This version of the article has been accepted for publication in *Psychological Research*, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at https://doi.org/10.1007/s00426-024-02044-6

Title: **Are Three Zebras More than Three Frogs: Examining Conceptual and Physical Congruency in Numerosity Judgements of Familiar Objects**

Mila Marinova [1,2,3]& Bert Reynvoet [2,3]

[1] Institute of Cognitive Science and Assessment, Department of Behavioural and Cognitive Sciences, Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Esch-Belval, Luxembourg.

[2]Brain and Cognition, KU Leuven, Leuven, Belgium

[3]Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Kortrijk Belgium.

Correspondence concerning this article should be addressed to Bert Reynvoet, Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. Tel: +32 56246177, bert.reynvoet@kuleuven.be

ORCID:

Mila Marinova: https://orcid.org/0000-0002-6875-7742

Bert Reynvoet: https://orcid.org/0000-0002-4898-2475

Abstract

Researchers in numerical cognition have extensively studied the number sense—the innate human ability to extract numerical information from the environment quickly and effortlessly. Much of this research, however, uses abstract stimuli (e.g., dot configurations) that are also strictly controlled for their low-level visual confounds, such as size. Nonetheless, individuals rarely extract numerical information from abstract stimuli in everyday life. Yet, numerical judgments of familiar objects remain poorly understood and understudied. In the current study, we examined the cognitive mechanisms underlying the numerical decisions of familiar objects. In two experiments, we asked adult participants (Experiment 1) and two groups of children (aged 7 – 9 years and 11 – 12 years, Experiment 2) to perform an animal numerosity task (i.e., "Which animal is more numerous?"), while the conceptual congruency (i.e., the congruency between an object's real-life size and its numerosity) and physical congruency (the congruency between the number of items and the total space they occupy on the screen) were manipulated. Results showed that the conceptual congruency effect (i.e., better performance when the animal with a larger size in real life is more numerous) and a physical congruency effect (i.e., better performance when the physically larger animal is more numerous) were present in adults and children. However, the effects differed across the age groups and were also a subject of developmental change. To our knowledge, this study is the first one to demonstrate that conceptual knowledge can interfere with numerosity judgements in a top-down manner. This interference effect is distinct from the bottom–up interference effect, which comes from the physical properties of the set. Our results imply that the number sense is not a standalone core system for numbers but is embedded in a more extensive network where both low-level and higher-order influences are possible. We encourage numerical cognition researchers to consider employing not only abstract but also familiar objects when examining numerosity judgements across the lifespan.

Introduction

The ability to extract numerical information rapidly and effortlessly from the environment is often referred to as number sense (Dehaene, 2011). This ability is already observed in infants (Xu & Spelke, 2000) and across several species in the animal kingdom (Nieder, 2021) and is therefore considered to be encapsulated in a core, innate system for number (Feigenson et al., 2004; Nieder & Dehaene, 2009). Despite its popularity, the idea of an innate core system for number and its role in developing and acquiring numerical and mathematical skills is contested, and there is no consensus yet about what the system represents (Clarke & Beck, 2020; Gebuis & Reynvoet, 2012; Leibovich et al., 2017; Lourenco & Aulet, 2023). For instance, it is debated whether number, as a property of a set, is extracted directly from the environment (i.e., as a primary feature) or is constructed based on non-numerical cues (e.g., size, surface area, density, etc.), or that its extraction depends depending on context and task characteristics (Clarke & Beck, 2020; Gebuis et al., 2016; Gennari et al., 2023; Laurenco & Aulet, 2023; Marinova et al., 2021). Another question, to be further clarified, is whether the results obtained in experimental settings – in which abstract stimuli are used (e.g., dot configurations) highly controlled for non-numerical features (e.g., the dots' size, density, circumference, etc.) - can be generalized to real-world decisions such as deciding how many cows there are in the meadow, or how many frogs there are in your pond (Odic & Oppenheimer, 2023; Reynvoet et al., 2019). Unlike the configuration of the dots, everyday objects such as animals or fruits possess in addition to external physical features (e.g., the physical size; the features typically controlled in former studies), also intrinsic features (i.e., conceptual knowledge about the object, including size). In the current study, we examined the influence of physical and conceptual knowledge on numerical judgments of familiar objects in adults (Experiment 1) and children (Experiment 2) at the beginning of primary school (aged 7 to 9 years), and children at the end of primary school (aged 11 to 12 years).

The Number Sense

In examining the number sense across the lifespan, researchers commonly use non-symbolic comparison tasks, in which individuals are instructed to numerically compare two visual sets of stimuli (e.g., a set of 10 dots vs a set of 20 dots) (Gebuis & Reynvoet, 2012; Halberda et al., 2012). Because of the natural correlation between numbers and other non-

numerical dimensions (e.g., if all dots are the same size, the more numerous dot configuration will also have higher total surface and density), scientists use various algorithms to ensure that decisions are based on the number (De Marco & Cutini, 2020; DeWind et al., 2015; Gebuis & Reynvoet, 2011; Guillaume et al., 2020; Halberda et al., 2008). Employing such algorithms enables teasing apart the non-numerical features from the numerosity of a set. Performance in a numerosity comparison task is typically reflected by a ratio effect – accuracy increases as the relative distance between two numerosities (n_1/n_2) increases (e.g., comparing 10 vs 20 dots is easier than 18 vs 20 dots). This discrimination ability is also subject of developmental changes: from childhood to adulthood, humans become progressively better and are able to distinguish more difficult ratios (Defever et al., 2013; Halberda et al., 2012; Piazza et al., 2018). The ratio effect is usually interpreted as evidence that comparison judgments are indeed based on numerical information and not on the perceptual features of the stimuli which are controlled in these studies.

However, the impact of the non-numerical features during numerosity decisions cannot be eliminated completely (Clayton et al., 2015; Reynvoet et al., 2021; Smets et al., 2015). Concretely, dot generation algorithms control for the non-numerical properties by creating at least two distinct types of trials – congruent (e.g., the numerically larger set occupies a larger total area) and incongruent (i.e., the numerically larger set occupies a smaller total area). Several studies have shown that participants' performance is not only described by a numerical ratio effect but also by a congruency effect: the congruent trials yield better performance than the incongruent trials (Fuhs & Mcneil, 2013; Gilmore et al., 2013; Reynvoet et al., 2021). Importantly, from a developmental point of view, the congruency effect decreases with age, indicating that younger children are more biased to non-numerical dimensions than older children and adults, although adults' performance is less but still affected by the non-numerical cues (Piazza et al., 2018; Smets et al., 2016; Starr et al., 2017)

The origin of this congruency effect remains under discussion (see Lourenco & Aulet, 2023 for a review). Some researchers argued that number is encoded indirectly and that the congruency effect is a by-product of combining non-numerical features (Gebuis & Reynvoet, 2012; Marinova et al., 2021). Alternatively, others have suggested that the congruency effect is the result of a response conflict between the independently processed numerical and non-numerical

information, where the latter needs to be inhibited in favor of a correct response (Piazza et al., 2018). Yet another possibility is that number and non-numerical features are processed holistically as integrated dimensions early in the visual stream (Lourenco & Aulet, 2023). Separating number from the other dimensions, required for numerosity decision tasks, recruits additional attentional and inhibitory mechanisms, resulting in a congruency effect. Irrespective of their differences, what all of the abovementioned accounts have in common is that they emphasize the competition between different features and therefore the need for inhibition to decrease the influence of irrelevant information, thus underlining the role of domain-general processes in explaining the congruency effect in numerosity decision tasks and numerosity processing in general (Wilkey & Ansari, 2020).

Comparative judgements of familiar objects

Most research on the number sense is based on abstract stimuli (i.e., dot configurations), resulting in a large gap with the numerosity perception of everyday familiar objects. Numerical decisions on familiar objects may differ in at least two ways from the typical numerosity comparison tasks with meaningless dots or shapes. First, familiarity of the object may enhance the individuation of that object because each individual object can be verbally labelled. Individuation refers to the process that an object is distinct from other objects and is an important step in enumeration and/or estimation of numerosities (Stoianov & Zorzi, 2012; Testolin et al., 2020). Evidence that familiarity can play a role in the individuation process comes from visual perception. For instance, Lupyan & Spivey, 2008) presented arrays of shapes and participants were instructed to indicate whether the display was homogeneous or contained an oddball. The shapes were 90° rotated numbers (“2” and “5”). In one condition, participants were informed about the nature of the shapes, whereas in another condition, they were told the stimuli were abstract shapes. Oddballs were detected faster in the condition in which participants were told that the shapes were rotated numbers (or when the participants spontaneously noticed that the shapes were known symbols) showing that familiarity influences visual search. In another study using electroencephalography , Gliga et al., (2010) observed enhanced oscillatory activity over the visual cortex in 1-year-old children when familiar objects were shown. Both results indicate that familiarity moderates early visual processing, possibly through top-down the influence of existing semantic knowledge. (Brady & Störmer, 2022; Jackson & Raymond, 2008; Le-Hoa Võ

& Wolfe, 2015; Xie & Zhang, 2017). Second, familiarity of the object is accompanied by the availability of conceptual knowledge of that object, including semantic knowledge about the size of the object in real life. This conceptual knowledge about the object's size in real life could be another dimension that is interfering with the decision. For instance, based on the seminal work of Pavio (1975), Gliksman et al. (2016) instructed participants to decide which object was larger on the screen (i.e., physical size) or, alternatively, larger in real life (i.e., conceptual size). Stimuli pairs were congruent (both conceptual and physical dimensions resulted in the same answer large objects, e.g., a large elephant vs a small mouse) or incongruent (e.g., a small elephant and large mouse). Congruency effects were observed regardless of the instructions, suggesting that both physical and conceptual sizes are processed quickly and effortlessly.

Against this background, some recent attempts were made to examine numerosity judgements with familiar objects and how they could possibly differ from numerosity decisions on dot patterns (Odic & Oppenheimer, 2023; Reynvoet et al., 2019). For instance, Odic & Oppenheimer (2023) contrasted performance in numerosity estimation task with either natural scenes or abstract stimuli (i.e., “How many chairs?”, “How many dots?”). However, contrary to their expectations, no advantage was observed for real-world scenes compared to abstract dot patterns. In another study, Reynvoet et al. (2019) compared participants' performance in a comparison task with dots and familiar objects. In the latter task, participants had to make a numerical comparison between pairs of physically large (apples, pears, oranges) and small fruits (grapes, strawberries, raspberries), presented in congruent (i.e., larger fruit is more numerous: 8 apples vs 4 strawberries), and incongruent conditions (i.e., larger fruit is less numerous: 4 apples vs 8 strawberries). The authors hypothesised that if familiarity leads to better object individuation, the congruency effect in the comparison task with familiar objects should be reduced. However, congruency effects were similar in both tasks, indicating that non-numerical dimensions also influence comparative judgements of familiar objects.

The current study

Both studies suggest that numerosity decisions on familiar objects are not different from numerosity judgements with dots: Numerosity comparison tasks with dots and familiar objects lead both to (similar) congruency effects. One caveat though is that the origin of the congruency effect as observed in Reynvoet et al. (2019) with familiar objects is yet unclear. In this study, the

size of small and large familiar objects as presented on the screen correlated with the natural size of the objects (the larger stimulus on the screen was also larger in reality), making it impossible to determine whether the congruency effect is caused by the physical size congruency (on the screen) or by the conceptual size congruency (retrieved from memory). This could possibly mean that although similar congruency effects were observed with dots and with familiar objects, their origin could still be different: physical size in the case of dots (since conceptual size is logically absent in dots) and conceptual size (or a combination of conceptual size and physical size) with familiar objects. Therefore, the current study aims to disambiguate the impact of physical and conceptual congruency effects on numerical comparison of familiar objects.

In this study, participants saw pictures with two sets of different animals and had to judge which animal was presented the most. In the *conceptual congruency condition*, one of the animals was a conceptually large animal (zebra) and the other a conceptually small animal (e.g., frog) and congruent (i.e., bigger animal (zebra) is more numerous) and incongruent conditions (i.e., smaller animal (frog) is more numerous) were created. The total area occupied by both type of animals was kept constant (i.e., when 5 zebras and 10 frogs were presented, the picture of the zebras was presented twice as large) and therefore, a congruency effect in this condition should be the result of conceptual size. In the *physical congruency condition*, the two different animals had a similar conceptual size (e.g., zebra-cow). Again, congruent (i.e., the more numerous animal occupies a larger total area on the screen), and incongruent trials (i.e., the more numerous animal occupies smaller total area) were compared. A congruency effect in this condition would as a consequence be caused by the physical size difference of the stimuli. The current design was implemented in two experiments. In Experiment 1, conceptual and physical congruency effects were examined in adults. In a second experiment, two samples of younger and older children were recruited to address developmental effects of both congruency effects.

Experiment 1 - Adults

Methods

Participants

A total of 50 adults participated in the study in exchange for course credits. One participant was removed due to low accuracy (see data preparation section), rendering the final sample to 49 participants aged between 18 and 40 (M_{age} = 20.31, SD = 4.07 years; 43 identified as females). We performed a power analysis to determine the sample size using the G*Power 3.1 software (Faul et al., 2007). To obtain the effect size, η_p^2 = 0.73 (Reynvoet et al., 2019), with α = .05 and power set at 90%, only 6 participants are needed. Consequently, power is guaranteed with the current sample size. Before the actual testing, written informed consent was sought from each participant. The ethical committee at the KU Leuven approved the study, its methods, and protocol (G-2021-4583).

Task, Stimuli, and Procedure

Stimuli preparation. The stimuli were four colored pictures of animals (zebra, cow, frog, and mouse). These pictures were taken from the picture database by Rossion and Pourtois (2004) and chosen based on a pilot study in which another sample of participants (*N* = 14) were presented with pictures of various animals (monkey, rabbit, zebra, spider, cat, butterfly, horse, fly, dog, fish, cow, frog, ant, mouse, and camel). Using a ten-point scale, participants had to rate the real-life size of animal. The animals were presented randomly, and their size on the screen of the animals was kept constant (i.e., 281×197 px). Based on these results, we selected two larger animals, rated similar in size (cow and zebra: 7.73 vs 7.69, $p > .05$) and two smaller animals rated similar in size (frog and mouse 2.81 vs 2.69, $p > .05$).

Stimuli manipulation. Based on the selection above, stimuli were created each containing two different types of animals. The two types of animals were presented intermixed on the screen (see Figure 1A). The participants had to decide which animal was more numerous. We created two main conditions in which either conceptual congruency or physical congruency was manipulated. In both conditions, the ratio between both sets of animals could be either 1.5 or 2. We also used numbers within (<4) and outside (>4) the subitizing range because it has been suggested that different perceptual mechanisms might underly the processing of small and large

sets which may impact congruency effects (Feigenson et al., 2004; but see Cheyette & Piantadosi, 2020). This resulted in the following numerosity pairs: 1-2, 2-3, 6-9 and 5-10. In the *conceptual congruency condition,* participants had to compare conceptually large animals and small animals (i.e., zebra – frog or cow – mice pair). Trials, where the conceptually bigger animal was more numerous, were considered congruent (e.g., 5 mice vs 10 cows), and trials, where the conceptually bigger animal was less numerous, were considered incongruent (e.g., 5 cows vs 10 mice; see Figure 1A).Total surface of the two sets of animals was always identical by reducing the more numerous animals with a factor corresponding to the numerical difference (ratio). For example, when 10 zebras and 5 frogs were presented (i.e., numerical ratio = 2), the width and height of each individual zebra were reduced with a factor sqrt(2). In this way, the individual surface of each zebra was two times smaller than the individual size of a frog. This resulted in the same total surface when all individual surfaces of the zebra and frog pictures were combined. For each combination (e.g., 5 mice vs 10 cows), eight different configurations were made manually with the animals positioned differently in the picture, resulting in 64 trials per pair.

In the *physical congruency condition*, the instructions were similar but the two types of animals had now the same conceptual size (e.g. zebra-cow), while the physical size was manipulated. In congruent trials the size of each individual animal was kept constant, resulting in a larger total area for the more numerous animal (e.g. when there were twice as many zebras than cows, the total area occupied by the zebras was also double). For incongruent trials, the size of the more numerous animal was reduced to obtain the inverse ratio in physical size as the numerical ratio. For example, when 10 zebras and 5 cows were presented (numerical ratio = 2), the width and height of each individual zebra were reduced by a factor of 2. In this way, the size of an individual zebra was 4 times smaller than that of a cow, and the total surface occupied by all the zebras was half of that of the cows. As in the conceptual congruency condition, for each pair of the design - congruency (congruent vs incongruent) × ratio (1.5 vs 2) × range (within vs outside of subitizing, eight different picture configurations were created, resulting in 64 trials per animal pair.

Procedure. Each trial started with a white fixation cross, presented in the center of a grey screen for 500 ms. Then the stimulus was presented for 1000 ms[1], followed by a black screen for maximum of 1000 ms. Participants were instructed to judge which animal was more numerous (e.g., "Are there more Frogs or Zebras?") by pressing the right or left arrow on a keyboard. Instructions emphasized both speed and accuracy. Responses were collected within a 2000 ms window – either during the stimulus's presentation or the blank screen's production. The next trial started after a 1500 ms intertrial interval, regardless of whether the participant responded within the 2000 ms time window or not. All participants performed four experimental blocks (i.e., one for each animal pair) of 64 trials each (i.e., 256 trials in total) with a short, self-paced break in between. Before each block, there were instructions and ten practice trials with feedback, followed by the 64 experimental trials without feedback. The stimuli were presented in a randomized blocked design. That is, participants started randomly either with the conceptual congruency or the physical congruency block, and the order of the subblocks (e.g., cow – mouse vs zebra – frog) was also random. PsychoPy Software (Peirce et al., 2019) was used to present stimuli and record the data. Participants self-administered the experiment online through pavlovia.org (Server Version: PSYCHOJS 2021.1.4) for which they received detailed written instructions. The study's materials, data set, and analyses, and supplementary information are available on the project's OSF page (https://osf.io/nxmjt/)

Results

Data preparation. The raw dataset consisted of 13056 data points (one point per experimental trial). We removed one participant who performed < 70% accuracy (N = 1), trials where no response was registered, duplicated data from a participant who erroneously did the experiment twice (N=1), which resulted in 12443 individual data points, which were used for the accuracy analysis. For the reaction time analyses, there were 10924 data points left after we filtered out the incorrect trials (N = 1519). We further excluded anticipation responses (RTs < 0.4s) and late response (RTs > 1.5s, i.e, (207 data points or 1.9%). This resulted in 10717 data points, which we used for the data analysis. Assumption checks showed that the RT data were normally distributed both for the conceptual congruency condition (*Skewness* = 0.92, *SE*= 0.03, *Kurtosis* =

[1] During the presentation, all images both experiments, were resized according to the participant's screens such that the height and the width of the image corresponded to the height of the participant's screen.

0.67, *SE* = 0.07, n = 5409), and the physical congruency condition (*Skewness* = 0.86, *SE*= 0.03, *Kurtosis* = 0.51, *SE* = 0.07, n = 5308). The data was then aggregated per subject and condition. Finally, across all conditions, there was a speed-accuracy trade-off, as evidenced by the presence of a positive correlation between the reaction times and accuracies, $r = 0.42$, $p = .003$. This correlation (also observed in Experiment 2) is likely due to the individual differences with some participants prioritizing fast over accurate responses and vice versa (for a review see Heitz, 2014). Consequently, we report separate analyses for each of these indices of performance.

Data analysis. Data are analyzed and reported in classical and Bayesian statistical frameworks whenever possible. We interpret the Bayes factors (BF), which is the ratio of the H_1 and H_0 likelihoods (BF_{10}). For analyses using more factors such as ANOVA (i.e., Bayesian Model Comparison), it is recommended that the model-averaged results be reported, such as the BFInclusion (BF_{Incl}). The latter reflects the predictive strength of the effect for the data by comparing all models that include the effect of interest to those without this effect (Van Den Bergh et al., 2020; van Doorn et al., 2020; Wagenmakers et al., 2018). By convention, the values of the BF are interpreted as the following: "anecdotal" (for values between < 1 and 3), "moderate" (for values between 3 and 10), "strong" (for values between 10 and 30), "very strong" (for values between 30 and 100), and "extreme" (for values > 100). To obtain both classical and Bayesian results, we used JASP statistical package v 0.18.3.0 (https://jasp-stats.org/) using default priors for the Bayesian Model Comparison (i.e., r scale fixed effects = 0.5, random effects = 1). See the supplementary materials for the breakdown of the interactions using Bayesian analyses. Confidence intervals (CI) for the classical effect sizes (e.g., η_p^2) are computed with an open-source calculator developed by James Uanhoro (https://effect-size-calculator.herokuapp.com/), whenever possible.

Table 1 depicts mean accuracies and the median reaction times per condition. Preliminary analyses indicated that including the animal pair as a factor leads to unexpected four-way interactions driven by the cow–mouse and zebra–frog pairs. Consequently, for each condition (conceptual size and physical size) and each animal pair (cow–mouse, zebra–frog, cow–zebra, and mouse–frog), we report the (Bayesian) repeated-measures ANOVA with congruency (congruent vs incongruent), ratio (1.5 vs 2.00), and range (subitizing vs non-subitizing) as within-subject factors separately for the accuracy and the reaction time data (see Figure 1B).

[INSERT TABLE 1 HERE]

Conceptual Congruency Condition

Cow – Mouse Pairs

Accuracy. Classical ANOVA showed only a significant ratio effect, $F(1,48) = 9.39$, $p = .004$, $\eta_p^2 = 0.16$, 90% CI [0.03, 0.31] with higher accuracies for ratio 2 compared to 1.5 (0.92 vs 0.89). The remaining main effects and interactions were not significant Fs < 3.89, ps > .05, $\eta_p^2 < 0.05$, 90% CIs [0, 0.21] (see Figure 1B).

The results from the Bayesian Repeated Measures ANOVA (i.e., Bayesian Model Comparison) were in line with the classical approach and showed that the model that best explained the data was the model including only the ratio effect, $BF_{10} = 4.11$. The model-averaged results further corroborated this and showed weak anecdotal evidence for including the effect of ratio, $BF_{Incl} = 1.83$.

Reaction Times. Classical ANOVA showed no congruency effect, $F(1,48) = 0.30$, $p = .59$, $\eta_p^2 < 0.01$, 90% CI [0, 0.08] (see Figure 1B). There was a significant ratio effect, $F(1,48) = 31.64$, $p < .001$, $\eta_p^2 = 0.40$, 90% CI [0.22, 0.53] showing faster RTs for ratio 2 compared to 1.50 (0.76 vs 0.80). There was a significant main effect of range, $F(1,48) = 13.20$, $p < .001$, $\eta_p^2 = 0.22$, 90% CI [0.07, 0.37], suggesting, somewhat unexpectedly, faster RTs for numbers in the non-subitizing range (0.77 vs 0.80). There were no significant interactions, Fs < 2.6, ps > .05, $\eta_p^2 < 0.05$, 90% CIs [0, 0.17] .

Bayesian Model Comparison showed that the model that best explained the data contained the main effects of Ratio and Range, $BF_{10} = 40553.25$. This was further corroborated by comparing the model-averaged results showing extreme evidence for the inclusion of ratio, $BF_{Incl} = 1183.55$, and moderate support for the inclusion of range, $BF_{Incl} = 10.80$.

Zebra – Frog Pairs

Accuracy. Classical ANOVA showed a main effect of congruency, $F(1,48) = 54.27$, $p < .001$, $\eta_p^2 = 0.53$, 90% CI [0.36, 0.64], yielding higher accuracies for congruent than incongruent trials (0.91 vs 0.83; see Figure 1B). There was also a main effect of ratio, $F(1,48) = 16.71$, $p < .001$, $\eta_p^2 = 0.26$, 90% CI [0.09, 0.41], and range, $F(1,48) = 25.43$, $p < .001$, $\eta_p^2 = 0.35$, 90% CI [0.17,

0.49]. The congruency × ratio interaction was significant, $F(1,48) = 10.91$, $p = .002$, $\eta_p^2 = 0.19$, 90% CI [0.05, 0.34]. Post-hoc comparison showed that the effect of congruency was present only for trials of ratio 1.5, $p_{bonf} < .001$, $d = 0.83$, 95% CI [0.45, 1.20], but not for trials of ratio 2, $p_{bonf} = 0.11$. The interaction between congruency and range was present but weak, $F(1,48) = 3.97$ $p = .05$, $\eta_p^2 = 0.08$, 90% CI [0.00, 0.21]. Post-hoc comparisons showed that the effect of congruency was present outside the subitizing range, $p_{bonf} < .001$, $d = 0.80$, 95% CI [0.35, 1.24], but not in the subitizing range, $p_{bon} = .24$. There was also a ratio × range, $F(1,48) = 5.66$, $p = .02$, $\eta_p^2 = 0.11$, 90% CI [0.009, 0.25], showing a significant ratio effect only for pairs outside the subitizing range, $p_{bonf} < .001$, $d = -0.56$, 95% CI [-0.92, -0.20] . The three-way congruency × ratio × range interaction was not significant, $F<1$, $p = 0.71$.

Accordingly, the Bayesian Repeated Measures ANOVA showed that the model that best explained the data was the model including all the main effects and all the first-order interactions, $BF_{10} = 1.881\times10^{12}$ (Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range). The model-averaged results also showed extreme support for including all main effects (all $BF_{Incl} > 100$), very strong support for the inclusion of the congruency × ratio interaction, $BF_{Incl} = 35.77$, and anecdotal to moderate support for the inclusion of congruency × range, $BF_{Incl} = 2.83$, and ratio × range interactions, $BF_{Incl} = 3.99$.

Reaction Times. Classical ANOVA showed a significant congruency effect (0.76 vs 0.84), $F(1,48) = 57.26$, $p < .001$, $\eta_p^2 = 0.54$, 90% CI [0.38, 0.65], a ratio effect, $F(1,48) = 29.40$, $p < .001$, $\eta_p^2 = 0.38$, 90% CI [0.20, 0.51],(0.82 vs 0.77), and an effect of number range $F(1,48) = 19.62$, $p < .001$, $\eta_p^2 = 0.29$, 90% CI [0.12, 0.44] with faster RTs for pairs within the subitizing range (0.77 vs 0.82; see Figure 1B). There was also a significant interaction between congruency and range, $F(1,48) = 13.96$, $p < .001$, $\eta_p^2 = 0.23$, 90% CI [0.07, 0.38]. Post-hoc test (Bonferroni) showed that the congruency effect was present in both number ranges, but larger outside the subitizing range. That is the congruent trials were significantly faster than the incongruent ones both within the subitizing, $p_{bonf} < .001$, $d = -0.37$, 95% CI [-0.68,-0.07], and at the outside of the subitizing range $p_{bonf} < .001$, $d = -0.88$, 95% CI [-1.25,-0.50]. The remaining interactions were not significant, Fs < 1.22, ps $> .05$, $\eta_p^2 < 0.02$, 90% CIs [0, 0.13].

Bayesian Repeated Measures ANOVA showed that the model that best explained the data was the model including all the main effects and the interaction between congruency and range, BF_{10} 7.130×10^{14} (Congruency + Ratio + Range + Congruency*Range). The model-averaged results also showed extreme support for including all main effects (all $BF_{Incl} > 100$) and very strong evidence for the interaction, $BF_{Incl} = 91.03$.

Physical Congruency Condition

Cow – Zebra Pairs

Accuracy. Classical ANOVA showed that all main effects were significant (see Figure 1B). There was a physical congruency (i.e., total surface) effect, $F(1,48) = 7.76$, $p = .008$, $\eta_p^2 = 0.14$, 90% CI [0.02, 0.29] (0.89 vs 0.86), ratio effect, $F(1,48) = 10.72$, $p = .002$, $\eta_p^2 = 0.18$, 90% CI [0.04, 0.33] (0.89 vs 0.86), and an effect of range, $F(1,48) = 19.07$, $p < .001$, $\eta_p^2 = 0.28$, 90% CI [0.12, 0.43] (0.90 vs 0.84). The congruency × ratio interaction was also significant, $F(1,48) = 13.89$, $p < .001$, $\eta_p^2 = 0.22$, 90% CI [0.07, 0.37]. Post-hoc comparison (Bonferroni corrected) showed that the effect of congruency was significant only for pairs with a ratio of 2.00, $p_{bonf} < .001$, $d = 0.54$, 95% CI [0.19, 0.88], but not for pairs with ratio 1.50, $p_{bonf} = 1.00$, $d = -0.06$, 95% [-0.37, 0.25]. There was also a ratio × range interaction, $F(1,48) = 12.85$, $p < .001$, $\eta_p^2 = 0.21$, 90% CI [0.06, 0.36]. The post-hoc comparison showed that the ratio effect was significant only for numbers in the non-subitizing range, $p_{bonf} < .001$, $d = -0.61$, 95% CI [-0.98, -0.24], but not for pairs within the subitizing range, $p_{bonf} = 1.00$, $d = 0.05$, 95% CI [-0.29, 0.38] . The remaining interactions were not significant Fs < 1, ps > .05, $\eta_p^2 < 0.01$, 90% CIs [0, 0.01].

Bayesian Model Comparison corroborated the classical results by showing that the model best explaining the data was the model including all the main effects and the interactions between congruency and ratio, and ratio and range (Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range), $BF_{10} = 5.294\times10^{6}$. The analysis of effects also supported this: main effect of congruency, $BF_{Incl} = 40.18$, ratio, $BF_{Incl} > 100$, range, $BF_{Incl} > 100$, congruency × ratio, $BF_{Incl} = 53.72$, ratio × range, $BF_{Incl} = 74.34$

Reaction Times. Classical ANOVA showed that all main effects were significant and in the expected directions (see Figure 1B): congruency, $F(1,48) = 22.80$, $p < .001$, $\eta_p^2 = 0.32$, 90% CI [0.15, 0.46], ratio, $F(1,48) = 31.37$, $p < .001$, $\eta_p^2 = 0.40$, 90% CI [0.21, 0.53], and range, $F(1,48) =$

4.81, p = .03, η_p^2 = 0.09, 90% CI [0.004, 0.23]. There were no significant interactions between the factors, Fs < 3.8, ps > .05, η_p^2 < 0.07, 90% CIs [0, 0.21].

The results from the Bayesian Model Comparison were largely in line with the classical ANOVA. The model that best explained the data was the model including all the main effects and the interaction between congruency and ratio, BF_{10} = 642012.03 (Congruency + Ratio + Range + Congruency*Ratio). The analysis of effects showed extreme evidence for the inclusion of the main effects of congruency and ratio, allBF_{Incl} > 100, but weak anecdotal evidence for the inclusion of main effect of range, and the congruency × ratio, interaction, allBF_{Incl} < 2.

Mouse – Frog Pairs

Accuracy. Classical ANOVA showed that all main effects were significant and in the expected directions (see Figure 1B): congruency, F(1,48) = 28.56, p <.001, η_p^2 = 0.37, 90% CI [0.19, 0.51], ratio, F(1,48) = 23.05, p < .001, η_p^2 = 0.32, 90% CI [0.15, 0.47], and range, F(1,48) = 8.68, p = .005, η_p^2 = 0.15, 90% CI [0.03, 0.30]. All the interactions were significant (all ps < .001), including the three-way congruency × ratio × range interaction, F(1,48) = 6.58, p = .013, η_p^2 = 0.12, 90% CI [0.014, 0.27]. Post hoc comparisons showed that the effect of congruency was significant only for small ratio pairs (i.e., 1.50) in the non-subitizing range (i.e., trials involving the quantity of 6 vs 9 animals), p_{bonf} < .001, d = 1.62 , 95% CI [0.86, 2.37]. Bayesian ANOVA showed the same outcomes with the model explaining the data best, being the model including all the main effects and interactions, BF_{10} = 2.917×10^{17} (Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range + Congruency*Ratio*Range). All showed extreme support for the inclusion of the effects and their interactions; all BF_{Incl} > 100.

Reaction Times. Classical ANOVA showed significant main effects of congruency, F(1,48) = 37.79, p < .001, η_p^2 =0.44, 90% CI [0.26, 0.57], and ratio, F(1,48) = 54.65, p < .001, η_p^2 = 0.53, 90% CI [0.36, 0.64] (see Figure 1B) . The congruency × range interaction was also significant, F(1,48) = 10.58, p = .002, η_p^2 = 0.18, 90% CI [0.04, 0.33]. Post-hoc test (Bonferroni correction applied) showed that the congruency effect was present only for number pairs outside of the subitizing range, p_{bonf} <.001, d= -0.65, 95%CI [-0.96, -0.33], but not for numbers within the subitizing range, p_{bonf} =.06. The ratio × range interaction was also significant, F(1,48) = 14.90, p < .001, η_p^2 = 0.24, 90% CI [0.08, 0.39]. Post-hoc test (Bonferroni correction applied) showed that

the ratio effect was present for numbers within the subitizing range, $p_{bonf} < .001$, $d = 0.70$, 95%CI [0.39, 1.00], but not for outside the subitizing range, $p_{bonf} = 0.14$. The remaining main effects and interactions were not significant, Fs < 4, ps > .05, $\eta_p^2 < 0.08$, 90% CIs [0, 0.].

Bayesian Model Comparison showed that the best model was the one that included all the main effects and the first-order interactions, $BF_{10} = 4.266 \times 10^{13}$ (Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range). The analysis of effects showed extreme support for the inclusion of all main effects, the ratio × range interaction, all BFIncl > 100, strong support for the inclusion of the congruency × range interaction, BFIncl = 34.53, and anecdotal evidence for the inclusion of congruency × ratio interaction, $BF_{Incl} = 2.88$.

Interim Discussion

Overall, the results of Experiment 1 showed that while the physical congruency effect (i.e., the effect of the total surface) was present in both animal pairs (cow – zebra and mouse – frog), the conceptual congruency effect (i.e., the effect of the semantic knowledge about the animal's real size) was present only in the zebra – frog pairs, but not in the cow – mouse (see Figure 1B). Possibly, this is due to some specific characteristics of the mouse picture. The mouse picture contains a large white space between the mouse's grey body and its tail (see Figure 1A). Because the entire picture was rescaled (occupied grey and non-occupied white parts) to create the different congruency conditions, our attempt to control the total surface might have failed in this condition (the total area of the mice when focusing on the grey parts are smaller than the total area of the cows). Therefore, in Experiment 2 with children, we only administered the two conditions without the mice (cow-zebra; zebra-frog).

We should also note that, due to the interrelatedness of different non-numerical dimensions, not all non-numerical dimensions could be controlled in the conceptual congruency condition. We consistently controlled total surface because space related non-numerical dimensions (e.g., density, convex hull) influence performance in an intermixed numerosity comparison task much less than size related characteristics since it is difficult to mentally compute density and/or convex hull in intermixed displays. By controlling for the total area of the animal type, however, we did not control individual size (i.e., the area occupied by one animal). By keeping total area constant, the individual size of the more numerous animal is indeed always smaller (see Figure

1A). It is important to note that this physical difference cannot account for the conceptual congruency effect because the more numerous animal is depicted smaller in both congruent and incongruent conditions. Finally, the congruency the physical congruency effect was pronounced for trials of ratio 2 (i.e., cow - zebra) or trials outside the subitizing range, or a combination of both (e.g., ratio 1.5 and in the non-subitizing range in the mouse frog trials). Somewhat similar, the conceptual congruency effect was also present and more pronounced for the difficult trials or ratio 1.5 and for trials outside of the subitizing range, but was absent or weaker for easier trails (i.e., trials of ratio 2 and trails within the subitizing range). These observations are in line with previous studies, suggesting an interplay between the numerical, non-numerical the more the uncertainty about a numerical decision increases, the more non-numerical information with the numerical decision (Clayton et al., 2015; Defever et al., 2013; Reynvoet et al., 2019, 2021). Furthermore, these results show that top-down conceptual information about the item can also interfere with numerical decision tasks. In sum, Experiment 1 provided evidence for distinct physical and conceptual congruency effects in adults. In Experiment 2, we take a developmental approach and evaluate how both congruency effects evolve through development.

[INSERT FIGURE 1AB HERE]

Experiment 2 - Children

Methods

Participants

A total of 148 children from Flemish primary schools participated in the study. Data from 11 participants was removed (see data preparation section), thus rendering the final sample to 137 participants, divided into two age groups. The first age group consisted of 53 children aged between 7 and 9 years (M_{age} =7.42, SD = 0.54 years, 24 boys and 29 girls). The second age group comprised 90 children aged between 11 and 12 years (M_{age} =11.26, SD = 0.44 years, 30 boys and 54 girls). We used the same parameters as Experiment 1 to determine the sample size. The sex ratio did not differ between the two age groups, X^2 (1, N = 137) = 1.25, p = 0.26. Before the actual testing, written informed consent was sought from children's legal representatives. The study, its methods, and protocol were approved by the ethical committee at the KU Leuven (G-2021-4583)

Task, Stimuli, and Procedure

The task and stimuli were identical to Experiment 1, with the following exceptions. Firstly, the stimulus was presented *until the response*. Secondly, children completed only two blocks – one conceptual congruency block (zebra-frog) and one physical congruency block (zebra–cow). Before each block, there were instructions and ten training trials with feedback. This was followed by 64 experimental trials without feedback for each condition, resulting in a total of 128 trials.

Children were tested in the second half of the school year between March and May. Because we anticipated potential COVID-19 hygiene restrictions at schools, we used pavlovia.org (Server Version: PSYCHOJS 2021.1.4) for the stimuli presentation and recording of the data. However, the majority of the testing was administered in school in groups of 4 -5 children in the presence of a research assistant. Only in one school, the researchers were not allowed to be physically present and parents were sent an email invitation for their child to self-administer the experiment. However, the response rate on this invitation was very low and only 3 participants self-administered the experiment for which their legal representatives received detailed written instructions with the instructions for their child. However, the researchers were

unable to control whether these children received these instructions and whether their parents helped them complete the experiment.

Results

Data preparation. The raw dataset consisted of 18944 data points (one point per experimental trial). First, we excluded all trials where the reaction times (RT) were < 0.4 s (anticipations or > 3.5 s (late responses), irrespective of the accuracy (n = 1185, 6.26%). This was done because the stimulus could stay on the screen until response, and thus, the longer response times could indicate counting. We also removed data from one participant due to missing age information, data from another 9 participants who performed > 2 SD below the mean accuracy for their age group, and data from 1 participant who had < 70% of the total trials left. This rendered the data set to 16512 individual data points on which the accuracy analyses were based. The data set for the RT analyses consisted of 14648 data points after filtering out the incorrect responses (n = 1864, 11.30%). Assumption checks showed that the RT data were normally distributed, for the conceptual congruency condition (*Skewness* = 0.39, *SE*= 0.05, *Kurtosis* = -0.09, *SE* = 0.10, n = 2565), and the physical congruency condition (*Skewness* = 0.41, *SE*= 0.05, *Kurtosis* = 0.01, *SE* = 0.10, n = 2597) for the Age group 1 (7-9 years old). The results were the same for the age group 2 (11-12 years old): conceptual congruency condition (*Skewness* = 0.81, *SE*= 0.04, *Kurtosis* = 0.65, *SE* = 0.07, n = 4726), and the physical congruency condition (*Skewness* = 0.75, *SE*= 0.04, *Kurtosis* = 0.50, *SD* = 0.07, n = 4760). We, therefore, did not remove any further data points. The data were then aggregated per subject and condition. Similarly to the data from Experiment 1, the reaction times and accuracies were positively correlated in both age groups, $r = .36$, $p = .008$, and $r = .38$, $p < .001$. We, thus, analysed them separately.

Data analysis. Table 2 depicts mean accuracies and the median reaction times per condition and age group. A two-tailed independent (Bayesian) student sample t-test, showed that overall, younger children (7 – 9 years olds) responded slower (1.88 vs 1.46), $t(135) = 7.71$, $p < .001$, $d = 1.35$, 95% CI [0.97, 1.73], $BF_{10} > 100$, and less accurately, $t(135) = -4.84$, $p < .001$, $d = -0.85$, 95% CI [-1.21, -0.49], $BF_{10} > 100$, compared to their older peers (11 – 12 years olds). However, the two groups differed in their sample size, which led to a violation of equality of variances, specifically for the accuracy data, $F(1, 135) = 20.31$, $p < .001$. Consequently, we performed

(Bayesian) repeated-measures ANOVA for each age group separately (7 – 9 years olds vs 11 – 12 years olds) and per condition (conceptual vs physical), with congruency (congruent vs incongruent), ratio (1.5 vs 2.00), and range (subitizing vs non-subitizing) as within-subject factors and age group (7-9 years olds vs 11-12 years olds) as a between-subject factor.

[INSERT TABLE 2 HERE]

Conceptual Congruency (Zebra – Frog Pairs) in the 7 – 9-year-old group.

Accuracies. All main effects were significant and in the expected directions: the effect of congruency, $F(1, 52) = 9.19$, $p = .004$, $\eta_p^2 = 0.15$, 90% CI [0.03,0.29], yielded higher accuracies for congruent trials, compared to incongruent (0.87 vs 0.82); the main effect of ratio, $F(1, 52) = 4.46$, $p = .04$, $\eta_p^2 = 0.08$, 90% CI [0.002, 0.21], showed higher accuracies for ratio 2 compared to 1.5 (0.86 vs 0,83); and the main effect of range, $F(1, 52) = 18.77$, $p < .001$, $\eta_p^2 = 0.27$, 90% CI [0.11, 0.41], yielded higher accuracies for trials in the subitizing range than those in the non-subitizing range (0.88 vs 0.81).

The congruency × ratio, $F(1,52) = 8.28$, $p = .006$, $\eta_p^2 = 0.14$, 90% CI [0.02, 0.28], was significant. Post-hoc tests (Bonferroni corrected), showed that the effect of congruency was present only for trials of ratio 1.5 (i.e., trials 2 vs 3 and 6 vs 9), $p_{bonf} < .001$, $d = 0.47$, 95% CI [0.14, 0.79], but not for ratio 2 (i.e., trials 1 vs 2 and 5 vs 10), $p_{bonf} = 1.00$. There was also congruency × range interaction, $F(1, 52) = 5.40$, $p = .02$, $\eta_p^2 = 0.09$, 90% CI [0.007, 0.23], where the effect of congruency was present only for trials outside of the subitizing range (6 vs 9 and 5 vs 10), $p_{bonf} = .002$, $d = 0.48$, 95% CI [0.11, 0.84], but not for trials within the subitizing range (1 vs 2 and 2 vs 3), $p_{bonf} = 1.00$. Finally, ratio × range, $F(1,52) = 6.91$, $p = .01$, $\eta_p^2 = 0.12$, 90% CI [0.016, 0.26] was also present, showing that the effect of ratio was present only for trials outside of the subitizing range, $p_{bonf} = .007$, $d = -0.35$, 95% CI [-0.64, -0.05]. The three-way interaction congruency × ratio × range was not significant, $F(1,52) = 1.47$, $p = .23$, $\eta_p^2 = 0.03$, 90% CI [0.00, 0.13] .

Bayesian Model Comparison showed that the model that best explained the data contained all the main effects and first-order interactions (Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range), BF10 = 37953.51. This was further corroborated by the model-averaged results: congruency, $BF_{Incl} = 38.28$, ratio, $BF_{Incl} = 7.82$, range, $BF_{Incl} > 100$, congruency × ratio, $BF_{Incl} = 11.24$, congruency × range, $BF_{Incl} = 7.05$, ratio × range, $BF_{Incl} = 6.61$.

Reaction times. Repeated measures ANOVA on the median reaction times showed that all main effects were significant and in the expected directions: the main effect of congruency, $F(1, 51) = 8.78$, $p = .005$, $\eta_p^2 = 0.15$, 90% CI [0.03, 0.29], yielded faster RTs for congruent compared to incongruent trials (1.87 vs 1.94); the main effect of ratio, $F(1, 51) = 48.19$, $p < .001$, $\eta_p^2 = 0.49$, 90% CI [0.31, 0.60], suggested faster RTs for ratio 2 compared to ratio 1.5 (1.84 vs 1.97); and finally, the main effect of range, $F(1, 51) = 9.08$, $p = .004$, $\eta_p^2 = 0.15$, 90% CI [0.03, 0.30] showed faster RTs for numbers in the subitizing range, compared to trials in the non-subitizing range (1.87 vs 1.94). The only significant interaction was congruency × range, $F(1, 51) = 6.46$, $p = .01$, $\eta_p^2 = 0.11$, 90% CI [0.013, 0.25], with the congruency effect being present only for trials in the non-subitizing range, $p_{bonf} = .001$, $d = -0.32$, 95% CI [-0.56, -0.08], but not for trials in the subitizing range, $p_{bonf} = 1.00$. The remaining interactions were not significant, ps > .05.

Bayesian Model Comparison supported the classical results showed that the model that best explained the data contained all main effects and only the interaction between congruency and range (Congruency + Ratio + Range + Congruency*Range), $BF10 = 1.10 \times 10^6$. This was further corroborated by comparing the model-averaged results showing extreme evidence for the inclusion of ratio, $BF_{Incl} > 100$, and moderate support for the inclusion of the remaining effects and interactions: congruency, $BF_{Incl} = 4.79$, range, $BF_{Incl} = 7.53$, and congruency × range, $BF_{Incl} = 4.29$.

Conceptual Congruency (Zebra–Frog Pairs) in the 11 – 12-year-old group.

Accuracies. The results from the repeated measures ANOVA show a significant main effect of congruency, $F(1, 83) = 7.00$, $p = .010$., $\eta_p^2 = 0.08$, 90% CI [0.01, 0.18] with higher accuracies for congruent than incongruent trials (0.92 vs 0.90).There was also a main effect of ratio, $F(1, 83) = 25.26$, $p < .001$, $\eta_p^2 = 0.23$, 90% CI [0.11, 0.35] yielding higher accuracies for ratio 2 compared to ratio 1.5 (0.93 vs 0.89). The main effect of range, $F(1, 83) = 47.02$, $p < .001$, $\eta_p^2 = 0.36$, 90% CI [0.23, 0.47] indicated higher accuracies for trials in the subitizing range than those in the non-subitizing range (0.94 vs 0.88). There was a congruency × ratio interaction, $F(1, 83) = 10.74$, $p = .002$, $\eta_p^2 = 0.12$, 90% CI [0.03, 0.22]. Post hoc comparison showed that the effect of conceptual congruency was present only for trials with a ratio of 1.5 (i.e., 2 vs 3 and 6 vs 9), $p_{bonf} < .001$, $d = 0.42$, 95% CI [0.14, 070], but not for ratio 2, $p_{bonf} = 1.00$. There was also a congruency × range interaction , $F(1, 83) = 9.77$, $p = .002$, $\eta_p^2 = 0.11$, 90% CI [0.02, 0.21] with the congruency effect being present only for trials in the non-subitizing range, $p_{bonf} < .001$, $d = 0.43$, 95% CI [0.14,

0.72], but not for trials in the subitizing range, p_{bonf} = 1.00. Finally, the ratio × range interaction was also significant, $F(1, 83) = 33.31$, $p < .001$, $\eta_p^2 = 0.29$, 90% CI [0.16, 0.40], yielding a significant ratio effect only in the non-subitizing range, $p_{bonf} < .001$, d = -0.71, 95% CI [-1.00, -0.42]. The three-way interaction congruency × ratio × range was not significant, $F(1, 83) = 3.44$, $p = .07$, $\eta_p^2 = 0.04$, 90% CI [0.00, 0.13].

Bayesian Model Comparison showed that the model that best explained the data, $BF_{10} = 1.721\times10^{17}$, was the full model including the main effects or congruency, $BF_{Incl} > 100$, ratio, $BF_{Incl} > 100$, range, $BF_{Incl} >100$, the congruency × ratio, $BF_{Incl} = 62.2$, and congruency × range, $BF_{Incl} = 56.33$, and congruency × ratio × range, $BF_{Incl} = 23.87$ interactions. Though this seems slightly contradictory with the classical results, the second-best model $BF_{10} = 1.095\times10^{17}$ was identical to the classical results, suggesting some evidence for the three-way interaction, but the interaction was probably unstable.

Reaction times. Results showed a significant main effect of congruency, $F(1, 83) = 36.93$, $p < .001$, $\eta_p^2 = 0.31$, 90% CI [0.18, 0.42] with faster RTs for congruent compared to incongruent trials (1.42 vs 1.52). There was a main effect of ratio $F(1, 83) = 97.41$, $p < .001$, $\eta_p^2 = 0.54$, 90% CI [0.42, 0.63], yielding faster RTs for ratio 2 compared to ratio 1.5 (1.41 vs 1.53), and main effect of range $F(1, 83) =58.38$, $p < .001$, $\eta_p^2 = 0.41$, 90% CI [0.28, 0.52] showing faster RTs for trials within the subitizing range than for trials in the non-subitizing range (1.39 vs 1.54). The interaction between congruency and ratio was significant, $F(1, 83) = 10.24$, $p = .002$, $\eta_p^2 = 0.11$, 90% CI [0.03, 0.22], with the congruency effect being present for both ratio 1.5, $p_{bonf} < .001$, d = -0.36, 95% CI [-0.52, -0.20] and ratio 2, $p_{bonf} = .04$, d = -0.15, 95% CI [-0.29, -0.04] but larger at ratio 1.5 . There was also congruency × range interaction, $F(1, 83) =28.49$, $p < .001$, $\eta_p^2 = 0.26$, 90% CI [0.13, 0.37], with the congruency effect being present only for trials in the non-subitizing range, $p_{bonf} < .001$, d = -0.48, 95% CI [-0.67, -0.29], but not for trial in the subitizing range, $p_{bonf} = 1.00$. The remaining interactions were not significant, all $p > .05$.

Bayesian Model Comparison showed that the model that best explained the data, $BF_{10} = 7.123\times10^{28}$, was the full model including the main effects or congruency, $BF_{Incl} > 100$, ratio, $BF_{Incl} > 100$, range, $BF_{Incl} >100$, the congruency × ratio, $BF_{Incl} = 12.39$, and .congruency × range, $BF_{Incl} > 100$ interactions.

Physical Congruency Condition (Cow - Zebra Pairs) in the 7 – 9-year-old group

Accuracies. Results of the repeated measures ANOVA showed that there was a reversed main effect of congruency, $F(1, 52) = 5.00$, $p = .03$, $\eta_p^2 = 0.09$, 90% CI [0.005, 0.22] with higher accuracies for incongruent than congruent trials (0.86 vs 0.83). There was a main effect of ratio, $F(1, 52) = 48.88$, $p < .001$, $\eta_p^2 = 0.49$, 90% CI [0.31, 0.60] with higher accuracies for ratio 2 than ratio 1.5 (0.88 vs 0.81). There was also a main effect of range, $F(1, 52) = 17.14$, $p < .001$, $\eta_p^2 = 0.25$, 90% CI [0.09,0.39] with higher accuracies for trials in the subitizing range than for trials in the non-subitizing range (0.88 vs 0.81). The interaction between congruency and ratio was significant, $F(1, 52) = 5.20$, $p = .03$, $\eta_p^2 = 0.09$, 90% CI [0.006, 0.22], with the reversed congruency effect being present only for trials of ratio 1.5, $p_{bonf} = .01$, $d = -0.34$, 95% CI [-0.64, -0.04], but not of ratio 2, $p_{bonf} = 1.00$. There was also a ratio × range interaction, $F(1, 52) = 5.52$, $p = .02$, $\eta_p^2 = 0.10$, 90% CI [0.007, 0.23], with the ratio effect being present in both ranges but larger in the non-subitizing range: subitizing, $p_{bonf} = .002$, $d = -0.30$, 95% CI [-0.54, -0.07], and non-subitizing, $p_{bonf} < .001$, $d = -0.56$, 95% CI [-0.82, -0.29]. The remaining interactions were not significant, all ps > .05.

Bayesian Model Comparison showed that the model that best explained the data, $BF_{10} = 6.173 \times 10^7$, included the main effects or congruency, $BF_{Incl} = 2.13$, ratio, $BF_{Incl} > 100$, range, $BF_{Incl} > 100$, and the congruency × ratio, $BF_{Incl} = 2.96$, and ratio × range, $BF_{Incl} = 2.86$, interactions.

Reaction times. There was no main effect of congruency, $F(1, 52) = 0.06$, $p = .80$, $\eta_p^2 = 0.001$, 90% CI [0.00, 0.05] There was a significant main effect of ratio, $F(1, 52) = 14.21$, $p < .001$, $\eta_p^2 = 0.22$, 90% CI [0.07, 0.36] yielding faster RTs for ratio 2 than ratio 1.5 (1.81 vs 1.90). There was also a significant ratio × range interaction, $F(1, 52) = 6.88$, $p = .01$, $\eta_p^2 = 0.12$, 90% CI [0.02, 0.26], which was further embedded in a significant congruency × ratio × range interaction, $F(1, 52) = 6.96$, $p = .01$, $\eta_p^2 = 0.12$, 90% CI [0.02, 0.26]. However, post hoc comparisons showed no significant congruency effects in none of the conditions, all $p_{bonf} > .05$.

Bayesian Model Comparison showed that the model that best explained the data, $BF_{10} = 37.51$, included the main effect of ratio, range, and ratio × range interaction. Though, the second-best model fit was the model including only the main effect of ratio, $BF_{10} = 34.98$. Indeed, the model-

averaged results did not provide clear support for the first model: ratio, $BF_{Incl} = 24.99$, range, $BF_{Incl} = 0.53$, and ratio × range, $BF_{Incl} = 1.96$.

Physical Congruency Condition (Cow - Zebra Pairs) in the 11 – 12-year-old group

Accuracies. Results of the repeated measures ANOVA showed no main effect of congruency, $F(1, 83) = 0.64$, $p = .43$, $\eta_p^2 = 0.008$, 90% CI [0.00, 0.07]. However, there was ratio effect, $F(1, 83) = 10.49$, $p = .002$, $\eta_p^2 = 0.11$, 90% CI [0.03, 0.22] with higher accuracies for ratio 2 than ratio 1.5 (0.93 vs 0.90). There was also a main effect of range, $F(1, 83) = 20.13$, $p < .001$, $\eta_p^2 = 0.20$, 90% CI [0.08, 0.31] with higher accuracies for trials in the subitizing range than for trials in the non-subitizing range (0.93 vs 0.90). The was a weak interaction between congruency and ratio, $F(1, 83) = 4.99$, $p = .03$, $\eta_p^2 = 0.06$, 90% CI [0.003, 0.15], but the effect of congruency was not significant neither for ratio 1.5, not for ratio 2, all $p_{bonf} = 1.00$. There was also a ratio × range interaction, $F(1, 83) = 14.69$, $p < .001$, $\eta_p^2 = 0.15$, 90% CI [0.05, 0.27], with the ratio effect being present only for the non-subitizing range, $p_{bonf} < .001$, $d = -0.53$, 95% CI [-0.83, -0.22], and but not for trials in the subitizing range, $p_{bonf} = 1.00$. The remaining interactions were not significant, all ps > .05.

Bayesian model comparison showed that the model that best explained the data, $BF_{10} = 410960.22$, was the model including the main effects of ratio, $BF_{Incl} > 100$, range, $BF_{Incl} > 100$, and the ratio × range interaction, $BF_{Incl} > 100$.

Reaction times. There was no main effect of congruency, $F(1, 83) = 0.31$, $p = .58$, $\eta_p^2 = 0.004$, 90% CI [0.00, 0.05]. There were, however, main effects of ratio, $F(1, 83) = 92.42$, $p < .001$, $\eta_p^2 = 0.53$, 90% CI [0.41, 0.61], and range, $F(1, 83) = 50.14$, $p < .001$, $\eta_p^2 = 0.38$, 90% CI [0.24, 0.48]: trials of ratio 2 were responded faster compared to trials of ratio 1.5 (1.38 vs 1.51), and trials in the subitizing range yielded faster RTs compared to trials in the non-subitizing range (1.39 vs 1.50). The only significant interaction was between congruency and range, $F(1, 83) = 4.68$, $p = .03$, $\eta_p^2 = 0.05$, 90% CI [0.002, 0.15], but there was no significant congruency effect in either of the ranges: subitizing, $p_{bonf} = .34$, and non-subitizing, $p_{bonf} = 1.00$.

Bayesian model comparison showed that the model that best explained the data, $BF_{10} = 3.600 \times 10^{18}$ was the model including the main effects of ratio, $BF_{Incl} > 100$, range, $BF_{Incl} > 100$, and the ratio × range interaction, $BF_{Incl} > 100$.

Interim Discussion

The results from Experiment 2 showed that, similarly to adults, a conceptual congruency effect is present in younger and older children. Decisions were more accurate and faster when the more numerous animal was larger in real life (e.g., 10 zebras and 5 frogs) than when the more numerous animal was smaller in real life (e.g., 10 frogs and 5 zebras), suggesting that stored semantic information about real-world size interferes with numerosity perception already at the beginning primary school (i.e., 7 years onwards).

Contrary to our expectations and the observations in adults (Experiment 1), in the 7 – 9 years old children, the physical congruency effect was reversed. The children were more accurate when the more numerous animal was presented on the screen smaller than the less numerous animal. However, we observed no physical congruency effect in the reaction times. In the older children (11-12 years old), we also did not observe physical congruency effect in the accuracies and reaction times. We elaborate on these unexpected findings in the general discussion.

Discussion

Humans have an innate number sense – the ability to extract numerical information from the environment rapidly and effortlessly (Dehaene, 2011). Most of the research on the number sense uses abstract stimuli (e.g., dot configurations), which are well-controlled for their visual confounds (e.g., size, total area, etc.). While this has resulted in valuable insights into how we perceive numbers, numerosity perception of familiar objects remains rarely examined and poorly understood. However, familiarity with an object may facilitate the individuation of each object in a set, thus enhancing the extraction of number resulting in less interference from non-numerical dimensions, such as the total surface an item occupies (i.e., the physical congruency between the number of items and the total space they occupy). Consequently, familiarity with an object can influence numerosity judgements in a bottom-up manner. Conversely, unlike abstract objects, familiar objects are also characterized by conceptual knowledge of that object - for instance, semantic knowledge about the object's actual size in real life (e.g., a zebra is larger than a frog). The latter could be an additional dimension that affects the numerical decision process in a top-down manner (i.e., the conceptual congruency between an object’s real-life size and its size in a set).

Although previous studies examining numerosity decisions with abstract and familiar objects (Odic & Oppenheimer, 2023; Reynvoet et al., 2019) observed similar performance signatures, it remains to be determined whether the cognitive mechanisms involved in the numerosity judgements of abstract and familiar objects are the same or not. Therefore, in the current study, we aimed to disentangle the physical (bottom-up) and conceptual (top-down) interference effects in numerosity judgments of familiar objects. To do so, we asked participants to perform an animal comparison task (i.e., which animal is more numerous), where the physical congruency (congruent = more numerous animal occupies larger total area, incongruent = the more numerous animal occupies smaller total area) and conceptual congruency (congruent = 5 frogs vs 10 zebras, incongruent = 5 zebras vs 10- frogs), were manipulated. We examined this question cross-sectionally in three populations: adults, children at the beginning (7 – 9 years old) and children at the end of primary school (11 – 12 years old).

[INSERT FIGURE 2 HERE]

Concerning the conceptual congruency effects, the current study is, to the best of our knowledge, the first one to demonstrate that conceptual knowledge can interfere with numerosity decisions with familiar objects. In both Experiments 1 and 2, while controlling for the physical size of the object on the screen, we observed a conceptual congruency effect indicating better performance when the more numerous animal was also larger in real life (e.g., 10 zebras and 5 frogs) than when the more numerous animal was smaller in real life (e.g., 10 frogs and 5 zebras). The conceptual congruency effect was already present in our youngest group (i.e., 7 – 9 years old) and increased throughout development (see Figure 2 and Figure 2S). Although the real-world size of the animal was irrelevant to the task, the presence of a conceptual congruency effect indicates that children and adults activate the real-world size of the object automatically. Automatic activation of the real-world size has been demonstrated before (Glicksman et al., 2016; Konkle & Olivia, 2012). For instance, Konkle and Oliva (2012) presented two familiar objects at different visual sizes on the screen and asked participants to decide which object was bigger (or smaller) on the screen. The results showed faster responses when the real-world size was congruent with the size on the screen (e.g., large elephant – small leaf) compared to when both size dimensions were incongruent (small elephant – large leaf; see also Glicksman et al., 2016). The current study supports and extends this observation by showing that real-world size of an object is indeed

automatically activated, also during decisions on the numerosity of a set of objects. In numerosity perception literature, there is currently an ongoing debate whether the origin of the congruency effects between numbers and non-numerical dimensions (such as size and space) occur early at the perceptual level or later at the decisional level (e.g., Picon et al., 2019; Lourenco & Aulet, 2023; Zorzi & Testolin, 2017). We would argue that the competition between numerosity and conceptual information is most likely situated at the decisional level. In other words, participants perform worse on incongruent trials because task-irrelevant top-down conceptual information (here, the animal's size in real life) competes with the bottom-up non-numerical information at a decision level. Altogether, these results suggest an interaction between conceptual and perceptual processing, an idea that has also been put forward in embodied cognition (Lakoff & Núñez, 2000).

Concerning the physical congruency, the results of Experiment 1 with adults and Experiment 2 with children show that the physical congruency effect in numerosity judgements of familiar objects is observed only in adults but not in children. At first glance, these results are rather unexpected in light of previous research with abstract stimuli (i.e., dot patterns). Concretely, previous results typically show a decrease in the physical congruency effect as age increases (Piazza et al., 2018; Rousselle & Noël, 2008; Soltész et al., 2010; Szucs et al., 2013): an observation framed as the filtering hypothesis (Piazza et al., 2018). According to the filtering hypothesis, as an individual's age increases, there is an increased ability to focus on the relevant number dimension and inhibit other interfering dimensions. Therefore, our findings that the physical congruency effect increases with age seem to contradict this hypothesis.

One way to explain the absence of a physical congruency effect in children is the facilitated individuation process. Because the objects can be verbally labelled, there is less room non-numerical information can interfere less. However, it is unclear why this physical congruency effect would re-appear in adults because they too, can use these verbal labels. Alternatively, a closer look at the data reveals an inconsistent mapping between non-numerical dimensions and numerosity through development. Concretely, while a standard congruency effect was present in adults (i.e., better performance on congruent than incongruent trials), the effect was not observed in the 11 – 12 years olds, and in the 7 – 9 years olds group, there was some evidence for a small reversed physical congruency effect (see Figure 2). Put differently, our youngest group performs better when the more numerous collections of animals occupy less total surface and are displayed

smaller (= incongruent trials). Defever and colleagues (2013) observed similar inconsistent mappings through development. In 6-7-year-old children, the researchers reported a bimodal distribution of the most pronounced congruency effects. Whereas some children performed better when the more numerous dot pattern had a larger total surface, density, and diameter but a smaller convex hull, other children were more accurate when the more numerous dot pattern had a smaller surface, density, and diameter but a larger convex hull. However, throughout development, more and more children start to associate a larger surface with more numerous dot patterns.

The gradual shift from relying on surface, density, and diameter to relying more on the total area could be the consequence of a local to global shift in perception (Nayar et al., 2015). Local processing is based on selective attention to individual elements, whereas global processing involves forming a coherent global structure (Navon, 1977). It is suggested that younger children rely more on local perceptual strategies, whereas global perception develops at a later age (e.g., Nayar et al., 2015). Consequently, the reversed physical congruency effect might be because younger children (7-9 years old) pay more attention to the individual objects and their size and since smaller objects are estimated as more numerous (Gebuis & Reynvoet, 2012), this may result in better performance in trials where the more numerous animal is smaller (i.e., the incongruent trials). However, through development, as attention shifts to the global structure, the performance on trials where the more numerous animal occupies a larger total area (i.e., the congruent trials) increases.

In the current study, we took a developmental cross-sectional approach to disentangle conceptual and physical congruency in numerosity judgements of familiar objects. In line with previous studies, our results showed a physical congruency effect on comparative judgements of familiar objects. More importantly, for the first time in numerosity perception literature, we also observed a conceptual congruency effect, suggesting that semantic information stored in the long-term memory can interfere with the numerosity decision of familiar objects. The effects show distinct signatures and are subject to developmental change. In sum, our results demonstrate that both (bottom-up) non-numerical characteristics and (top-down) conceptual knowledge may influence numerosity comparison. These findings further suggest that our number sense is not a standalone core system for number but is possibly embedded in a larger network where both low-level perceptual and higher-level conceptual information can influence individuals' numerosity

judgements. Additionally, based on our results, future developmental research should account for these top-down influences by shifting from abstract to real-world stimuli.

Declarations

Conflict of interest. The authors declare no conflict of interest.

Ethics approval. This research was conducted and written according to APA ethical standards. The ethical committee at the KU Leuven approved the study's design, methods, and protocol (G-2021-4583). Prior to participation, written informed consent was sought from each participant or their legal representative.

Contributions:

Study design conceptualization: BR

Design, methods, and experimental protocol implementation: MM & BR

Data curation and Data analysis: MM with input from BR

Manuscript preparation and revisions: MM & BR

Funding:

MM: Luxembourg National Research Fund (C20/SC/14629532/SymNumDev)

BR: KU Leuven Research Fund (C14/23/057)

Acknowledgements:

The authors would like to thank Lotte Dehaes and Silke Geerens for their help with the data collection.

References:

Aulet, L. S., & Lourenco, S. F. (2021). Numerosity and cumulative surface area are perceived holistically as integral dimensions. *Journal of Experimental Psychology: General*, *150*(1), 145–156. https://doi.org/10.1037/xge0000874

Brady, T. F., & Störmer, V. S. (2022). The Role of Meaning in Visual Working Memory: Real-World Objects, But Not Simple Features, Benefit From Deeper Processing. *Journal of Experimental Psychology: Learning Memory and Cognition*, *48*(7), 942–958. https://doi.org/10.1037/xlm0001014

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272. https://doi.org/10.1038/s41562-020-00946-0

Clarke, S., & Beck, J. (2020). The number sense represents (rational) numbers. *Behavioral and Brain Sciences*, 4–5. http://eprints.soton.ac.uk/252625/2/bbs.html

Clayton, S., Gilmore, C., & Inglis, M. (2015). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica*, *161*, 177–184. https://doi.org/10.1016/j.actpsy.2015.09.007

De Marco, D., & Cutini, S. (2020). Introducing CUSTOM: A customized, ultraprecise, standardization-oriented, multipurpose algorithm for generating nonsymbolic number stimuli. *Behavior Research Methods*, *52*(4), 1528–1537. https://doi.org/10.3758/s13428-019-01332-z

Defever, E., Reynvoet, B., & Gebuis, T. (2013). Task- and age-dependent effects of visual stimulus properties on children's explicit numerosity judgments. *Journal of Experimental Child Psychology*, *116*(2), 216–233. https://doi.org/10.1016/j.jecp.2013.04.006

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265. https://doi.org/10.1016/j.cognition.2015.05.016

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Fuhs, M. W., & Mcneil, N. M. (2013). ANS acuity and mathematics ability in preschoolers from low-income homes: Contributions of inhibitory control. *Developmental Science*, *16*(1), 136–148. https://doi.org/10.1111/desc.12013

Gebuis, T., Cohen Kadosh, R., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta Psychologica*, *171*, 17–35. https://doi.org/10.1016/j.actpsy.2016.09.003

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods*, *43*(4), 981–986. https://doi.org/10.3758/s13428-011-0097-5

Gebuis, T., & Reynvoet, B. (2012). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642–648. https://doi.org/10.1037/a0026218

Gennari, G., Dehaene, S., Valera, C., & Dehaene-Lambertz, G. (2023). Spontaneous supra-modal

encoding of number in the infant brain. *Current Biology*, *33*(10), 1906-1915.e6. https://doi.org/10.1016/j.cub.2023.03.062

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., Simms, V., & Inglis, M. (2013). Individual Differences in Inhibitory Control, Not Non-Verbal Number Acuity, Correlate with Mathematics Achievement. *PLoS ONE*, *8*(6), 1–9. https://doi.org/10.1371/journal.pone.0067374

Gliga, T., Volein, A., & Csibra, G. (2010). Verbal Labels Modulate Perceptual Object Processing in 1-Year-Old Children Teodora. *Journal of Cognitive Neuroscience*, *22*(12), 2781–2789. https://doi.org/10.1162/jocn.2010.21427

Gliksman, Y., Itamar, S., Leibovich, T., Melman, Y., & Henik, A. (2016). Automaticity of Conceptual Magnitude. *Scientific Reports*, *6*(November 2015), 1–7. https://doi.org/10.1038/srep21446

Guillaume, M., Schiltz, C., & Van Rinsveld, A. (2020). NASCO: A new method and program to generate dot arrays for non-symbolic number comparison tasks. *Journal of Numerical Cognition*, *6*(1), 129–147. https://doi.org/10.5964/jnc.v6i1.231

Halberda, J., & Feigenson, L. (2008). Developmental Change in the Acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-Year-Olds and Adults. *Developmental Psychology*, *44*(5), 1457–1465. https://doi.org/10.1037/a0012682

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150. https://doi.org/10.3389/fnins.2014.00150

Jackson, M. C., & Raymond, J. E. (2008). Familiarity Enhances Visual Working Memory for Faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556–568. https://doi.org/10.1037/0096-1523.34.3.556

Konkle, T., & Oliva, A. (2012). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 23–37. https://doi.org/10.1037/a0020413

Lakoff, G., & Núñez, R. (2000). Where mathematics comes from: How the embodied mind brings mathematics into being. New York, NY: Basic Books.

Le-Hoa Võ, M., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 72–81. https://doi.org/10.1111/nyas.12667

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*. https://doi.org/10.1017/S0140525X16000960

Lourenco, S. F., & Aulet, L. S. (2023). A Theory of Perceptual Number Encoding. *Psychological Review*, *130*(1), 155–182. https://doi.org/10.1037/rev0000380

Lupyan, G., & Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel

stimuli. *Current Biology*, *18*(10), R410–R412. https://doi.org/10.1016/j.cub.2008.02.073

Marinova, M., Fedele, M., & Reynvoet, B. (2021). Weighted numbers. *Behavioral and Brain Sciences*, *44*(e196), 2–5. https://doi.org/https://doi.org/10.1017/S0140525X21001059

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

Nayar, K., Franchak, J., Adolph, K., & Kiorpes, L. (2015). From local to global processing: The development of illusory contour perception. *Journal of Experimental Child Psychology*, *131*, 38–55. https://doi.org/10.1016/j.jecp.2014.11.001

Nieder, A. (2021). Neuroethology of number sense across the animal kingdom. *The Journal of Experimental Biology*, *224*(6), jeb218289. https://doi.org/10.1242/jeb.218289

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, *32*, 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550

Odic, D., & Oppenheimer, D. M. (2023). Visual numerosity perception shows no advantage in real-world scenes compared to artificial displays. *Cognition*, *230*(April 2022), 105291. https://doi.org/10.1016/j.cognition.2022.105291

Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition*, *181*(August), 35–45. https://doi.org/10.1016/j.cognition.2018.07.011

Reynvoet, B., Ribner, A. D., Elliott, L., Van Steenkiste, M., Sasanguie, D., & Libertus, M. E. (2021). Making sense of the relation between number sense and math. *Journal of Numerical Cognition*, *7*(3), 308–327. https://doi.org/10.5964/jnc.6059

Reynvoet, B., Vos, H., & Henik, A. (2019). Comparative Judgment of Familiar Objects Is Modulated by Their Size. *Experimental Psychology*, *65*(6), 353–359. https://doi.org/10.1027/1618-3169/a000418

Rousselle, L., & Noël, M. P. (2008). The Development of Automatic Numerosity Processing in Preschoolers: Evidence for Numerosity-Perceptual Interference. *Developmental Psychology*, *44*(2), 544–560. https://doi.org/10.1037/0012-1649.44.2.544

Smets, K., Moors, P., & Reynvoet, B. (2016). Effects of Presentation Type and Visual Control in Numerosity Discrimination: Implications for Number Processing? *Frontiers in Psychology*, *7*(February), 1–14. https://doi.org/10.3389/fpsyg.2016.00066

Smets, K., Sasanguie, D., Szücs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology*, *27*(3), 310–325. https://doi.org/10.1080/20445911.2014.996568

Sokolowski, H. M., Fias, W., Bosah Ononye, C., & Ansari, D. (2017). Are numbers grounded in a general magnitude processing system? A functional neuroimaging meta-analysis. *Neuropsychologia*, *105*(November 2016), 50–69. https://doi.org/10.1016/j.neuropsychologia.2017.01.019

Soltész, F., Szucs, D., & Szucs, L. (2010). Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: A developmental study. *Behavioral and Brain Functions*, *6*, 1–14. https://doi.org/10.1186/1744-9081-6-13

Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222–233. https://doi.org/10.1016/j.cognition.2017.07.004

Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative

models. *Nature Neuroscience*, *15*(2), 194–196. https://doi.org/10.1038/nn.2996

Szucs, e. D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology*, *4*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2013.00444

Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2020). Visual sense of number vs. sense of magnitude in humans and machines. *Scientific Reports*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-66838-5

Van Den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E. J., Derks, K., Dablander, F., Gronau, Q. F., Kucharský, Š., Gupta, A. R. K. N., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E. J. (2020). A tutorial on conducting and interpreting a bayesian ANOVA in JASP. *Annee Psychologique*, *120*(1), 73–96. https://doi.org/10.3917/anpsy1.201.0073

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E. J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-020-01798-5

Van Rinsveld, A., Wens, V., Guillaume, M., Beuel, A., Gevers, W., De Tiège, X., & Content, A. (2021). Automatic Processing of Numerosity in Human Neocortex Evidenced by Occipital and Parietal Neuromagnetic Responses. *Cerebral Cortex Communications*, *March*, 1–12. https://doi.org/10.1093/texcom/tgab028

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wilkey, E. D., & Ansari, D. (2020). Challenging the neurobiological link between number sense and symbolic numerical abilities. *Annals of the New York Academy of Sciences*, *1464*(1), 76–98. https://doi.org/10.1111/nyas.14225

Xie, W., & Zhang, W. (2017). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory and Cognition*, *45*(4), 677–689. https://doi.org/10.3758/s13421-016-0679-7

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), 1–11. https://doi.org/10.1016/S0010-0277(99)00066-9

Tables and Figures

Table 1.Mean accuracies (proportion correct) and the median reaction times (in seconds) with their corresponding standard deviations (SD), depicted per condition for Experiment 1 with adults.

			Conceptual Congruency					Physical Congruency			
			Mean Accuracy		Median RT			Mean Accuracy		Median RT	
			Congruent	Incongruent	Congruent	Incongruent		Congruent	Incongruent	Congruent	Incongruent
Animal Pair	Ratio	Range									
Cow-Mouse	1.5	Subitizing (trials 2-3)	0.91 (0.11)	0.87 (0.13)	0.80 (0.15)	0.83 (0.17)	Cow-Zebra	0.90 (0.11)	0.92 (0.11)	0.79 (0.12)	0.82 (0.12)
		Non-subitizing (trials 6-9)	0.90 (0.13)	0.88 (0.13)	0.78 (0.15)	0.79 (0.15)		0.81 (0.15)	0.80 (0.15)	0.83 (0.17)	0.84 (0.15)
	2	Subitizing (trials 1-2)	0.92 (0.10)	0.89 (0.12)	0.77 (0.12)	0.78 (0.14)		0.93 (0.10)	0.87 (0.15)	0.73 (0.12)	0.81 (0.12)
		Non-subitizing (trials 5-10)	0.92 (0.11)	0.95 (0.10)	0.76 (0.15)	0.74 (0.14)		0.92 (0.10)	0.85 (0.15)	0.77 (0.11)	0.80 (0.12)
Zebra-Frog	1.5	Subitizing (trials 2-3)	0.93 (0.09)	0.85 (0.17)	0.77 (0.11)	0.83 (0.13)	Mouse-Frog	0.89 (0.11)	0.87 (0.15)	0.82 (0.14)	0.87 (0.12)
		Non-subitizing (trials 6-9)	0.87 (0.14)	0.71 (0.20)	0.79 (0.13)	0.90 (0.15)		0.91 (0.11)	0.69 (0.20)	0.75 (0.14)	0.86 (0.18)
	2	subitizing (trials 1-2)	0.92 (0.11)	0.91 (0.11)	0.73 (0.11)	0.76 (0.13)		0.89 (0.12)	0.92 (0.09)	0.74 (0.12)	0.76 (0.12)
		Non-subitizing (trials 5-10)	0.91 (0.12)	0.84 (0.18)	0.74 (0.12)	0.85 (0.15)		0.93 (0.09)	0.88 (0.13)	0.74 (0.13)	0.81 (0.15)

Table 2. Mean accuracies (proportion correct) and the median reaction times (in seconds) with their corresponding standard deviations (SD), depicted per condition and age group for Experiment 2.

			Conceptual Congruency Zebra – Frog Pairs				Physical Congruency Cow – Zebra Pairs			
			Mean Accuracy		Median RT		Mean Accuracy		Median RT	
Age Group	Ratio	Range	Congruent	Incongruent	Congruent	Incongruent	Congruent	Incongruent	Congruent	Incongruent
7 – 9 years-olds N = 53	1.5	Subitizing (trials 2-3)	0.90 (0.15)	0.87 (0.15)	1.95 (0.38)	1.95 (0.42)	0.84 (0.18)	0.86 (0.15)	1.95 (0.37)	1.90 (0.37)
		Non-subitizing (trials 6-9)	0.85 (0.17)	0.71 (0.25)	1.93 (0.38)	2.05 (0.46)	0.71 (0.24)	0.81 (0.18)	1.86 (0.49)	1.91 (0.41)
	2	Subitizing (trials 1-2)	0.87 (0.16)	0.89 (0.14)	1.78 (0.37)	1.79 (0.45)	0.91 (0.17)	0.90 (0.16)	1.72 (0.33)	1.81 (0.42)
		Non-subitizing (trials 5-10)	0.85 (0.19)	0.83 (0.16)	1.82 (0.38)	1.96 (0.43)	0.86 (0.19)	0.87 (0.19)	1.89 (0.45)	1.84 (0.38)
11 – 12 years-olds N = 84	1.5	Subitizing (trials 2-3)	0.94 (0.09)	0.93 (0.10)	1.43 (0.41)	1.48 (0.37)	0.93 (0.09)	0.94 (0.10)	1.46 (0.36)	1.47 (0.39)
		Non-subitizing (trials 6-9)	0.88 (0.15)	0.78 (0.19)	1.50 (0.39)	1.73 (0.46)	0.85 (0.13)	0.89 (0.13)	1.58 (0.44)	1.54 (0.40)
	2	Subitizing (trials 1-2)	0.94 (0.10)	0.95 (0.08)	1.35 (0.32)	1.32 (0.36)	0.94 (0.10)	0.93 (0.11)	1.29 (0.36)	1.36 (0.34)
		Non-subitizing (trials 5-10)	0.93 (0.13)	0.91 (0.13)	1.41 (0.40)	1.55 (0.43)	0.93 (0.09)	0.93 (0.11)	1.44 (0.39)	1.43 (0.35)

Figure 1A. Example of an experimental trial of ratio 2 (5 vs 10), as a function of condition and congruency in Experiment 1.

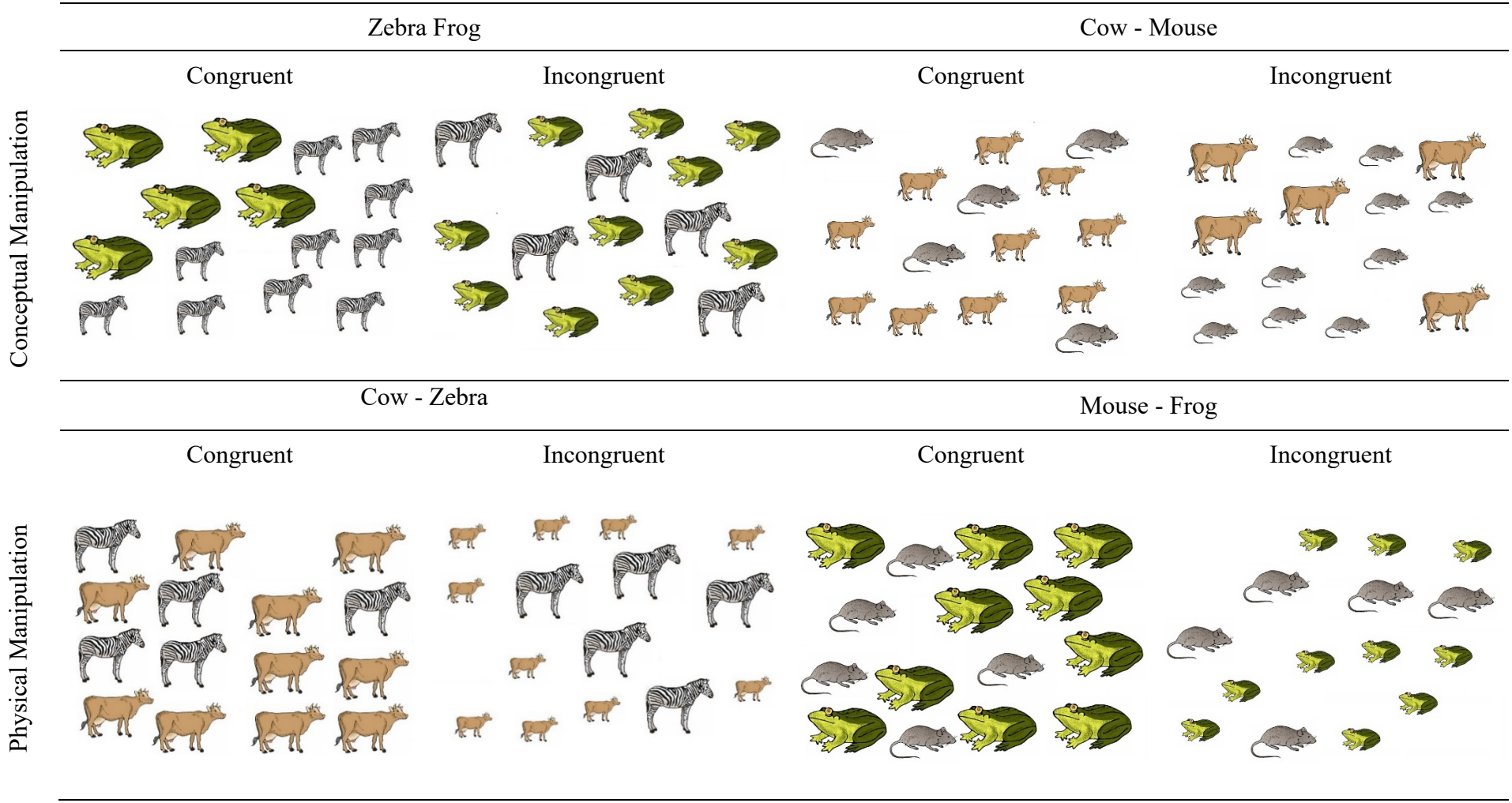
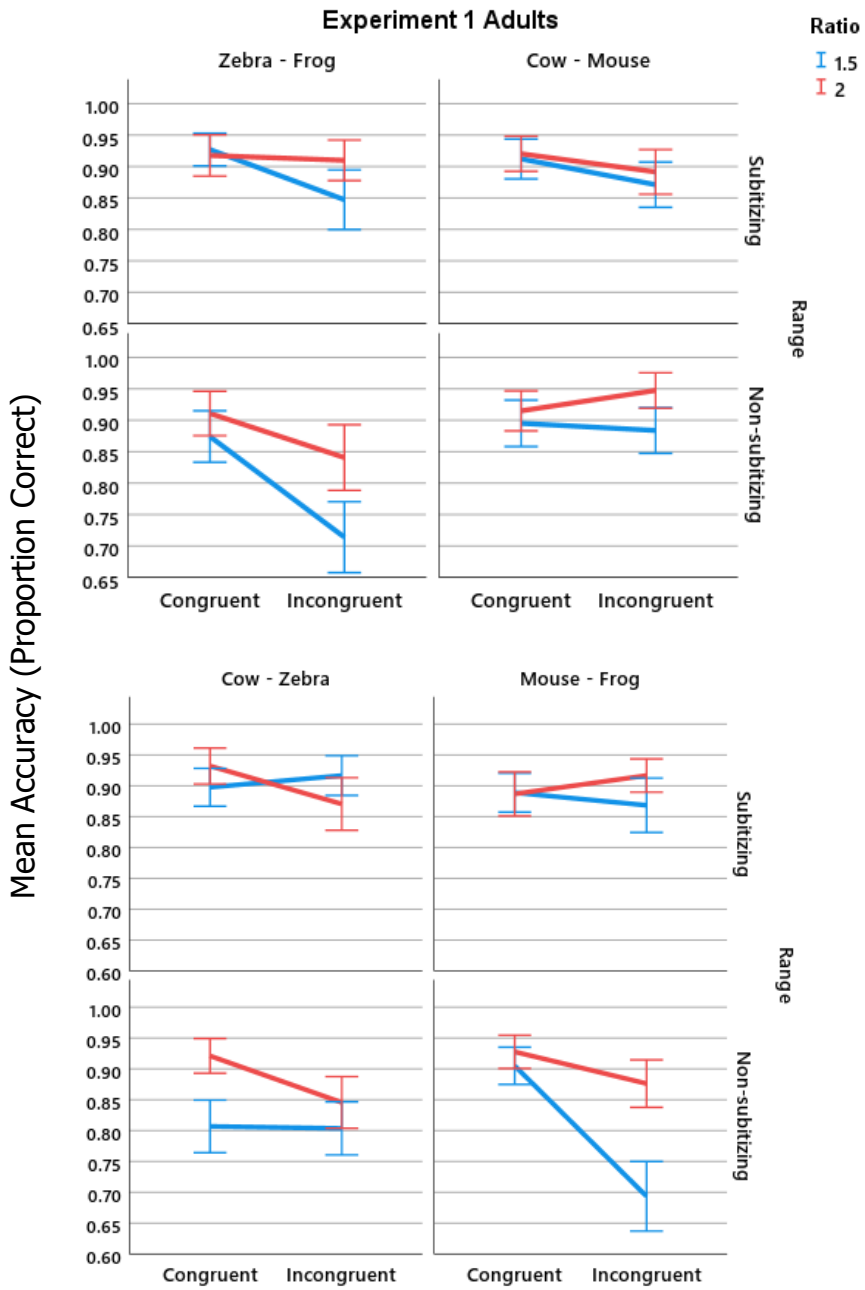
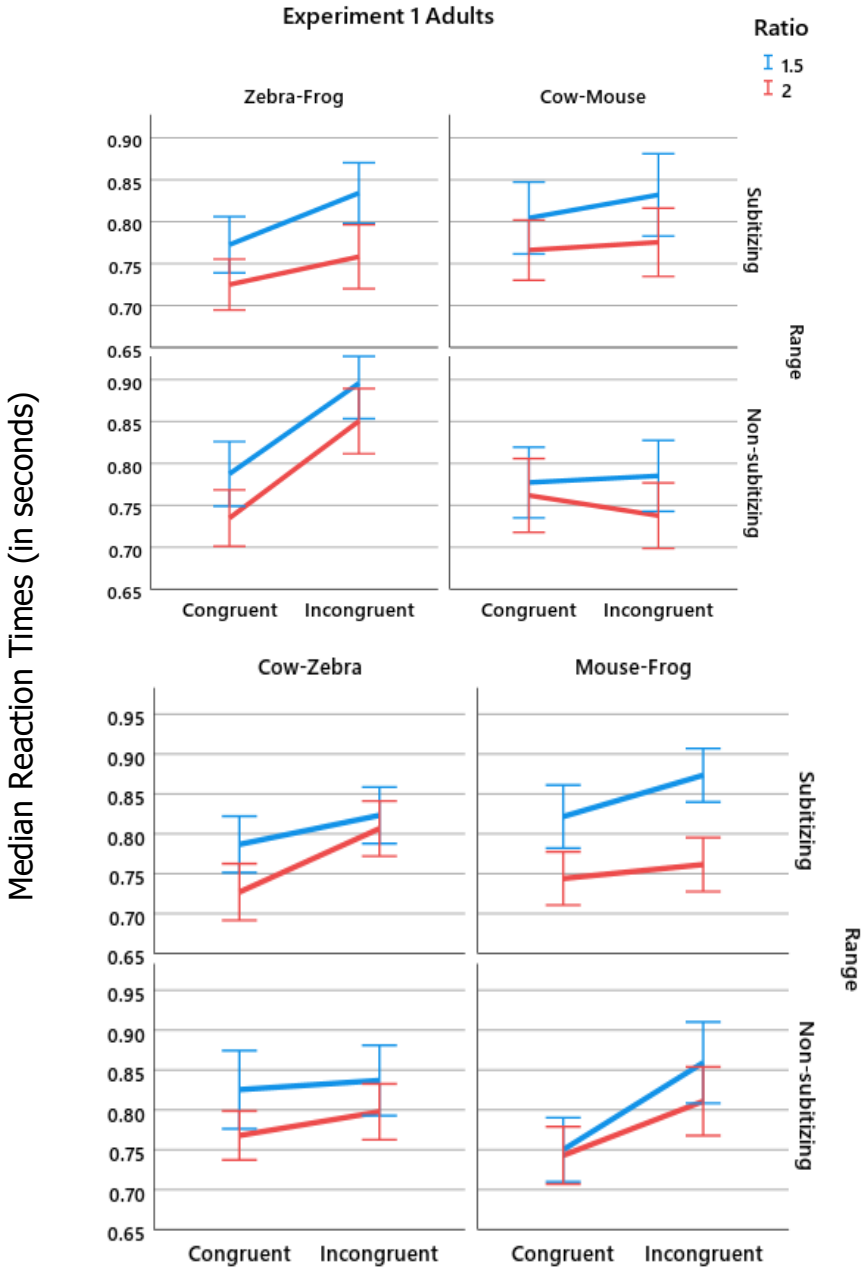
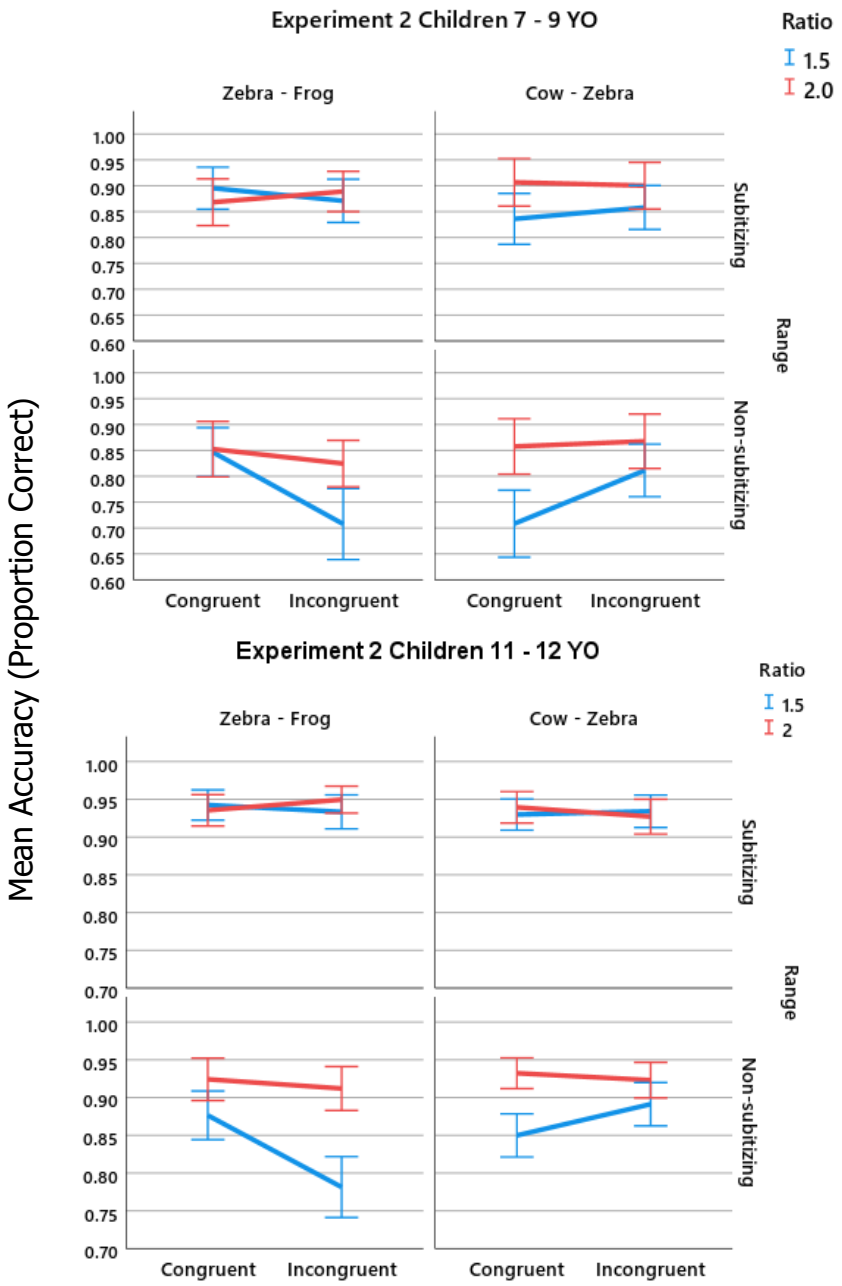


Figure1B. Congruency effects in the mean accuracy and median reaction times, depicted per animal pair, ratio and range for Experiments 1 and 2. Vertical bars denote 95% Confidence Intervals.






Mean Accuracy (Proportion Correct)
Experiment 2 Children 7 - 9 YO
Ratio
1.5
2.0
Zebra - Frog
Cow - Zebra
Subitizing
Non-subitizing
Range
Congruent
Incongruent
Experiment 2 Children 11 - 12 YO
Ratio
1.5
2
Zebra - Frog
Cow - Zebra
Subitizing
Non-subitizing
Range
Congruent
Incongruent

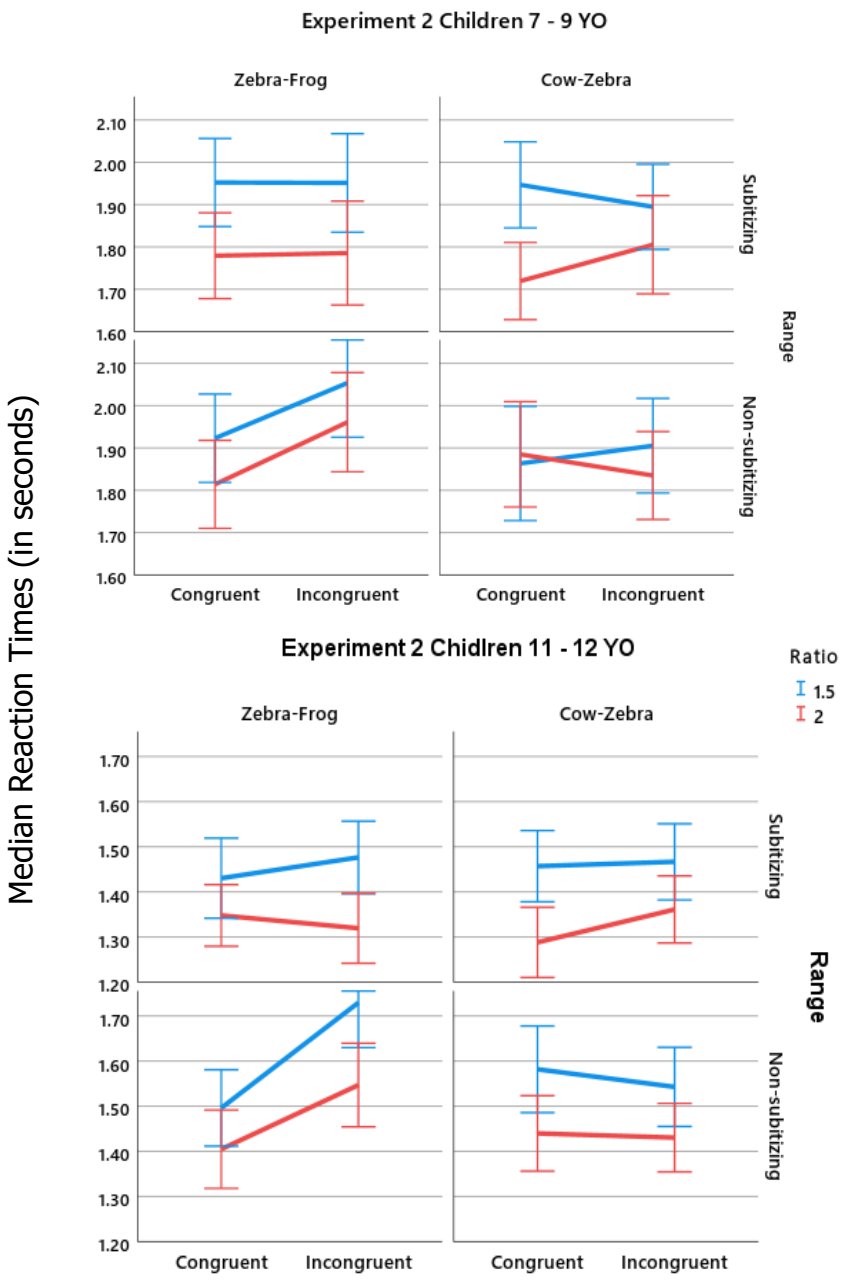
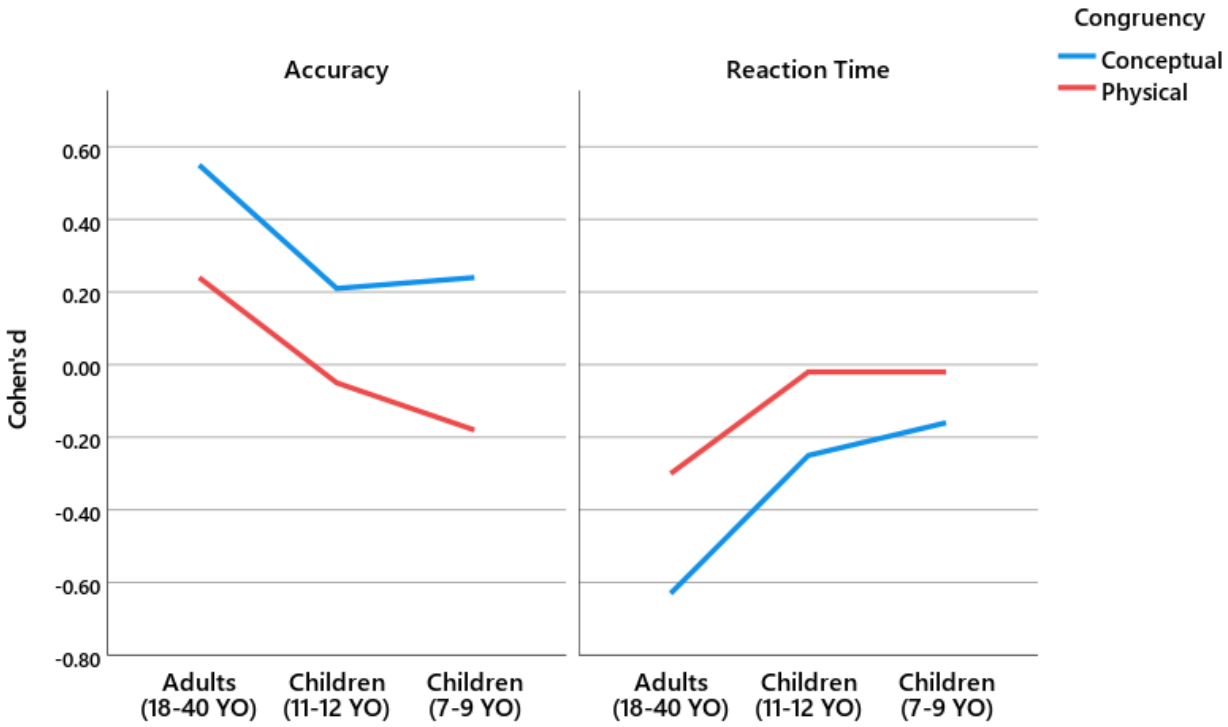

Median Reaction Times (in seconds)
Experiment 2 Children 7 - 9 YO
Ratio
1.5
2
Zebra-Frog
Cow-Zebra
Subitizing
Non-subitizing
Range
Congruent
Incongruent
Experiment 2 Chidlren 11 - 12 YO
Ratio
1.5
2
Zebra-Frog
Cow-Zebra
Subitizing
Non-subitizing
Range
Congruent
Incongruent

Figure 2. Visual depiction of the size of the conceptual (zebra - frog) and physical (cow - zebra) congruency effects in each age group.



.

Note 1. The effect size is depicted as Cohen's d following two-tailed paired samples -test (congruent vs incongruent). For the interpretation, negative values inaccuracy indicates a reversed congruency effect, while positive values indicate a standard congruency effect. In the reaction times, negative values indicate standard congruency effect (i.e., faster RTs for congruent trials)

Are Three Zebras More than Three Frogs: Examining Conceptual and Physical Congruency in Numerosity Judgements of Familiar Objects

SUPPLEMENTARY BAYESIAN ANALYSES

Mila Marinova [1,2,3]& Bert Reynvoet [2,3]

[1] Institute of Cognitive Science and Assessment, Department of Behavioural and Cognitive Sciences, Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Esch-Belval, Luxembourg.

[2]Brain and Cognition, KU Leuven, Leuven, Belgium

[3]Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Kortrijk Belgium.

Correspondence concerning this article should be addressed to Bert Reynvoet, Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. Tel: +32 56246177, bert.reynvoet@kuleuven.be

ORCID:

Mila Marinova: https://orcid.org/0000-0002-6875-7742

Bert Reynvoet: https://orcid.org/0000-0002-4898-2475

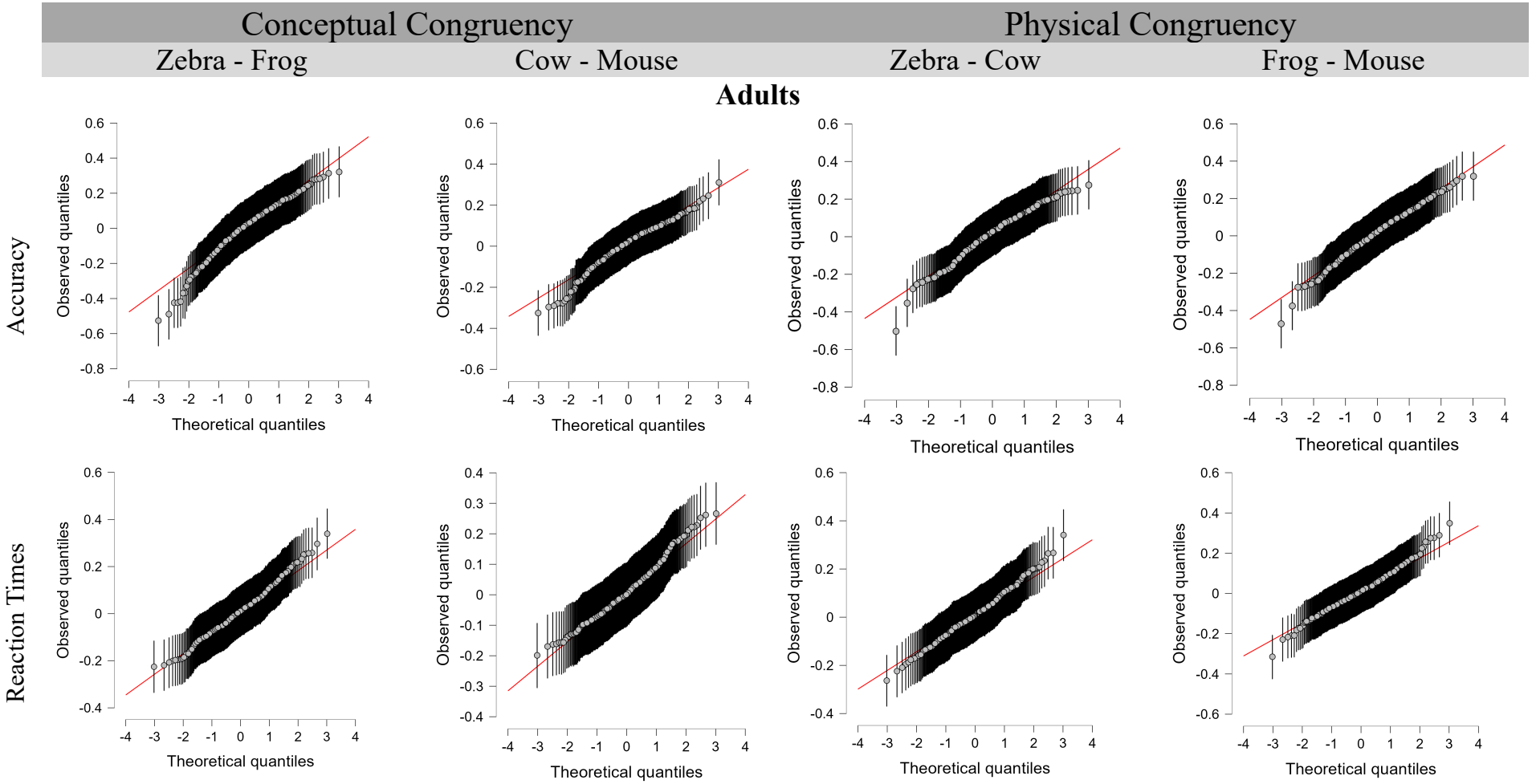
SUPPLEMENTARY BAYESIAN ANALYSES

We interpret the Bayes factors (BF), which is the ratio of the H_1 and H_0 likelihoods (BF_{10}). For analyses using more factors such as ANOVA (i.e, Bayesian Model Comparison), it is recommended to report the model-averaged results such as the BFInclusion (BF_{Incl}). The latter reflects the predictive strength of the effect for the data by comparing all models that include the effect of interest to those without this effect (Van Den Bergh et al., 2020; van Doorn et al., 2020; Wagenmakers et al., 2018). By convention, the values of the BF are interpreted as the following: “anecdotal” (for values between < 1 and 3), “moderate” (for values between 3 and 10), “strong” (for values between 10 and 30), “very strong” (for values between 30 and 100), and “extreme” (for values > 100). To obtain both classical and Bayesian results, we used JASP statistical package v 0.18.3.0 (https://jasp-stats.org/) using default priors for the Bayesian Model Comparison (i.e., r scale fixed effects = 0.5, random effects = 1).

Normality assumption checks for Bayesian ANOVA are depicted in Figure 1.S. The Outcomes of the Bayesian ANOVA (i.e., Bayesian Model Comparison—all models) for each experiment and condition are depicted in Table 1S below, alongside the Analysis of Effects Across All Models. For presentation purposes, we depicted only the first three best models. To disentangle the interactions, we used a Bayesian paired-samples t-test and/or performed a post-hoc Bayesian Model Comparison.

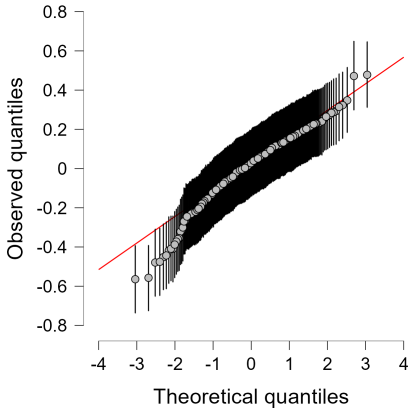
The Figure 2S depicts the mean size of the conceptual and physical congruency effects in each participant's accuracy and reaction times data computed as the difference between the congruent and incongruent trials.

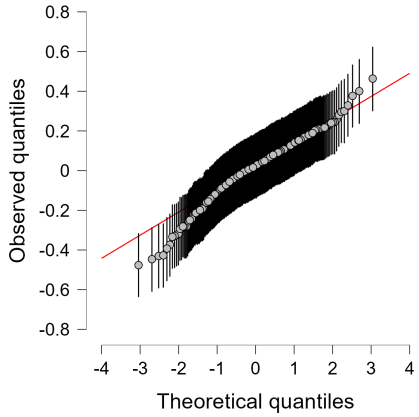
Figure 1S.Q-Q plot visualising the observed residuals against the residuals of a standard normal distribution for each animal pair and depicted per age group. The data is perfectly normally distributed if all the points are on the red line. The vertical bars through each point represent 95% Central Credible Interval.



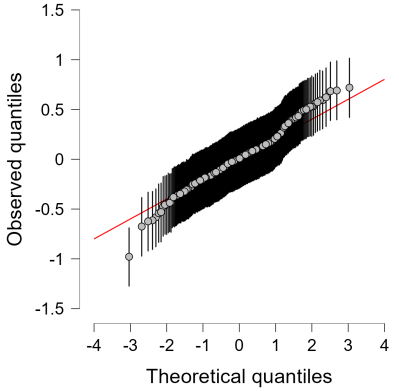
Children Aged 7 – 9 years

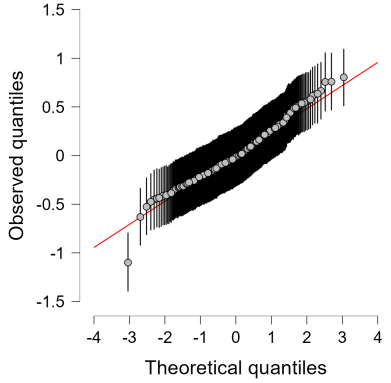
Accuracy





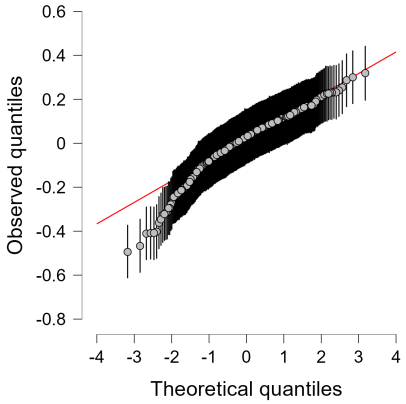
Reaction Times

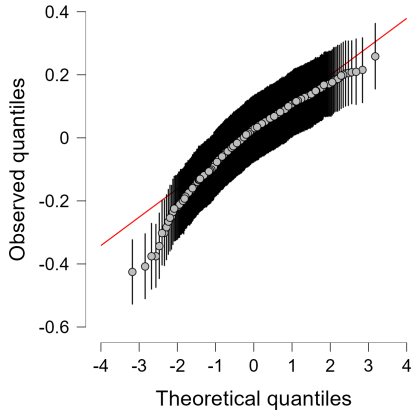




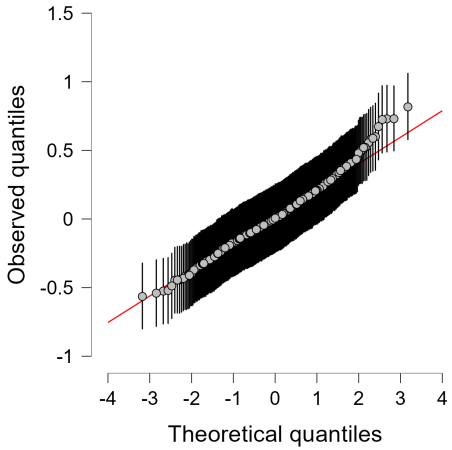
Children Aged 11 – 12 years

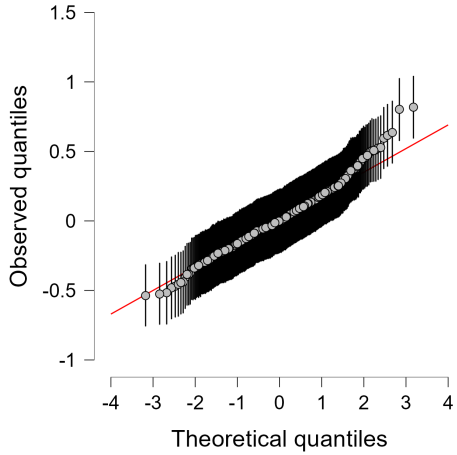
Accuracy





Reaction Times





SUPPLEMENTARY ANALYSES

Table 1S. Bayesian Model Comparison, Analysis of Effects (i.e., model-averaged results) and the post-hoc t-test for each animal pair, and each age group performed on the accuracies and the reaction times.

Conceptual Congruency Zebra-Frog Accuracy Adults

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	9.103×10^{6}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	1.881×10^{12}	Ratio	1755.31
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	1.565×10^{12}	Range	3664.85
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	1.062×10^{12}	Congruency*Ratio	35.78
		Congruency*Range	2.83
		Ratio*Range	3.99
		Congruency*Ratio*Range	1.33

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	690620.04
CongruentRatio2 vs InCongruentRatio2	3.31
CongruentSubitizing vs InCongruentSubitizing	1.88
CongruentNon-Subitizing vs InCongruentNon-Subitizing	2078.54
Ratio1.5Subitizing vs Ratio2Subitizing	0.74
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	114.07

Conceptual Congruency Zebra-Frog Reaction Times Adults

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	1.692×10^{8}
Congruency + Ratio + Range + Congruency*Range	7.130×10^{14}	Ratio	3795.17
Congruency + Ratio + Range + Congruency*Range + Ratio*Range	1.694×10^{14}	Range	6159.64
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range	1.654×10^{14}	Congruency*Ratio	0.55
		Congruency*Range	91.03
		Ratio*Range	0.56
		Congruency*Ratio*Range	0.30

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentSubitizing vs InCongruentSubitizing	24.51
CongruentNon-Subitizing vs InCongruentNon-Subitizing	1.16×10^{8}

Conceptual Congruency Cow – Mouse Accuracy Adults

Models	BF_{10}	Analysis of Effects	BF_{incl}
Null model (incl. subject and random slopes)	1.00	Congruency	0.18
Ratio	4.11	Ratio	1.83
Ratio + Range	1.18	Range	0.18
Congruency + Ratio	1.17	Congruency*Ratio	0.22
		Congruency*Range	0.20
		Ratio*Range	0.23
		Congruency*Ratio*Range	0.05

Conceptual Congruency Cow–Mouse Reaction Times Adults

Models	BF_{10}	Analysis of Effects	BF_{incl}
Null model (incl. subject and random slopes)	1.00	Congruency	0.16
Ratio + Range	40553.25	Ratio	1183.55
Ratio + Range + Ratio*Range	11268.10	Range	10.80
Congruency + Ratio + Range	7053.52	Congruency*Ratio	0.28
		Congruency*Range	0.27
		Ratio*Range	0.61
		Congruency*Ratio*Range	0.04

Physical Congruency Zebra-Cow Accuracy Adults

Models	BF_{10}	Analysis of Effects	BF_{incl}
Null model (incl. subject and random slopes)	1.00	Congruency	40.18
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	5.294×10^{6}	Ratio	2438.61
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	1.427×10^{6}	Range	4506.48
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range + Congruency*Ratio*Range	288075.13	Congruency*Ratio	53.72
		Congruency*Range	0.69
		Ratio*Range	74.34
		Congruency*Ratio*Range	0.72

Two-tailed paired samples *t*-test	BF_{10}

CongruentRatio1.5 vs InCongruentRatio1.5	0.19
CongruentRatio2 vs InCongruentRatio2	319.50
Ratio1.5Subitizing vs Ratio2Subitizing	0.16
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	1732.23

Physical Congruency Zebra-Cow Reaction Times Adults

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)		Congruency	327.26
Congruency + Ratio + Range + Congruency*Ratio	642012.03	Ratio	1518.04
Congruency + Ratio + Range + Congruency*Range	544069.54	Range	1.20
Congruency + Ratio + Range	525124.97	Congruency*Ratio	1.88
		Congruency*Range	1.20
		Ratio*Range	0.39
		Congruency*Ratio*Range	0.17

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	0.74
CongruentRatio2 vs InCongruentRatio2	6221.93

Physical Congruency Frog – Mouse Accuracy Adults

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	5.738×10^{11}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range + Congruency*Ratio*Range	2.917×10^{17}	Ratio	1.967×10^{8}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	2.005×10^{16}	Range	1.068×10^{8}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range	2.738×10^{14}	Congruency*Ratio	25568.43
		Congruency*Range	1.666×10^{6}
		Ratio*Range	2462.42
		Congruency*Ratio*Range	258.01

Post-hoc Bayesian ANOVA per Range

Subitizing

There was no effect congruency, BF_{incl}= 0.17, or ratio, BF_{incl}= 0.34 or an interaction, BF_{incl}= 0.21

Non-Subitizing

There were effects congruency, BF_{incl}= 3.222×10^{11}, ratio, BF_{incl}= 8.153×10^{8} and an interaction, BF_{incl}= 419191.39

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	2.135×10^{8}
CongruentRatio2 vs InCongruentRatio2	2.85

Physical Congruency Frog – Mouse Reaction Times Adults

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	1.027×10^{6}
Congruency + Ratio + Range + Congruency*Range + Congruency*Range + Ratio*Range	4.266×10^{13}	Ratio	1.247×10^{8}
Congruency + Ratio + Range + Congruency*Range + Ratio*Range	3.798×10^{13}	Range	111.29
Congruency + Ratio + Range + Congruency*Range + Congruency*Range + Ratio*Range + Congruency*Ratio*Range	8.266×10^{12}	Congruency*Ratio	2.88
		Congruency*Range	34.53
		Ratio*Range	138.63
		Congruency*Ratio*Range	1.70

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	16234.64
CongruentRatio2 vs InCongruentRatio2	11.79
CongruentSubitizing vs InCongruentSubitizing	15.35
CongruentNon-Subitizing vs InCongruentNon-Subitizing	61230.33
Ratio1.5Subitizing vs Ratio2Subitizing	1.03×10^{7}
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	1.73

Conceptual Congruency Accuracy Age 7 – 9 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	38.28
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	37593.51	Ratio	7.82
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range + + Congruency*Ratio*Range	18064.15	Range	1246.09
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	13713.88	Congruency*Ratio	11.24
		Congruency*Range	7.05
		Ratio*Range	6,61
		Congruency*Ratio*Range	3.78

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	100.86
CongruentRatio2 vs InCongruentRatio2	0.15
CongruentSubitizing vs InCongruentSubitizing	0.15
CongruentNon-Subitizing vs InCongruentNon-Subitizing	11.55
Ratio1.5Subitizing vs Ratio2Subitizing	0.15
Ratio1.5NonSubitizing vs Ratio2NonSubitizing	16.55

Conceptual Congruency Reaction Times Age 7 – 9 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	4.79
Congruency + Ratio + Range + Congruency*Range	1.10×10^{7}	Ratio	19567.38
Congruency + Ratio + Range + Congruency*Range + Ratio*Range	408888.23	Range	7.53
Congruency + Ratio*Range	392009.90	Congruency*Ratio	0.37
		Congruency*Range	4.29
		Ratio*Range	0.77
		Congruency*Ratio*Range	0.11

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentSubitizing vs InCongruentSubitizing	0.15
CongruentNon-Subitizing vs InCongruentNon-Subitizing	109.69

Physical Congruency Accuracy Age 7 – 9 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	2.13
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	6.173×10^{7}	Ratio	331434.24
Congruency + Ratio + Range + Congruency*Ratio	5.515×10^{7}	Range	178.52
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	5.425×10^{7}	Congruency*Ratio	2.96
		Congruency*Range	1.43
		Ratio*Range	2.86
		Congruency*Ratio*Range	1.22

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	9.12
CongruentRatio2 vs InCongruentRatio2	0.15

Ratio1.5Subitizing vs Ratio2Subitizing	99.14
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	318322.62

Physical Congruency Reaction Times Age 7 – 9 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	0.09
Ratio + Range + Ratio*Range	37.51	Ratio	24.99
Ratio	34.98	Range	0.53
Ratio + Range	10.36	Congruency*Ratio	0.17
		Congruency*Range	0.15
		Ratio*Range	1.96
		Congruency*Ratio*Range	0.87

Two-tailed paired samples *t*-test	**BF_{10}**
Ratio1.5Subitizing vs Ratio2Subitizing	330.94
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	0.19

Conceptual Congruency Accuracy Age 11 – 12 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	2266.83
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range + Congruency*Ratio*Range	1.721×10^{17}	Ratio	7.228×10^{8}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	1.095×10^{17}	Range	1.138×10^{12}
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	1.003×10^{16}	Congruency*Ratio	62.27
		Congruency*Range	56.33
		Ratio*Range	109939.43
		Congruency*Ratio*Range	23.87

Post-hoc Bayesian ANOVA per Range

Subitizing

There was no effect congruency, BF_{incl}= 0.25, but a main effect of ratio, BF_{incl}= 66037, and no interaction, BF_{incl}= 0.73

Non-Subitizing

There were effects congruency, BF_{incl}= 8.56×10^{7}, ratio, BF_{incl}= 383822.22, and an interaction, BF_{incl}= 6.47

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	1.200×10^{6}
CongruentRatio2 vs InCongruentRatio2	636.62

Conceptual Congruency Reaction Times Age 11 – 12 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	4.81×10^{9}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range	7.123×10^{28}	Ratio	1.46×10^{11}
Congruency + Ratio + Range + Congruency*Ratio + Congruency*Range + Ratio*Range	1.350×10^{28}	Range	1.48×10^{12}
Congruency + Ratio + Range + Ratio*Range	1.281×10^{28}	Congruency*Ratio	12.39
		Congruency*Range	34393.95
		Ratio*Range	0.46
		Congruency*Ratio*Range	0.38

Two-tailed paired samples *t*-test	**BF_{10}**
CongruentRatio1.5 vs InCongruentRatio1.5	1.99×10^{6}
CongruentRatio2 vs InCongruentRatio2	5.22
CongruentSubitizing vs InCongruentSubitizing	0.13
CongruentNon-Subitizing vs InCongruentNon-Subitizing	6.46×10^{7}

Physical Congruency Accuracy Age 11 – 12 years

Models	**BF_{10}**	**Analysis of Effects**	**BF_{incl}**
Null model (incl. subject and random slopes)	1.00	Congruency	0.19
Ratio + Range + Ratio*Range	410960.22	Ratio	505.41
Congruency + Ratio + Range + Congruency*Ratio + Ratio*Range	86464.95	Range	11636.47
Congruency + Ratio + Range + Ratio*Range	61378.07	Congruency*Ratio	0.57
		Congruency*Range	0.25
		Ratio*Range	167.09
		Congruency*Ratio*Range	0.41

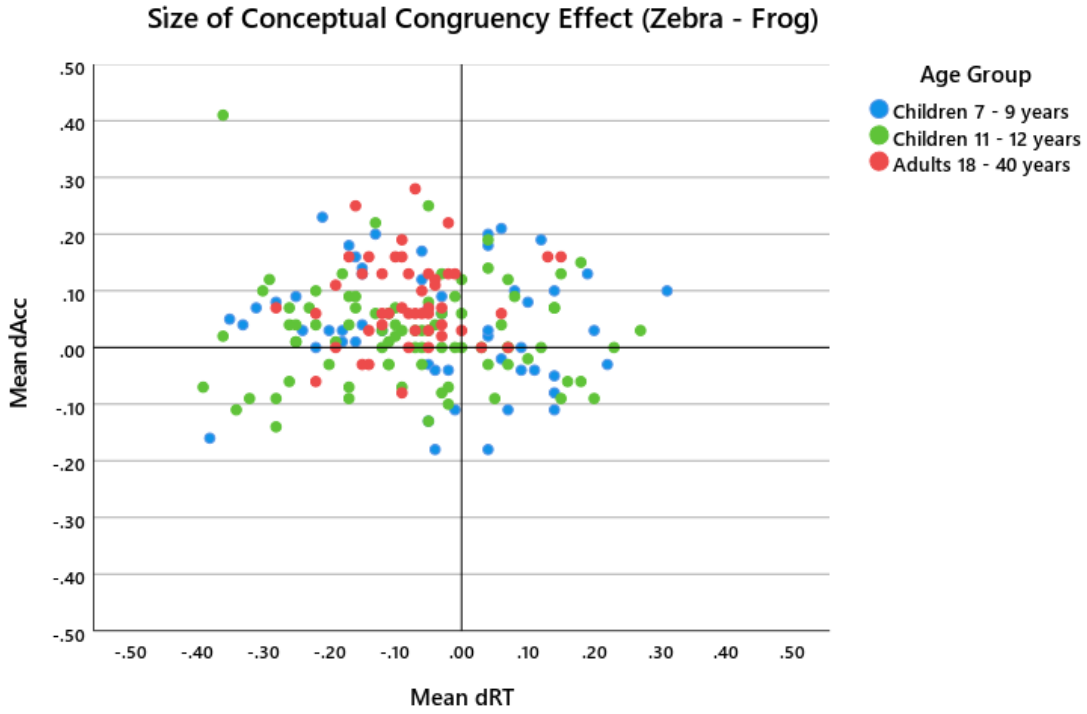
Two-tailed paired samples *t*-test	**BF_{10}**
Ratio1.5Subitizing vs Ratio2Subitizing	0.12
Ratio1.5Non-Subitizing vs Ratio2Non-Subitizing	1482.75

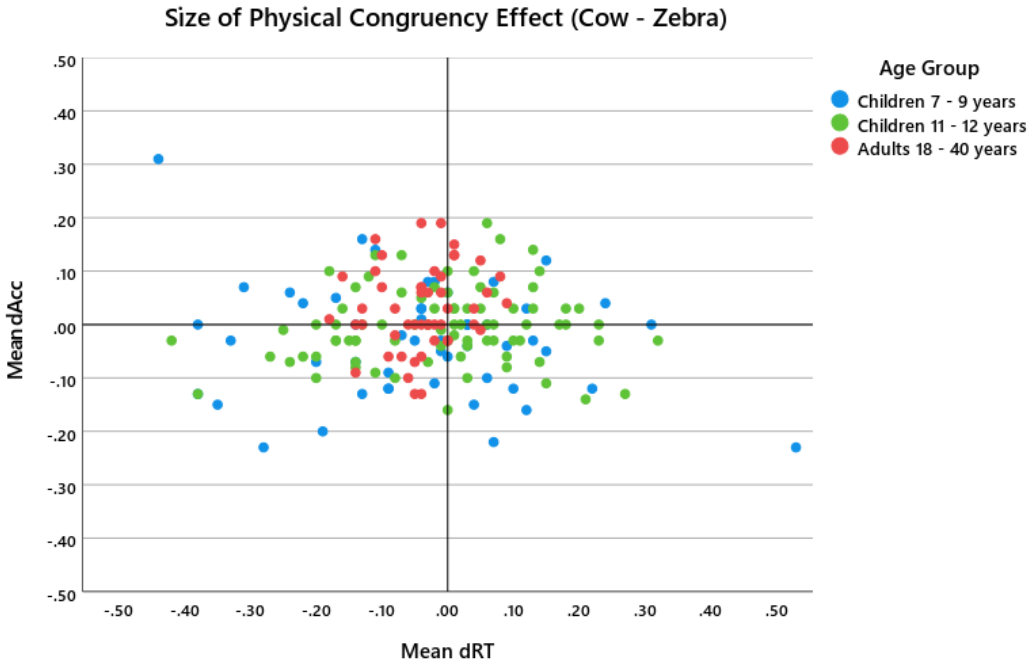
Physical Congruency Reaction Times Age 11 – 12 years

Models	BF_{10}
Null model (incl. subject and random slopes)	1.00
Ratio + Range	3.600×10^{18}
Ratio + Range + Ratio*Range	1.018×10^{18}
Congruency + Ratio + Range + Congruency*Range	5.775×10^{17}

Analysis of Effects	BF_{incl}
Congruency	0.15
Ratio	1.904×10^{11}
Range	6.039×10^{6}
Congruency*Ratio	0.27
Congruency*Range	0.42
Ratio*Range	0.53
Congruency*Ratio*Range	0.03

Figure 2S. The mean size of conceptual (upper panel) and physical (lower panel) congruency effects computed as the difference between congruent and incongruent trials and plotted as a correlation between the accuracy(dAcc) and reaction times (dRT) across the three age groups and per participant.





Note. For the interpretation, if most of the dots are shifted in the upper left quadrant, this indicates the classical congruency effect (i.e., better performance in congruent than incongruent trials