial models found a positive relationship between the magnitude of the previous trial and the response in the current trial, which was consistent with the analysis of oddball trials. Note that the weights for previous-trial magnitude differed between experiments: A weight of over .2, as we saw in Experiment 2, is rather large when compared to the literature (Cicchini et al., 2014), whereas such effects were small in Experiment 3 and nonexistent in Experiment 4. These findings from the latter two experiments also indicate that in our design, effects of previous trials alone were unlikely to be the sole or even main cause of the response function's shape, as it was similarly nonlinear here as in Experiment 2.

EXPERIMENT 5: DOES CALIBRATION GENERALIZE BETWEEN RESPONSES TO SYMBOLIC AND NONSYMBOLIC STIMULI?

In our final experiment, we sought to explore whether the responses to nonsymbolic stimuli could be linearized (i.e., correctly calibrated) through feedback, and if this could be done for responses to symbolic and nonsymbolic stimuli independently. This served the broader purpose of providing a stronger test of whether the observed shape of the

TABLE 2.
Linear, Logarithmic, and Power Models Fit to Data from All Experiments

Exp	Condition	lin	log	Power	Previous trial weight	Previous trial weight by trial type
1	Symbolic stimuli	**1.07*x* − 7.90** **R^2 = .99**	60.85 × log(*x*) − 163.28 R^2 = .80	0.27 × *x*^1.27 R^2 = .98	***b* = −0.0139** **ΔAIC = −2.71**	-
2a	Symbolic stimuli	**1.06*x* − 8.23** **R^2 = .99**	60.23 × log(*x*) − 161.93 R^2 = .81	0.32 × *x*^1.23 R^2 = .99	*b* = 0.0344	**b_{same} = 0.0380** **b_{diff} = 0.0041**
2b	Nonsymbolic stimuli	**0.69*x* + 17.88** **R^2 = .91**	42.67 × log(*x*) − 96.12 R^2 = .87	1.10 × *x*^0.94 R^2 = .89	ΔAIC = −7.65	**ΔAIC = −9.66**
3a	Symbolic stimuli	**1.07*x* − 10.74** **R^2 = .99**	60.15 × log(*x*) − 163.82 R^2 = .79	0.31 × *x*^1.23 R^2 = .99	*b* = 0.0175	**b_{same} = 0.241** **b_{diff} = −0.0283**
3b	Nonsymbolic stimuli	0.58*x* + 49.50 R^2 = .77	37.34 × log(*x*) − 53.17 R^2 = .80	**12.66 × *x*^0.48** **R^2 = .82**	ΔAIC = 0.83	**ΔAAC = v1.66**
4a	Symbolic stimuli	**1.02*x*–6.26** **R^2 = .99**	58.02 × log(*x*) − 154.57 R^2 = .79	0.43 × *x*^1.16 R^2 = .99	*b* = −0.0193	**b_{same} = −0.0061** **b_{diff} = −0.0777**
4b	Nonsymbolic stimuli	0.36*x* + 67.13 R^2 = .50	23.04 × log(*x*) + 3.09 R^2 = .52	**32.50 × *x*^0.26** **R^2 = .54**	ΔAIC = 0.99	**ΔAIC = −5.19**
5a	response bar, pre-FB	0.65*x* + 32.13 R^2 = .84	41.03 × log(*x*) − 80.84 R^2 = .78	**4.69 × *x*^0.66** **R^2 = .85**	***b* = 0.0424** **ΔAIC = −11.43**	-
	response bar, FB	0.82*x* + 17.59 R^2 = .93	59.92 × log(*x*) − 163.15 R^2 = .88	**2.97 × *x*^0.77** **R^2 = .94**	***b* = 0.2398** **ΔAIC = −314.70**	
	response bar, post-FB	**0.79*x* + 24.86** **R^2 = .93**	49.29 × log(*x*) − 109.54 R^2 = .91	1.56 × *x*^0.91 R^2 = .91	***b* = 0.0440** **ΔAIC = −26.9**	
5b	Number response, pre-FB	0.50*x* + 16.13 R^2 = .88	30.91 × log(*x*) − 67.22 R^2 = .83	**1.91 × *x*^0.77** **R^2 = .89**	***b* = 0.0579** **ΔAIC = −35.32**	
	Number response, FB	**0.84*x* + 8.59** **R^2 = .94**	59.77 × log(*x*) − 170.18 R^2 = .86	2.10 × *x*^0.83 R^2 = .94	***b* = 0.3258** **ΔAIC = −463.97**	
	Number response, post-FB	0.64*x* + 21.63 R^2 = .89	40.20 × log(*x*) − 87.54 R^2 = .86	**2.10 × *x*^0.81** **R^2 = .90**	***b* = 0.0648** **ΔAIC = −39.36**	

Note. Bold indicates the best-fitting model. ΔAIC given relative to the simplest model. FB = feedback, AIC = Akaike information criterion (Akaike, 1974; Burnham & Anderson, 2004). For details on the conditions see Table 1 and Figure 1.

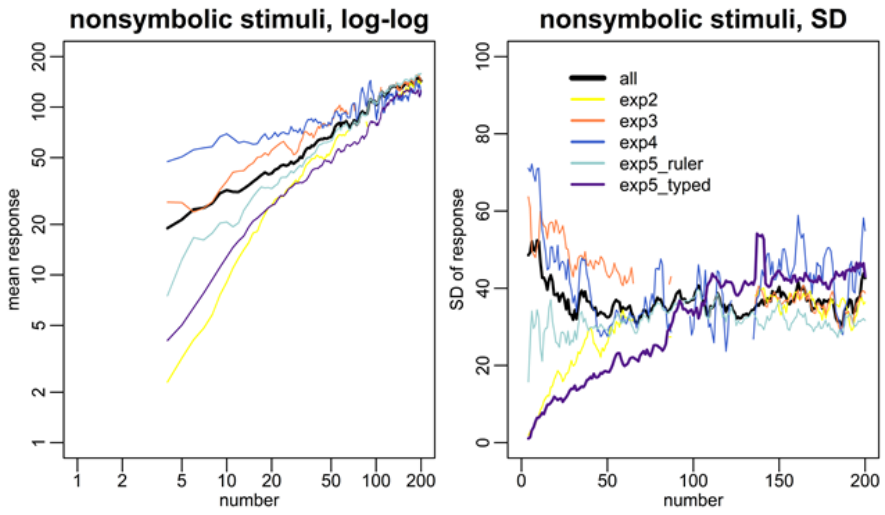

FIGURE 3.

Exploring variability in responses to nonsymbolic stimuli. Left = responses in a log-log plot. A power function would be linear in such a plot; Right = SD by number, for all experiments.

responses is better understood as a phenomenon of mapping input-to-representation or representation-to-response by comparing sequential effects between stimulus types to calibration effects between tasks. To this end, we conducted an experiment in which participants conducted both a ruler-based task like the one described in Experiment 3, but with randomised starting positions, and a classic numerosity-judgement task in which they typed in the estimated number of dots for each stimulus (here called the typed-response task, see the Our Study section in the Introduction). Both tasks included feedback blocks to allow participants to calibrate their responses, with feedback being either (a) veridical, (b) systematically lower, or (c) systematically higher than the actual magnitude. This feedback was given independently for each task, thus allowing us to investigate not only the effect of feedback in each task on responses in the same task, but also its influence on responses in the other task. Knowing whether feedback effects would generalize between tasks with the same stimuli but different responses would allow us to see whether calibration through feedback would influence the mapping from the stimulus to the internal magnitude estimate, or rather something else (most likely the mapping from the internal estimate to the response).

Methods

PARTICIPANTS

Due to the larger number of conditions, as well as to enable us to test our predictions more conclusively in a single experiment, we increased the number of participants to a total of 36 (M_{age} = 24 years, age range 18 to 36; 26 females). These were again recruited from the same participant pool as in the Experiments 2-4.

PROCEDURE

All participants completed two numerosity-estimation tasks: a ruler-based task like the one used in Experiment 3, and a simple typed-response task, in which participants entered their estimate via a standard computer keyboard. Stimuli were the same nonsymbolic stimuli as in Experiments 2, 3, and 4. Participants were again allowed 4 s to respond in the ruler-based task. The same time limit was set for the typed-response task; however, due to a programming error in the MATLAB program, the timer was reset when participants started typing, in effect giving them 4 s to start typing, and 4 s thereafter (mean response times in this task were 2570 ms, with an *SD* of 702 ms).

For both the ruler-based task and the typed-response task, we included one middle block each where participants received feedback about the correctness of their response. Thus, each participant completed three blocks—prefeedback, feedback, postfeedback—for both the ruler-based task and the typed-response task. This feedback could be either veridical (reflecting a 1-to-1 mapping of stimulus to the position on the response bar or the number to be entered), or distorted by an amount of either +15 or −15. This feedback was consistent across all trials within one block, but varied independently for the two tasks, and feedback conditions were counterbalanced between participants so that one third of participants was assigned to each feedback condition in each task. The order of tasks was also counterbalanced. Feedback in the typed-response task was presented as a red number appearing on the screen once the participant hit the return key, to the right of the number the participant had just entered. In the ruler-based task, feedback was given through a red square mark of the same size as the adjustment mark, appearing on the horizontal bar at the location corresponding to the correct magnitude (or the correct magnitude +15 / −15). Participants received 10 practice trials with feedback on a response bar right before the ruler-based task, along with instructions on

how the feedback worked. In the −15 and +15 feedback blocks, stimuli were restricted so that feedback fell in the 15-85 and 115-185 ranges, such that feedback was never too close to the bounds of the response bar, or on the "wrong" side of the mid-point, so as to not make the manipulation too obvious (see Barth & Paladino, 2011). No oddball blocks were included. Each block consisted of 120 trials, resulting in a total of 720 trials per participant.

Results

Overall, 860 trials had to be removed because no valid answer had been given (3.3%) or for being outside the allowed response times. A further 330 trials (1.3%) were excluded as outliers. One participant had to be removed from analysis for not understanding the task. During debriefing, all participants were asked if they had considered the feedback to be accurate. Six out of 36 participants said that they had not, which included one participant who had received veridical feedback. Following this, participants received information about whether they had in fact received veridical or distorted feedback. Removing the six participants who had believed the feedback to be inaccurate during the experiment and prior to debriefing from analysis did not substantially change the results (the following analyses include those participants).

Similar to the analysis of the other experiments, we first investigated how to best model the data from each task. We fit separate models for responses given in blocks prior to feedback, during feedback, and after feedback had been presented. In the ruler-based task, prefeedback data were fitted best by a power model, $y = 4.69 \times x^{0.66}$; $R^2 = .85$, which fit marginally better than a linear model, $R^2 = .85$ to .84, with postfeedback data being better fit linearly, $y = 0.79x + 24.86$; $R^2 = .93$. As was the case in most previous tasks, the data were fit better when including previous trial magnitude as predictor, both before ($\Delta AIC = -11.43$; $b = 0.0424$) and after feedback ($\Delta AIC = -26.9$; $b = 0.0440$). During feedback, a power function gave the best fit, $y = 2.97 \times x^{0.77}$; $R^2 = .94$, and, unsurprisingly, there was a very strong previous-trial effect, $\Delta AIC = -314.70$; $b = 0.2399$. In all three phases, the differences in goodness of fit between the power model and the linear model were marginal, with the logarithmic function doing substantially worse (see Table 2). For the typed-response task, the data were fit best by a power function, in both the prefeedback, $y = 1.91 \times x^{0.77}$; $R^2 = .89$, and postfeedback blocks, $y = 2.10 \times x^{0.81}$; $R^2 = .90$. Including the previous trial also improved the fit, $\Delta AIC = -35.32$, $b = 0.0579$ for prefeedback and $\Delta AIC = -39.36$, $b = 0.0648$ for postfeedback. With feedback, the linear model did best, $y = 8.59 + 0.84x$, $R^2 = .94$, but only marginally better than the power model (see Table 2), and showed a very strong effect of previous trials, $\Delta AIC = -463.97$; $b = 0.3258$. Again, model fits were virtually equally good for linear and power functions, but worse for logarithmic fits.

To investigate further the effects of feedback on relative error, and to see if the feedback in the respective other task mattered at all, we further conducted a mixed ANOVA with two between-subjects factors of symbolic feedback, as well as nonsymbolic feedback (3 levels each: +15, −15, 0), as well as the within-subject factor of trial type (ruler-based response or typed response). Unsurprisingly, this revealed a main effect of trial-type, $F(1, 27) = 33.24$, $p < .001$, but, perhaps surprisingly, no interaction of Trial Type and either Feedback factor (Trial Type × Verbal Feedback: $F[2, 27] = 1.35$, $p = .275$; Trial Type × Nonverbal Feedback: $F[2, 27] = 0.30$, $p = .746$), which would have indicated a general impact of feedback on the response. Post-hoc two-sample t tests revealed that no response was influenced by feedback in the other task ($p > .24$ in each instance), but also that there was only very weak evidence, if any, for an effect of same-task feedback (ruler-based task, feedback +15 vs. 0: $t[20.20] = 0.92$, $p = .370$; feedback −15 vs. 0: $t[21.26] = -0.56$, $p = .580$; typed response task, feedback +15 vs. 0: $t[20.95] = 0.38$, $p = .706$; feedback −15 vs. 0: $t[15.94] = -2.57$, $p = .021$). This is supported by visual inspection (see Figure 4).

Discussion

We conducted Experiment 5 for two main purposes: to be able to compare responses obtained from different tasks with nonsymbolic stimuli and to investigate the effect of different types of feedback on the shape of the response functions. The shape we found for the typed response tasks was indeed different from what was found in previous experiments—note that the overestimation for small numbers, found in all response functions to nonsymbolic stimuli, was not found here. The best fitting model was a power function (albeit not by much, see Table 2), as has been proposed by several authors (Izard & Dehaene, 2008; Krueger, 1972; Nieder & Miller, 2003). We can also see (see Figure 3) that this was the only task where variability increased almost linearly with stimulus magnitude—a typical feature of magnitude estimation. In the ruler-based task, we found a similar, albeit somewhat steeper and more linear response function than in previous experiments, even before any feedback had been given. It is possible that the added practice trials had an effect here.

With regards to the feedback we introduced to help participants calibrate their responses, we can see that response functions were markedly steeper after feedback was given. Feedback linearized the responses somewhat, although the response function was still quite far from a veridical function, and it was not clear whether this was still the case once feedback was removed. Still, ruler-based responses were also fit better in blocks after feedback had been given (but was not given any longer). This was not true of typed responses. Importantly, feedback effects did not transfer between tasks, indicating that any calibration was task-specific.

GENERAL DISCUSSION

Two classic findings were reproduced in our experiments: ruler-based responses to symbolic magnitude stimuli exhibited a linear shape in adult participants (Anobile, Cicchini, & Burr, 2012; Siegler & Opfer, 2003), and typed responses to nonsymbolic magnitudes were fit best by a power function (Izard & Dehaene, 2008; Krueger, 1972; Nieder & Miller, 2003). Still, it should be noted that the predictions of these models were remarkably similar, as exponents of the power functions tended to be close to 1 and intercepts of the linear functions close to 0. We also investigated the ruler-based responses to nonsymbolic mag-

nitudes, which were best fit by a power function in three out of four experiments.

What remains to be explained is the cause of the shape of these responses. Our experiments (see Table 1 and Figure 1) were designed to investigate three questions, namely, whether responses would depend on previous trials and whether this could explain the response function shape, whether response functions for different input (in the same task) and output (with the same input type) would differ, and whether giving feedback to calibrate the response would linearize the response function.

With regard to the first question of whether these were dynamic effects brought about by the effects of previous trials, we found the strongest effects to be mostly static. There were previous-trial effects, but these were not strong enough to explain much of the variance—and, importantly, not robust to variations (such as flipping the response bar or randomising the starting position in Experiment 4) that the shape of the response function was robust to: The previous-trial effect we found virtually disappeared when different types of stimuli were presented on the current and previous trial, respectively, and disappeared entirely or was even reversed with a randomly flipped response bar in Experiment 4. This touches on another debate, the question of whether a single semantic representation is underlying the processing of magnitudes of different notations (Dehaene, 1992; Walsh, 2003), or whether some magnitudes may be explained as sensory features (Arrighi, Togoli, & Burr, 2014). Our data do not speak strongly against either hypothesis, although a strong version of an underlying magnitude representation would probably predict less distinct interaction patterns depending on whether the stimulus type was the same or different in consecutive trials.

Regarding the second question of telling apart the different roles of input and output mapping in creating the typical shape of response functions: While a lot of research has focused on the different properties of symbolic and nonsymbolic magnitude processing, the contribution of output format has not been investigated as much beyond the question of methodological confounds (Barth & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011). Although using a ruler-based task may have some problems, it is also quite clear that it provides the possibility of a useful, very direct instrument of measuring responses to symbolic and nonsymbolic stimuli in a comparable way. In this study, we have laid out some key differences between ruler-based task and typed-response tasks, finding that both the mean responses (see Figure 2) and the measured variability (see Figure 3) differ between the two. However, we also found that effects of input type are much more pronounced than effects of the output measure (compare top and bottom rows of Figure 2).

We also used the two different tasks (a ruler-like response bar without a symbolic component and a typed response) to investigate whether giving feedback on one type of response would have any effect on responses in another type of response, which we did in Experiment 5, thus trying to answer the third question: Can we linearize responses to nonsymbolic magnitudes through calibration, and if so, what do we calibrate? We found some, but not very strong linearization (see Figure 4), and no impact of feedback in one task on responses in the other task. Since both tasks used identical stimuli, we conclude that what is calibrated is not the input mapping from stimulus to internal

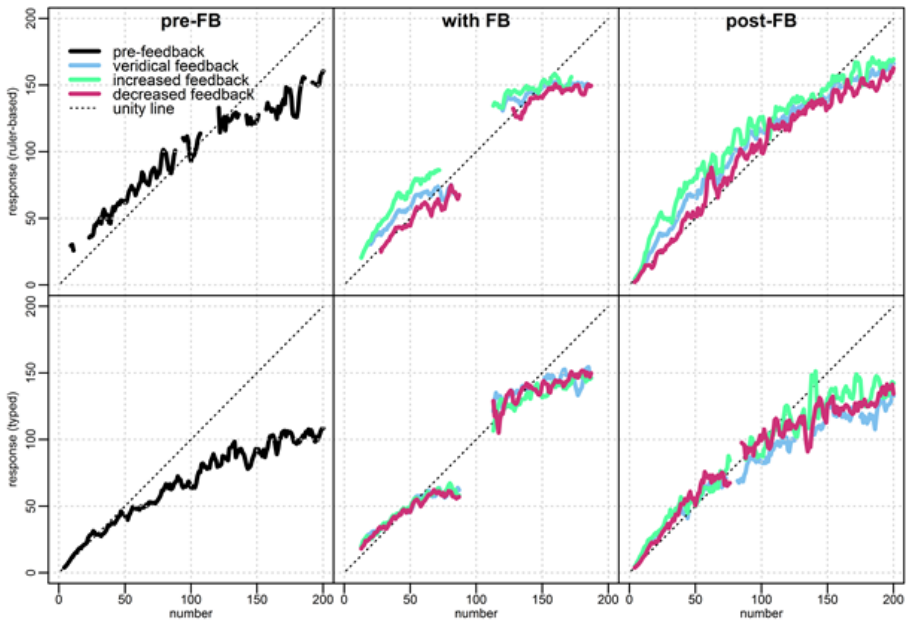

FIGURE 4.

Responses in Experiment 5, by feedback. Prefeedback panels show data from all groups. Dashed lines depict veridical performance. Top row = ruler-based task. Bottom row = typed number response; Left column = before feedback was introduced; Middle column = feedback blocks; Right column = post-feedback blocks without feedback.

representation, but the output mapping from internal representation to response. We also see that this is not sufficient to completely linearize the response. Our data also allow the conclusion that the nonlinear shape is arguably a function of input-to-representation mapping mechanisms, as it is seen in both tasks, and we have demonstrated that the ruler-based task in itself does not lead to a nonlinear response function.

We should also offer a word of caution on our stimuli. As mentioned in the Experiment 2 section, several sensory stimulus features were somewhat informative about the numerosities presented. Indeed, responses based solely on the circumference of the clouds would have allowed a participant to judge the magnitudes quite well (explaining 59% of the variance, assuming perfect perception of this aspect). Participants scored substantially higher than this, however (see Table 2), so that any reliance on purely sensory cues would have had to be a combination of several cues (see, e.g., Gebuis et al., 2014). However, the result of such a combination of cues would be a sort of nonsymbolic magnitude, and while the discussion about how to define nonsymbolic magnitude is an interesting one, it is beyond the scope of this paper. Another caveat is the fact that our stimuli were always displayed centrally. This would have directed participants' attention to the centre of the screen, as well as providing a potential point of reference, and thereby could have biased responses towards the middle of the response bar. We accounted for this by validating the method with Arabic digits (Experiment 1), but of course the visual processing of Arabic digits is much easier than that of clouds of dots—thus, it is possible for such a mechanism to affect nonsymbolic, but not symbolic trials. Of course, typed responses would not be affected by this mechanism and produced the same nonlinear response function (Experiment 5).

We conclude that the nonlinear shape of the number line is largely robust to calibration even through direct, veridical feedback, and that features of both stimuli and output measures contribute to it. Small effects of feedback do not transfer to different response types using the same stimuli, indicating that calibration affects the mapping from representation to output. Serial dependencies exist between trials, but are too weak to explain the shape of the responses in such tasks.

ACKNOWLEDGEMENTS

We thank Alistair Yousefi and Larissa Brockmann for their help collecting data. KKK was supported by a PhD scholarship as part of a grant to VHF by the Deutsche Forschungsgemeinschaft (DFG) within the international graduate research training group on Cross-Model Interaction in Natural and Artificial Cognitive Systems (CINACS; DFG IKG-1247).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. In T. Kailath, D. Q. Mayne, & R. K. Mehra (Eds.), *IEEE Transactions on Automatic Control AC-19* (pp. 716–723). New York, NY: The Institute of Electrical and Electronics Engineers.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition, 122*, 454–459. doi: 10.1016/j.cognition.2011.11.006

Arrighi, R., Togoli, I., & Burr, D. C. (2014). A generalized sense of number. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20141791. doi: 10.1098/rspb.2014.1791

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: evidence against a representational shift. *Developmental Science, 14*, 125–135. doi: 10.1111/j.1467-7687.2010.00962.x

Burnham, K. P., & Anderson, R. P. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304. doi: 10.1177/0049124104268644

Cantlon, J. F., Cordes, S., Libertus, M. E., & Brannon, E. M. (2009). Comment on "Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures." *Science, 323*(5910), 38. doi: 10.1126/science.1164773

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 7867–7872. doi: 10.1073/pnas.1402785111

Cipolotti, L., & Butterworth, B. (1995). Toward a multiroute model of number processing: Impaired number transcoding with preserved calculation skills. *Journal of Experimental Psychology: General, 124*, 375–390. doi: 10.1037/0096-3445.124.4.375

Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic Bulletin & Review, 18*, 331–338. doi: 10.3758/s13423-011-0059-z

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*, 698–707. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11848588

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1–42. doi: 10.1016/0010-0277(92)90049-N

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences, 7*, 145–147. doi: 10.1016/S1364-6613(03)00055-X

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371–396. doi: 10.1037/0096-3445-.122.3.371

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2009). Response to comment on "Log or linear? Distinct intutions of the number scale in Western and Amazonian indigene cultures." *Science, 323*(5910), 38. doi: 10.1126/science.1164878

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western Amazonian indigene cultures. *Science, 320*(5880), 1217–1220. doi: 10.1126/science.1156540

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*,

487–506. doi: 10.1080/02643290244000239

Fechner, G. T. (1860). Elemente der Psychophysik. Leipzig, Germany: Breitkopf und Haertel.

Fischer, M. H., & Shaki, S. (2015). Two steps to space for numbers. *Frontiers in Psychology, 6*:612. doi: 10.3389/fpsyg.2015.00612

Gebuis, T., Gevers, W., & Cohen Kadosh, R. (2014). Topographic representation of high-level cognition: Numerosity or sensory processing? *Trends in Cognitive Sciences, 18*, 1–3. doi: 10.1016/j.tics.2013.10.002

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods, 43*, 981–986. doi: 10.3758/s13428-011-0097-5

Gebuis, T., & Reynvoet, B. (2012). The role of visual information in numerosity estimation. *PloS One, 7*, e37426. doi: 10.1371/journal.pone.0037426

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112. doi: 10.1007/BF02289823

Haubensak, G. (1992). The consistency model: A process model for absolute judgements. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 303–309. doi: 10.1037/0096-1523.18.1.303

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70. doi: 10.2307/4615733

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition, 106*, 1221–1247. doi: 10.1016/j.cognition.2007.06.004

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, D. (2007). What's new in Psychtoolbox-3? *Perception 36*, 1-16. doi: 10.1068/v070821

Krueger, L. E. (1972). Perceived numerosity. *Perception & Psychophysics, 11*, 5–9. doi: 10.3758/BF03212674

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude"—The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences, 40*, e164. doi: 10.1017/S0140525X16000960

Link, T., Huber, S., Nuerk, H. C., & Moeller, K. (2014). Unbounding the mental number line—New evidence on children's spatial representation of numbers. *Frontiers in Psychology, 4*:1021. doi: 10.3389/fpsyg.2013.01021

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111*, 1–22. doi: 10.1037/0096-3445.111.1.1

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition, 44*, 107–157. doi: 10.1016/0010-0277(92)90052-J

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing: Evidence from dyscalculia. *Brain and Cognition, 4*, 171–196. doi: 10.1016/0278-2626-(85)90069-7

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

Nieder, A. (2016). The neural code for number. *Nature Reviews Neuroscience, 17*, 366–382. doi: 10.1016/B978-0-12-385948-8.00008-6

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron, 37*, 149–157. doi: 10.1016/S0896-6273(02)01144-3

Opfer, J. E., Siegler, R. S., & Young, C. J. (2011). The powers of noise-fitting: Reply to Barth and Paladino. *Developmental Science, 14*, 1194–1204. doi: 10.1111/j.1467-7687.2011.01070.x

Parducci, A., & Perret, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology, 89*, 427–452. doi: 10.1037/h0031258

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237–243. doi: 10.1111/1467-9280.02438

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153–181. doi: 10.1037/h0046162

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology, 78*, 392–402. doi: 10.2307/1420573

Trick, L. M. (2008). More than superstition: Differential effects of featural heterogeneity and change on subitizing and counting. *Perception & Psychophysics, 70*, 743–760. doi: 10.3758/PP.70.5.743

Trick, L. M., & Pylyshyn, Z. M. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101*, 80–102. doi: 10.1037/0033-295X.101.1.80

van der Lubbe, R. H. J., Schölvinck, M. L., Kenemans, J. L., & Postma, A. (2006). Divergence of categorical and coordinate spatial processing assessed with ERPs. *Neuropsychologia, 44*, 1547–1559. doi: 10.1016/j.neuropsychologia.2006.01.019

van Oeffelen, M. P., & Vos, P. G. (1982). A probabilistic model for the discrimination of visual number. *Perception & Psychophysics, 32*, 163–170. doi: 10.3758/BF03204275

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7*, 483–488. doi: 10.1016/j.tics.2003.09.002

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*, 130–137. doi: 10.1111/1467-9280.00120

RECEIVED 16.02.2017 | ACCEPTED 14.05.2018