Journal of Experimental Psychology: Human Perception and Performance

Recent Is More: A Negative Time-Order Effect in Nonsymbolic Numerical Judgment

Ronald van den Berg, Marcus Lindskog, Leo Poom, and Anders Winman

Online First Publication, March 6, 2017. http://dx.doi.org/10.1037/xhp0000387

Journal of Experimental Psychology:
Human Perception and Performance

© 2017 American Psychological Association
0096-1523/17/$12.00 http://dx.doi.org/10.1037/xhp0000387

Recent Is More: A Negative Time-Order Effect in Nonsymbolic Numerical Judgment

Ronald van den Berg, Marcus Lindskog, Leo Poom, and Anders Winman
University of Uppsala

Humans as well as some nonhuman animals can estimate object numerosities—such as the number of sheep in a flock—without explicit counting. Here, we report on a negative time-order effect (TOE) in this type of judgment: When nonsymbolic numerical stimuli are presented sequentially, the second stimulus is overestimated compared to the first. We examined this "recent is more" effect in two comparative judgment tasks: larger–smaller discrimination and same–different discrimination. Ideal-observer modeling revealed evidence for a TOE in 88.2% of the individual data sets. Despite large individual differences in effect size, there was strong consistency in effect direction: 87.3% of the identified TOEs were negative. The average effect size was largely independent of task but did strongly depend on both stimulus magnitude and interstimulus interval. Finally, we used an estimation task to obtain insight into the origin of the effect. We found that subjects tend to overestimate both stimuli but the second one more strongly than the first one. Overall, our findings are highly consistent with findings from studies on TOEs in nonnumerical judgments, which suggests a common underlying mechanism.

Public Significance Statement

This study shows that numerical judgments are influenced by a time-order bias: When viewing two sets of objects after one another, the number of objects in the second set is systematically overestimated. We characterize the empirical properties of this effect in a range of numerical judgment tasks and discuss the implications of our findings for studies on numerical cognition.

Keywords: numerical cognition, time-order effects, sensation weighting, approximate number system, ideal-observer models

Supplemental materials: http://dx.doi.org/10.1037/xhp0000387.supp

Imagine walking in the countryside. In a large field you spot two flocks of sheep, one with only white sheep and one with only black, and you make a snapshot judgment of whether there are more white than black sheep. Later, you encounter another two flocks of sheep. This time they emerge from a tunnel, one flock after the other, separated by a brief time interval. Once again, you test your judgment skills by deciding which of the two flocks is more numerous. Will the temporal order in which the two flocks were encountered have influenced your impression of their number? Studies of numerical judgment have thus far implicitly assumed that this is not the case. Here, we question this assumption and find that it is incorrect: When humans make numerical judgments of two sequentially presented stimuli, the latter is generally overestimated relative to the former, a phenomenon that we term the "recent is more" effect.

Estimating the number of sheep in a flock without explicitly counting them is an instance of nonsymbolic numerosity estimation. It has been suggested that humans as well as nonhuman animals are equipped with a dedicated *approximate number system* to support nonsymbolic numerical judgments (Dehaene, 1997; Feigenson, Dehaene, & Spelke, 2004), but other studies have challenged this claim and have suggested instead that numerical estimates are partially or even fully derived from higher level visual cues (Gebuis & Reynvoet, 2012a, 2012b, 2012c; Tokita & Ishiguchi, 2013). Regardless of the underlying mechanism, it is quite generally agreed upon that humans can make fast estimates of nonsymbolic numerosities without the use of explicit counting. The precision of such estimates has been found to obey Weber's law, meaning that they become increasingly imprecise as numerosity increases (Dehaene, 2003; Dehaene, Dehaene-Lambertz, & Cohen, 1998; Mechner, 1958; Whalen, Gallistel, & Gelman, 1999). Moreover, studies with children have suggested that the precision of a person's numerical judgments—typically quantified

Ronald van den Berg, Marcus Lindskog, Leo Poom, and Anders Winman, Department of Psychology, University of Uppsala.

Marcus Lindskog and Anders Winman designed the experiments and collected the data. Ronald van den Berg analyzed the data. All authors contributed to writing the article and approved the final version for submission.

This research was sponsored by the Swedish Research Council (Vetenskapsrådet). We thank Åke Hellström and Marc Buehner for their helpful comments.

Correspondence concerning this article should be addressed to Ronald van den Berg, Department of Psychology, Uppsala University, Von Kraemers allé 1, SE-752 37 Uppsala, Sweden. E-mail: ronald.vandenberg@psyk.uu.se

as a Weber fraction—is predictive of mathematical ability (Halberda, Mazzocco, & Feigenson, 2008; Inglis, Attridge, Batchelor, & Gilmore, 2011), but findings from studies on adults have been mixed (Gebuis & van der Smagt, 2011; Inglis et al., 2011; Price, Palmer, Battista, & Ansari, 2012).

Two types of task are commonly used to measure the precision of a subject's numerical judgments. In *estimation tasks,* subjects directly report the numerosity of a stimulus (e.g., a collection of dots). Such estimates are often biased: In some contexts, subjects may systematically underestimate numerosity, whereas in others they may overestimate it (Crollen, Castronovo, & Seron, 2011). In *binary-decision tasks,* subjects are presented with two stimulus arrays—either simultaneously or sequentially—and make a comparative judgment, such as "Which array contained more dots?" or "Did both arrays contain the same number of dots?" Studies on numerical judgment have rarely, if ever, addressed possible effects of bias in these tasks: It is commonly assumed either that if a bias exists, it is the same in both stimulus arrays and cancels out at the decision stage or is canceled out by counterbalancing the stimuli. Either way, any unaddressed bias impoverishes an experimenter's estimate of the precision of a subject's numerical estimates, because errors due to bias would be interpreted as errors due to imprecise numerical estimates.

The assumption that numerical comparisons are free of bias is questionable when numerosities are judged sequentially, because it is well documented that the temporal order in which stimuli are presented often does induce a bias. Fechner was probably the first to report on this, when he found that the probability of a correct weight-comparison judgment depended on whether an incremented weight was lifted before or after the standard weight (Fechner, 1860). Since then, *time-order effects* (TOEs) have been found in a broad range of judgments, such as successive comparison of temporal intervals (Allan, 1977; Eisler, Eisler, & Hellström, 2008; Jamieson & Petrusic, 1975), line lengths (Tresselt, 1944), tone loudness (Postman, 1946), auditory pitch (Tresselt, 1948), and visual contrast (Alcalá-Quintana & García-Pérez, 2011). A frequent finding in such studies has been that the second stimulus is estimated with a relatively larger magnitude (i.e., judged as longer, louder, or brighter than the first stimulus), but effects of opposite direction have also been found. The sign and magnitude of the TOE typically change with the stimulus magnitude, and these changes are further modulated by the length of the interstimulus interval (ISI; Hellström, 1985).

Surprisingly little is known about TOEs in judgments of nonsymbolic numerical stimuli. Several studies have reported indications of TOEs in pigeons (e.g., Fetterman & MacEwen, 1989; Santi, Lellwitz, & Gagne, 2006), and one study with humans has reported that a group of extensively trained subjects discriminated 15 from 16 visual objects more accurately when the larger set was displayed last (Becker, 1957). However, a detailed study of TOEs in human numerical judgment is currently lacking.

Here, we report on a series of experiments aimed at characterizing both the prevalence and properties of TOEs in human numerical judgment. We quantify the effect in three commonly used numerical judgment tasks, examine how it interacts with stimulus magnitude and interstimulus interval, report on individual differences, and compare our results with those of previous reports of TOEs in other kinds of judgment. Our findings demonstrate that TOEs are highly prevalent and share similarities with TOEs in nonnumerical judgments, which suggests a common underlying mechanism.

Study 1: Time-Order Effects in Two Numerical Judgment Tasks

Method

Larger–smaller judgment task. On each trial, arrays of yellow and blue dots were presented sequentially and the subject reported which of the two arrays contained more dots (see Figure 1a, left side). Each array was viewed for 200 ms, and the two arrays were separated by a 300-ms interstimulus interval. Subjects were tested on numerosity ratios 1:2, 3:4, 5:6, 7:8, and 9:10. Constraining the total number of dots in a pair to the range [10, 30], the 12 unique pairs of numerosities were (5, 6), (5, 10), (6, 8), (6, 12), (7, 8), (7, 14), (8, 16), (9, 10), (9, 12), (10, 12), (12, 16), and (14, 16). Each subject finished 40 trials at each ratio, giving a total of 200 trials per subject. At each trial, a random numerosity pair consistent with the ratio chosen for that trial was chosen from the available pairs. Both the color order (blue first vs. yellow first) and numerosity order (larger first vs. smaller first) were counterbalanced and randomized from trial to trial. The dots varied randomly in size, with radii ranging from .25 to .50 degrees of visual angle. They were randomly placed inside a central square area of 17 × 17 degrees of visual angle, with the constraint that no two dots should overlap. To reduce the potential use of perceptual cues, we matched dot arrays for total area on half of the trials and for average dot size on the other half of the trials (Halberda et al., 2008). Subjects reported which dot array—blue or yellow—was more numerous by pressing a color-coded keyboard button. They did not receive feedback about their performance.

Same–different judgment task. This task was identical to the larger–smaller judgment task except for the following differences. On half of the trials, both arrays had the same number of dots. The task of the subjects was to report whether the number of blue dots was the same as or different from the number of yellow dots. In total, there were 21 unique pairs of numerosities: 12 "different" pairs, which were the same as those used in the larger–smaller judgment task, and nine "same" pairs in which both arrays contained 5, 6, 7, 8, 9, 10, 12, 14, or 16 dots.

Subjects. A total of 30 undergraduate students (10 male) from Uppsala University with a mean age of 26.1 years ($SD = 6.6$) were recruited to participate in this study. Subjects performed both tasks in a single experimental session.[1] The order in which subjects performed the two tasks was counterbalanced. Subjects received a cinema voucher or course credits for their participation. One subject was excluded from the analyses because of unreliable model parameter estimates (see the online supplemental materials for details on the exclusion criterion).

Models. We fitted three ideal-observer models to individual data sets, with two goals in mind: (1) to obtain

[1] The subjects also performed both tasks with parallel stimulus presentations and an arithmetic fluency test. The data of those experiments were collected for purposes that are not relevant to the present study and will be presented elsewhere.

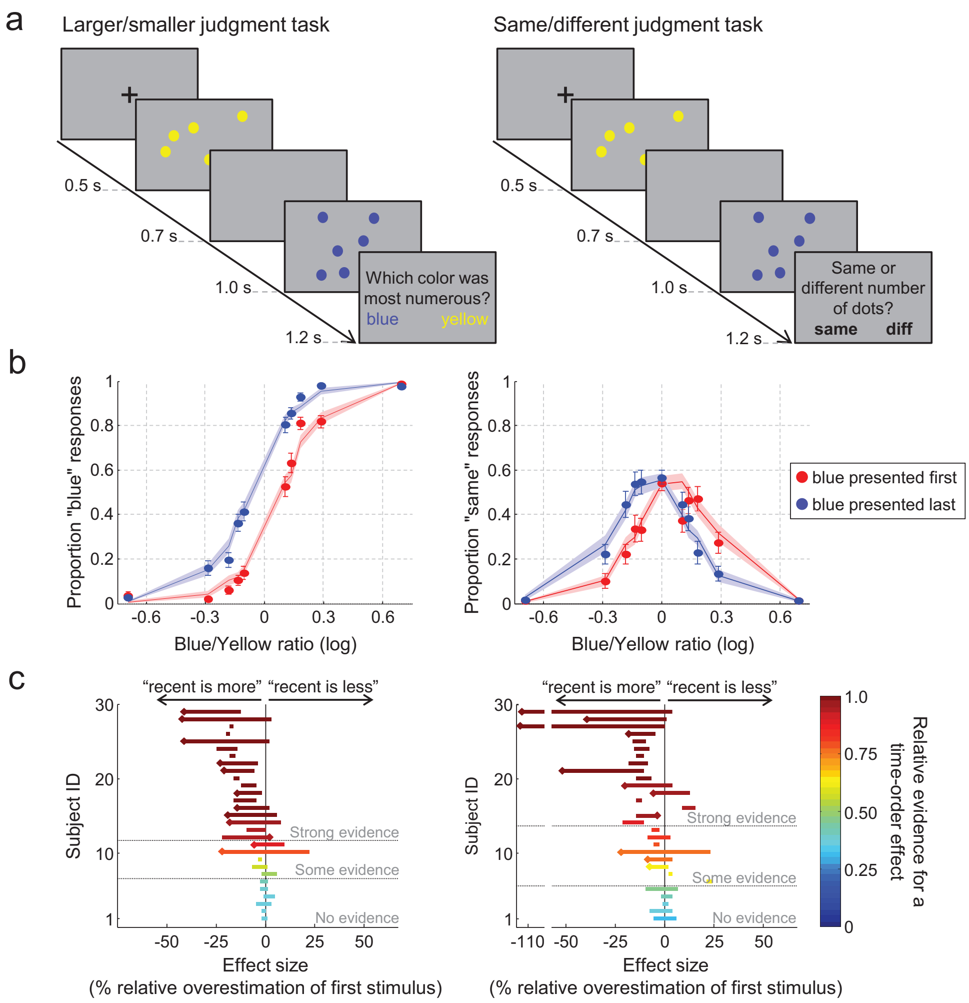

Figure 1. Design and results of Study 1. In all three panels, the left figure refers to the larger–smaller judgment task and the right figure to the same–different judgment task. Panel a: Cartoons of trial sequences in both tasks. diff = different. Panel b: Subject-averaged data (markers) and fits of the ideal-observer model with a magnitude-dependent time-order effect (curves). Error bars and shaded areas represent 1 *SEM* across subjects. The data and fits are split by the temporal order in which the blue and yellow array were presented. The separation between these curves indicates a negative time-order effect: The numerosity of the second presented stimulus was overestimated relative to that of the first. The nonsmoothness of the model predictions is due to unequal distributions of stimulus magnitudes across blue–yellow ratios. Panel c: Individual-subject estimates of time-order effects, based on the best-fitting parameter values of the model with a magnitude-dependent time-order effect. The estimates are sorted by strength of evidence for the existence of a time-order effect (indicated by shades of gray in the bars [or color in the online version of the figure]). Negative effects correspond to "recent is more," and positive ones to "recent is less." Due to the dependence on stimulus magnitude, estimated effect sizes varied across trials. Each bar indicates the range of estimated effects across all of a subject's trials. Hence, one end corresponds to the predicted effect size at trials with the smallest stimulus magnitude (11), and the other to the predicted effect size at trials with the largest stimulus magnitude (30). When a bar is marked with a diamond at one end, it means that the model with a magnitude-dependent time-order effect received more evidence than did the models with a fixed or no time-order effect. The location of the marker indicates direction of the magnitude effect: A marker on the left indicates that the time-order effect decreased with stimulus magnitude, and a marker on the right indicates an increase. Note that in most cases, a decrease in effect size meant an increase in effect *strength* (from negative to more negative) See the online article for the color version of this figure.

subject-level estimates of TOEs and (2) to perform subject-level hypothesis testing by means of formal model comparison. Here, we give a brief description of the most general version of our model—the other two models were constrained variants of it and are introduced later. The most general version of our model was derived from the following four assumptions (see the online supplemental materials for mathematical details):

1. Numerosities are internally represented on a logarithmic scale and corrupted by Gaussian noise (Dehaene, 2003; Nieder & Miller, 2003). The standard deviation of the noise distribution is a free parameter, σ (see Figure S1a in the online supplemental materials).

2. Subjects use a decision strategy based on signal detection theory (Green & Swets, 1966). For the larger–smaller judgment task, this simply means that the observer reported "blue" whenever the observed number of blue dots was larger than the observed number of yellow dots and reported "yellow" otherwise (see Figure S1b in the online supplemental materials). For the same–different task, the decision rule was to report "same" whenever the difference in observed numerosities was within a range $[-c, c]$, where criterion c is a free parameter (see Figure S1c).

3. Observations may be subject to a color-induced bias. This bias is captured by a free parameter δ_{blue}, whose value is added to the subject's estimate of the numerosity of the blue array.

4. Observations may be subject to a bias induced by the presentation order of stimuli (i.e., a TOE), whose sign and magnitude may depend on stimulus magnitude (Hellström, 1985; Michels & Helson, 1954; Needham, 1935; Woodrow, 1935). This bias is captured by a model variable denoted $\delta_{1\text{st}}$, whose value is added to the subject's observation of the numerosity of the first of the two presented stimuli. We relate $\delta_{1\text{st}}$ linearly to stimulus magnitude, which we define as the average numerosity in both arrays (in log units), such that

$$\delta_{1\text{st}} = \alpha + \beta \cdot \left[\frac{\log(N_{\text{blue}}) + \log(N_{\text{yellow}})}{2}\right],$$

where α and β are free parameters.

The parameters of greatest interest are α and β, because they characterize the magnitude of the observer's TOE, $\delta_{1\text{st}}$, and its dependence on stimulus magnitude. When $\delta_{1\text{st}}$ has a positive value, the observer overestimates the numerosity of the first array relative to that of the second, which by convention corresponds to a *positive* TOE (Fechner, 1860).

Note that in this model, the effect of adding $\delta_{1\text{st}}$ to the observed numerosity of the first array is equivalent to subtracting the same value from the observed numerosity of the second array. Therefore, we could estimate only *relative* biases with the current experiment. For example, a positive estimate of $\delta_{1\text{st}}$ would be consistent with an overestimation of the numerosity of the first array, an underestimation of that of the second array, or a combination of both.

Parameter fitting. The model for the larger–smaller task has four free parameters: σ, δ_{blue}, α, and β. The model for the same–different task has one additional parameter, namely decision criterion c. We used Matlab's *fminsearch* function to estimate the maximum-likelihood values of these parameters (MATLAB Optimization Toolbox, 2015).

Estimation of time-order effect sizes. Due to the dependence on stimulus magnitude, the predicted effect size in our model varies across trials. To get a measure of a subject's effect size, we computed the predicted values of $\delta_{1\text{st}}$ (see Assumption 4 earlier in the text) across all of the subject's trials, using her best-fitting estimates of parameters α and β. For presentation purposes, we converted the $\delta_{1\text{st}}$ values to percentage overestimation.[2] We report both the range and mean of these values.

Group-level hypothesis testing. We used Bayesian *t* tests to quantify the amount of evidence for group-level hypotheses (JASP Team, 2016; Morey & Rouder, 2015; Rouder, Morey, Speckman, & Province, 2012), which is a principled alternative to the frequentist *t* test (Rouder, Speckman, Sun, Morey, & Iverson, 2009). The outcome of such a test is a Bayes factor, that is, the ratio between the evidence in favor of the null hypothesis and the evidence in favor of the alternative hypothesis. Bayes factors are more intuitive than are p values and have several other advantages. First, they do not suffer from the fallacy that "absence of evidence is not evidence of absence": Unlike a large *p* value, a large Bayes factor can perfectly be interpreted as strong evidence in favor of the null hypothesis. Second, and related to this, whereas *p* values randomly vary between 0 and 1 if the null hypothesis is true, the expected Bayes factor increases monotonically as one gathers more data.

Subject-level hypothesis testing. We tested two hypotheses at the level of single subjects, the second one being a strong version of the first one: (1) a TOE is present and (2) a TOE is present with a strength that depends on stimulus magnitude. To measure the evidence for these hypotheses, we fitted three models to each data set: (1) the "full" model (described earlier), which has a TOE that depends on stimulus magnitude; (2) a reduced variant, in which the TOE is constant (i.e., $\beta = 0$); and (3) a yet further reduced variant without a TOE (i.e., $\alpha = 0$ and $\beta = 0$). We refer to these models as M_2, M_1, and M_0, respectively. For all three models, we computed the Akaike information criterion (AIC), which is a likelihood-based measure of goodness of fit that takes into account differences in numbers of free parameter (Akaike, 1974). For ease of interpretation, we converted the AIC values to Akaike weights (Burnham, Anderson, & Huyvaert, 2011; Wagenmakers & Farrell, 2004). These weights sum to 1 and represent the relative evidence for a model.

Considerations about statistical power. Statistical power—that is, the probability of rejecting a null hypothesis given that a specific alternative hypothesis is true—is a concept that is useful in the context of null-hypothesis significance testing (NHST). However, as described earlier, we avoided any form of NHST by using Bayesian and model-based approaches instead. In a sense, Bayes factors and Akaike weights are themselves measures of power: The more a Bayes factor deviates from 1 and the closer an Akaike weight is to 0 or 1, the less likely it is that one erroneously accepts a false hypothesis or rejects a true one. Moreover, we performed most of our analyses at the level of single subjects, which means that every subject can be considered as a study replication. For these reasons, a power analysis was not possible or necessary.

[2] For example, a $\delta_{1\text{st}}$ value of .15 would translate to 16.2% relative overestimation. To see this, recall that numerosities are internally represented in logarithmic units. Hence, if the internal representation of the numerosity of the first array equals $n_{1\text{st}}$ and there is no bias, then the subject's numerosity estimate is $N_{\text{unbiased}} = \exp(n_1)$. In the example of $\delta_{1\text{st}} = .15$, the numerosity estimate is $N_{\text{biased}} = \exp(n_{1\text{st}} + .15) = \exp(n_{1\text{st}}) \times \exp(.15) \approx N_{\text{unbiased}} \times 1.162 = N_{\text{unbiased}} + 16.2\%$.

Data and code sharing. All data and analysis code are publicly available at https://uppsalacognitionlab.github.io/RecentIsMore

Results

Group-level analysis of time-order effects on task accuracy. As a first test of whether a TOE exists in our data, we computed for both tasks the percentage of correct responses separately for the trials in which the more numerous array was presented first and for the trials in which it was presented last.[3] The prediction is that if there is no TOE, there should obviously be no reliable difference in accuracy on both types of trial. Alternatively, in the case of a negative TOE (i.e., an underestimation of the numerosity of the first stimulus array relative to that of the second), observed numerosity differences will have been magnified on trials in which the more numerous array was presented last and reduced on trials in which it was presented first, predicting increased performance in the former subset of trials and decreased performance in the latter. Opposite effects on accuracy are predicted in the case of a positive TOE (i.e., an overestimation of the numerosity of the first stimulus array relative to that of the second).

In the larger–smaller judgment task, subjects were correct on 92.0% ± 1.2% of the trials in which the more numerous array was presented last, versus 75.8% ± 2.1% of the trials in which it was presented first (throughout the article, X ± Y stands for mean plus or minus the standard error of the mean across subjects). We used a Bayesian paired samples *t* test to quantify the statistical evidence for the hypothesis that presentation order has an effect on accuracy. We found a Bayes factor (BF) of 12.6×10^3 in favor of the hypothesis that accuracy was higher when the more numerous array was presented last (relative to the hypothesis that accuracy was lower or equal when the more numerous array was presented last). Similarly, accuracy in the same–different task was 77.0% ± 2.4% when the more numerous array was presented last, versus 64.2% ± 3.0% when it was presented first. Also here, a Bayesian *t* test strongly favored the hypothesis that there is a negative TOE ($\text{BF} = 77.9$). These results suggest that the numerosity of the second array was on average overestimated relative to that of the first, a negative TOE that we refer to as the recent-is-more effect. In the psychometric curves, the effect is clearly visible as a time-order-induced shift (see Figure 1b). The trends in Figure 1b also show that the effect cannot be explained as a simple response bias: Although the shifts in the larger–smaller task could potentially be explained as such (e.g., "When uncertain, report the color of the last presented array"), no response-bias explanation can account for the order-induced separation of curves in the same–different task.

Characterization of time-order effects at the level of individual subjects. We next characterized the identified effects in more detail, by seeking answers to the following four questions: (1) What proportion of subjects exhibits evidence for a TOE? (2) What proportion of identified TOEs depend on stimulus magnitude? (3) What is the dominant sign of the TOEs: Do all individual effects reflect the negative group-level effect ("recent is more"), or are there also subjects with a positive TOE? and (4) Are the effect sizes task dependent? We approached these questions by fitting three models to the data (see the Method section for details): one without a TOE (M_0), one with a TOE that did not depend on stimulus magnitude (M_1), and one with a TOE that depended (linearly) on stimulus magnitude (M_2). For each subject, we quantified the amount of evidence for each model in terms of Akaike weights, which sum to 1 and represent the relative evidence for a model. For example, if model M_0 has an Akaike weight of .05, M_1 a weight of .70, and M_2 a weight of .25, then there is strong evidence for a TOE but not for a dependency of effect size on stimulus magnitude (see the Method section for details).

Model fits. Overall, the most flexible model, M_2, accounted well for the data (see Figure 1b), which justifies using the model as a tool to identify and characterize TOEs in these data. Note that the nonsmoothness of the model fits is caused by the fact that the model predictions depend on both the numerosity ratio and the stimulus magnitude; the curves would be smooth if the stimulus magnitude were held constant across ratios or if predictions did not depend on stimulus magnitude (as in models M_0 and M_1).

What proportion of subjects exhibits evidence for a time-order effect? Both models M_1 and M_2 incorporate a TOE. The evidence in favor of a hypothesis that is represented by two models is obtained by summing the Akaike weights for those models (Wagenmakers & Farrell, 2004).[4] Therefore, we computed the evidence in favor of the hypothesis that a TOE was present in a subject's data set by summing the Akaike weights of models M_1 and M_2. If this summed weight exceeded the weight of the model without a TOE (i.e., if it exceeded .50), then the evidence was in favor of the hypothesis; moreover, if it exceeded .90, then we considered the evidence to be strong (Burnham et al., 2011). Using this approach, we found evidence of a TOE for 23 subjects in the larger–smaller judgment task and for 24 subjects in the same–different judgment task (out of a total of 29 subjects); 18 and 16 of these cases, respectively, constituted strong evidence (see Figure 1c). Hence, we found, combined across tasks, evidence of a TOE in 81.0% of the individual data sets (47 out of 58 cases).

What proportion of the identified time-order effects depend on stimulus magnitude? Next, we examined what proportion of the identified effects exhibited evidence for a dependence on stimulus magnitude, again by studying the Akaike weights: If the weight of model M_2 was larger than both the Akaike weight of M_0 and that of M_1, then the evidence was in favor of a dependence; again, when this weight exceeded .90, we considered the evidence to be strong. We found evidence of a dependence for 12 subjects in the larger–smaller judgment task (out of 23 who showed evidence for a TOE; see earlier) and for 11 (out of 24) in the same–different judgment task; four and six of these cases, respectively, constituted strong evidence. Hence, for nearly half of the identified TOEs (23 out of 47), there was evidence for a dependence on stimulus magnitude. The direction of the dependency was strongly consistent: In 21 of these 23 cases, the TOE became more negative or less positive with stimulus magnitude (indicated in Figure 1c with diamond markers located on the left end of a bar).

What is the dominant sign of the time-order effects? Although several subjects showed both negative and positive TOEs—depending on the stimulus magnitude—the large majority of effects were negative: In 85.1% (40 of the 47) of the data sets

[3] In the same–different task, this analysis could obviously be done on only the "different" trials.

[4] If this seems unfair, one should realize that when a hypothesis is shared by multiple models, the total evidence for the hypothesis is not artificially increased but is instead spread out over these models.

that contained evidence for a TOE, the sign of the average effect was negative, that is, a recent-is-more effect (see Figure 1c). Averaged across all subjects, the estimated effect size was 8.5% ± 1.5% relative overestimation of the second stimulus in the larger–smaller judgment task and 8.6% ± 2.9% relative overestimation of the second stimulus in the same–different judgment task.

Is the time-order effect task dependent? Finally, we refitted the model to both data sets simultaneously with a single set of parameters (σ, δ_{blue}, α, β, and c) to test for task dependence of the observed effects. We found that for 15 of the 29 subjects, a comparison based on Akaike weights favored the model in which all parameters were shared across the two tasks (in eight of those cases, the evidence was strong). This suggests that the TOE (as well as the internal noise and color bias) was very similar in both tasks. Moreover, finding that a single set of parameters can account for multiple data sets provides support for the plausibility of a model (Lee, 2011).

Discussion

The results of Study 1 (summarized in Table 1) provide clear evidence for the presence of a TOE in two different numerical judgment tasks: When comparing two sequentially presented numerical stimuli, subjects tended to overestimate the second stimulus compared to the first one. In the terminology of Fechner (1860), this would classify as a negative TOE. Overall, the TOE results of Study 1 are consistent with those observed in other types of judgment (Hellström, 1985): First, we found that the effects were predominantly negative, that is, a relative underestimation of the first stimulus compared to the second; second, the magnitude of the effect depended on the magnitude of the stimulus (the larger the stimulus, the stronger the TOE tended to be); third, there were rather large individual differences in effect sizes. Moreover, we found that for the majority of subjects we were able to explain the data from both tasks using a single set of parameters, which suggests that TOEs in numerical judgments do not strongly depend on the task. This latter finding contrasts that in a previous study that found a difference between TOEs in larger–smaller versus same–different judgments using a sound duration judgment task (Dyjas & Ulrich, 2014).

In both experiments of Study 1, we held the interstimulus interval (ISI) constant at 300 ms. However, studies of TOEs in nonnumerical judgments have reported that ISI is an important factor to consider, because it has been found to interact with the effect that stimulus magnitude has on the TOE (Hellström, 1979, 2003). Both negative and positive TOEs are typically found at every ISI, with the magnitude and sign of the effect changing with stimulus magnitude. It is intriguing that the direction of this change has often been found to depend on the ISI. Next, in Study 2 we examine how ISI affects TOEs in numerical comparisons and whether it interacts with stimulus magnitude.

Study 2: Effect of ISI on Time-Order Effects in Numerical Judgment

Method

Larger–smaller judgment task. The task was the same as in Study 1 (see Figure 1a, left side), except for the following differences. In three separate experimental sessions, each subject was tested with interstimulus intervals (ISIs) of 50 ms, 300 ms, and 2,000 ms. The order of the sessions was randomized across subjects. Presented numerosity ratios were 3:4, 5:6, 7:8, and 9:10 and consisted of the following pairs: (5, 6), (6, 8), (7, 8), (9, 10), (9, 12), (10, 12), (12, 16), and (14, 16). Each pair was repeated 30 times, giving a total of 240 trials per ISI per subject. The dots varied randomly in size, with radii ranging from .50 to .90 degrees of visual angle. They were randomly placed inside a central square area of 13 × 13 degrees.

Subjects. Eighty-five undergraduate students (30 male) from Uppsala University with a mean age of 25.2 years ($SD = 5.8$) were recruited for this study. They received a cinema voucher or course credits for their participation. Three subjects were excluded from the analyses because of unreliable parameter estimates (see the online supplemental materials for details on the exclusion criterion).

Analysis. We used the same statistical methods and models as in Study 1. In addition, we fitted the sensation-weighting model (Hellström, 1979, 2000, 2003), which has been used previously to account for TOEs in a broad range of nonnumerical comparative

Table 1
Prevalence and strength of time-order effects in Studies 1 and 2

Study and task	ISI (ms)	No. data sets with evidence for . . .			Estimated average effect size (% overestimation of 2nd stimulus)	Accuracy difference between largest 2nd and largest 1st trials (%)	Estimated internal Weber fraction (σ)
		Absence of TOE[a]	Presence of TOE[b]	Presence of TOE that depends on stimulus magnitude[c]			
Study 1							
L–S	300	6	23	12	8.5 ± 1.2	16.2 ± 2.8	.134 ± .013
S–D	300	5	24	11	8.6 ± 2.9	12.8 ± 3.4	.112 ± .009
Study 2							
L–S	50	11	71	30	6.3 ± 1.0	13.2 ± 1.9	.168 ± .012
L–S	300	10	72	38	6.65 ± .80	15.2 ± 1.6	.158 ± .009
L–S	2,000	4	78	61	10.3 ± 1.1	20.7 ± 1.7	.165 ± .008
All combined		36	268	152	7.91 ± .55	16.0 ± .9	.156 ± .005

Note. ISI = interstimulus interval; TOE = time-order effect; L–S = larger–smaller; S–D = same–different.
[a] Number of data sets with Akaike weight for model M_0 greater than .50. [b] Number of data sets with Akaike weight for model M_0 smaller than .50. [c] Number of data sets with Akaike weight for models $M_0 < .50$ and $M_2 > M_1$.

judgments. Inspired by the theory offered by Michels and Helson (1954), this model postulates that the inputs to the comparison process are not the stimulus observations themselves but weighted averages between each stimulus observation and a reference level (see Figure S2 in the online supplemental materials for a graphical illustration). Accordingly, the subjective difference between two stimuli takes the form $d = [s_1n_1 + (1–s_1)\psi_R] – [s_2n_2 + (1–s_2)\psi_R]$, where s_1 and s_2 are the "sensation weights," n_1 and n_2 are the noisy stimulus observations, and ψ_R is the reference level (see the online supplemental materials for more details). The model predicts a TOE when $s_1 \neq s_2$. One of the main merits of the model is that it automatically predicts an effect of stimulus magnitude. Moreover, it can describe the interaction effect of ISI and stimulus magnitude on TOEs as relative changes in weights s_1 and s_2 induced by differences in the ISI.

Results

General observations. At all three ISIs, the psychometric curves contained clear evidence for a TOE (see Figure 2a). To estimate individual effect sizes, we fitted the same three models as in Study 1: a model without a TOE (M_0), a model with a constant TOE (M_1), and a model with a TOE that depended (linearly) on stimulus magnitude (M_2). Overall, the most flexible model (M_2) accounted well for the data (see Figure 2a). As in Study 1, we found evidence for a TOE in the large majority of data sets, and the estimated effects were mainly negative (see Figure 2b).

Group-level analysis of the effect of ISI. To assess at the group level whether there was evidence for a main effect of ISI and an interaction with stimulus magnitude, we performed a Bayesian analysis of variance (ANOVA) with δ_{1st} (see the Method section of Study 1) as the dependent variable, ISI and stimulus magnitude as fixed factors, and subject number as a random factor. This statistical test quantifies the evidence for each of five models: H_0, the null model (no main effects and no interaction); H_1, a model with a main effect of only ISI; H_2, a model with a main effect of only stimulus magnitude; H_3, a model with both main effects but no interaction; and H_4, a model with both main effects and an interaction. We found that the model with both main effects and an interaction was strongly favored over the four alternative models, with Bayes factors of 1.21×10^{123} (H_4 vs. H_0), 1.97×10^{103} (H_4 vs. H_1), 1.98×10^{85} (H_4 vs. H_2), and 8.40×10^{62} (H_4 vs. H_3). Hence, at the group level, the data contain overwhelmingly strong evidence for an interaction effect of ISI and stimulus magnitude on TOE size.

Subject-level analyses of the effect of ISI. To obtain insight into the nature of the interaction, we next examined the subject-level model fits in more detail. Using the same method as in Study 1, we found evidence for a TOE for 71, 72, and 78 of the subjects in the ISI = 50, 300, and 2,000 ms conditions, respectively (out of a total of 82 subjects); in 51, 52, and 70 of these cases, respectively, the evidence was strong. The nature of the interaction effect becomes clear when looking at the estimated relation between stimulus magnitude and TOE for these subjects (see Figure 2c): At the shortest ISI, the TOE slightly increases on average with stimulus magnitude ($\beta = .016 \pm .018$), but this reverses at the intermediate ISI ($\beta = -.058 \pm .016$) and more strongly so at the longest ISI ($\beta = -.19 \pm .012$). Note, however, that at all three ISIs the TOE is predominantly negative (relative overestimation of the second stimulus), such that a decrease means that the effect became more negative, that is, stronger. Note also that there is considerable individual variability at each ISI. We found that for 39 out of 82 subjects, a model with all parameters shared between the three ISIs provides a better fit, in terms of Akaike weights, than does fitting them separately. Hence, for almost half of the subjects, there is no evidence for an interaction effect, meaning that the interaction is driven by a subset of the subjects.

Fits of the sensation-weighting model. The interaction effect reported earlier shares similarities with previously reported interaction effects in comparative judgments of nonnumerical stimuli (Hellström, 1985). The currently most comprehensive model to account for such effects is the sensation-weighting model (Hellström, 1979, 2000, 2003). The essence of this model is that the inputs to the comparison process are not the stimulus observations themselves but weighted averages of each stimulus observation and a reference level (see the Method section and the online supplemental materials for details). When the sensation weights are not equal to each other, the model predicts a TOE that depends on stimulus magnitude. Moreover, it describes the interaction effect between ISI and stimulus magnitude as a reversal of the relative size of the sensation weights.

We fitted a slightly modified version (see the online supplemental materials) of the general formulation of the sensation-weighting model (Hellström, 1979, 1985, 2000; Patching, Englund, & Hellström, 2012) and found that it fitted the data equally well[5] as model M_2. Hence, the main difference between the models does not lie in their quantitative predictions (which are apparently near-identical for this experiment) but is purely conceptual: M_2 captures the TOE as a simple bias and remains agnostic about the origin of this bias, whereas the sensation-weighting model accounts for the TOE through an interplay between sensation weights and an adaptation level. We found that, consistent with results of earlier work that used the sensation-weighting model, this model describes the interaction effect of ISI and stimulus magnitude on the TOE in Study 2 as a change in the relative values of the sensation weights: the longer the ISI, the larger the weight of the second stimulus compared to that of the first (see Figure 3).

Discussion

The results of Study 2 are summarized in Table 1. We draw four conclusions from these results. First, they are broadly consistent with those from Study 1: We found strong evidence for TOEs, most of these were negative (i.e., recent is more), and there was considerable variability in individual effect sizes. Second, they reveal an interaction between ISI and stimulus magnitude on TOE size: At short ISI, TOEs became on average weaker (less negative) as a function of stimulus magnitude, whereas at higher ISIs they tended to become increasingly stronger (more negative) with larger stimulus magnitudes. Third, this interaction effect is like effects found in other types of comparative judgment (Hellström,

[5] Across all 246 fits (82 subjects × 3 ISI conditions), the average maximum-likelihood difference was .028 ± .007 and the maximum absolute difference was .77. This suggests that the models make identical quantitative predictions for this task.

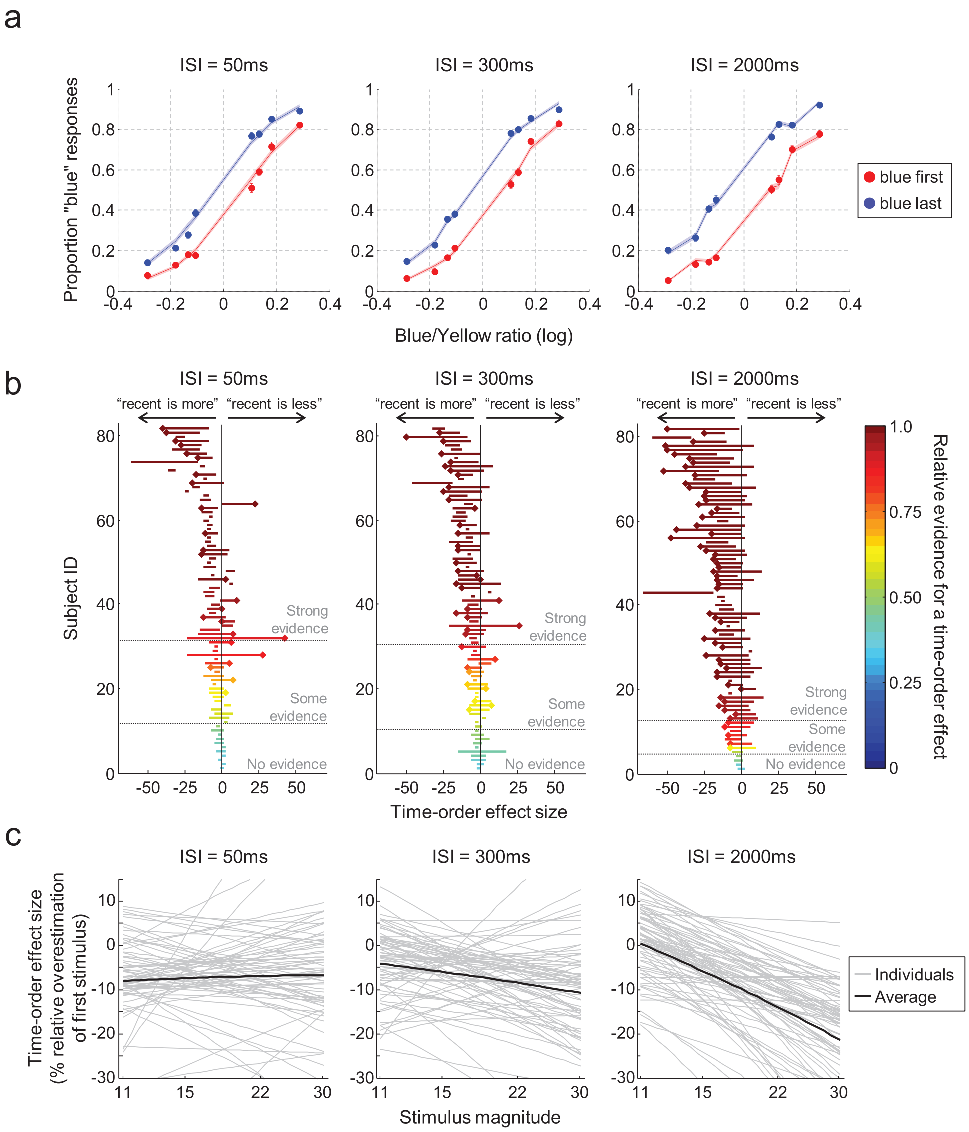

Figure 2. Results of Study 2. Panel a: Subject-averaged data (markers) and fits of the ideal-observer model with a magnitude-dependent time-order effect (curves). ISI = interstimulus interval. Panel b: Individual-subject estimates of time-order effect sizes, based on the maximum-likelihood parameter values of the ideal-observer model with a magnitude-dependent time-order effect. When a bar is marked with a diamond at one end, it means that the model with a magnitude-dependent time-order effect received more evidence than did the models with a fixed or no time-order effect. The location of the marker indicates direction of the magnitude effect: A marker on the left indicates that the time-order effect decreased with stimulus magnitude, and a marker on the right indicates an increase. Note that in most cases, a decrease meant that the effect became more negative, causing a stronger time-order effect. Panel c: Effects of stimulus magnitude and ISI on the time-order effect. Despite considerable variability in individual estimates (thin gray lines), the time-order effect is on average (thick black lines) predominantly negative at every ISI (i.e., subjects tend to overestimate the second stimulus relative to the first one). How the time-order effect changes with stimulus magnitude depends on ISI: At the shortest ISI, the time-order effect on average increases with stimulus magnitude, but this reverts into a decrease at the two longer ISIs. Note that although the model assumes a linear relation between stimulus magnitude and time-order effect, the displayed relations are slightly curved. This is due to the transformation from $\delta_{1\text{st}}$ to percentage overestimation, which is nonlinear (see the Method section of Study 1). See the online article for the color version of this figure.

1985). Fourth, the sensation-weighting model (Hellström, 1979, 1985, 2000) accounts well for these data. When one combines these observations with the similarities that already pointed out in the discussion of Study 1, it starts to appear that the TOE in numerical judgment is in many ways comparable with TOEs found in other types of comparative judgment and may thus share the same underlying mechanism. We come back to this point in the General Discussion section.

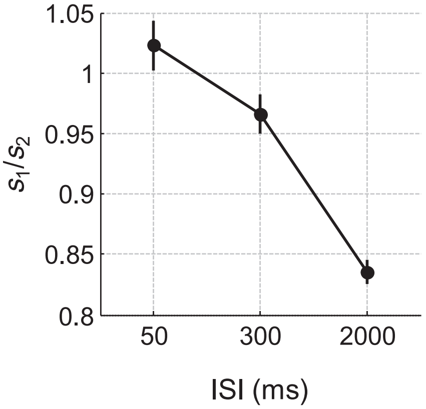

Figure 3. Estimated sensation-weight ratios in the sensation-weight model fitted to the data of Study 2.

Studies 1 and 2 provided a characterization of the prevalence and main properties of a TOE in comparative numerical judgments. However, due to the design of the experiments, these studies provided little information about the nature of the effect: It could be caused by a misjudgment of the first stimulus, a misjudgment of the second stimulus, or a combination of both. Alternatively, it is even possible that both stimuli are perceived without any bias and that the effect is entirely rooted in the comparison process. To obtain insight into the nature of the TOE, in Study 3 we used a numerical estimation task in which subjects provided estimates of both array numerosities without having to compare them.

Study 3: Time-Order Effects in a Sequential Numerical Estimation Task

Method

Stimuli and task. On each trial, subjects were presented with either one or two dot arrays, which we refer to as single-judgment and double-judgment trials, respectively (see Figure 4a). Exposure time of each array was 200 ms, and in double-judgment trials, the arrays were separated by a 300-ms interstimulus interval. Dots were blue in one array and yellow in the other. The color order was always the same for each subject but randomized between subjects. Each array contained 8, 11, or 14 dots, giving nine possible pairs (e.g., 8–8, 8–11). Each pair was presented eight times in random order. Intermixed with those 72 trials were 24 single-judgment

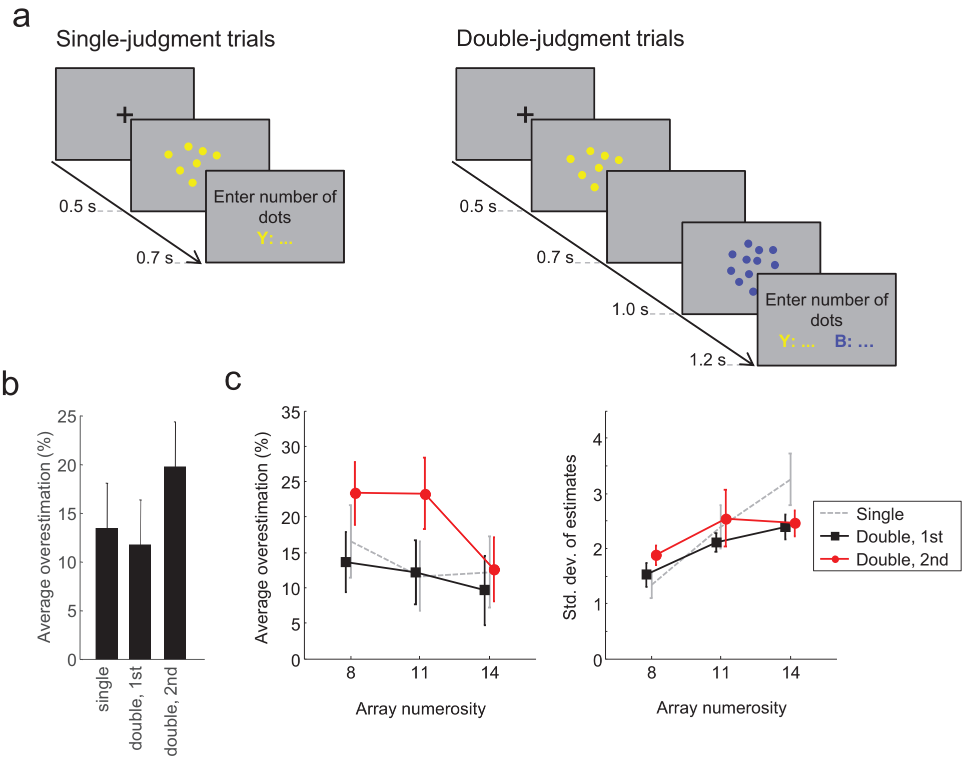

Figure 4. Design and results of Study 3: estimation of sequentially presented numerosities. Panel a: Cartoons of trial procedures in single-judgment (left) and double-judgment (right) trials. Y = yellow; B = blue. Panel b: Comparison of biases. Subjects on average overestimated numerosities in all three judgment types but by different magnitudes. Error bars represent 1 *SEM*. Panel c: Comparison of the average amount of overestimation (left) and variability (right) in responses between the first (black) and second (gray [red in the online version of the figure]) judgment in double-judgment trials, as a function of array numerosity. Error bars represent 1 *SEM*. Std. dev. = standard deviation. See the online article for the color version of this figure.

trials in which each possible numerosity appeared eight times. The displays were controlled for average dot size in half of the double-judgment trials and for cumulative area in the other half (Halberda et al., 2008). Subjects reported the number of dots in the arrays by entering a number with the keyboard. The input box on the screen was color-coded and always occurred in a left–right fashion corresponding to first–second temporal position. When subjects missed a stimulus, they could indicate this by entering an error code instead of an estimate (this happened on only six trials, which were excluded from the analyses).

Subjects. Twenty undergraduate students (12 female), with an average age of 24.8 years ($SD = 5.49$) were recruited for the study. They received a cinema voucher or course credits for their participation.

Results

Numerosities are overestimated in all three types of judgment. Subjects on average overestimated the numerosities by 13.5% ± 4.6% in the single-judgment trials and by 11.8% ± 4.3% and 19.8% ± 4.6% on the first and second judgments, respectively, in double-judgment trials (see Figure 4b). In all three cases, a Bayesian one-sample *t* test supported the hypothesis that the average bias was larger than 0 (BFs = 11.3, 8.23, and 200, respectively). These results reveal a general overestimation bias that is commonly observed in numerical judgments (Crollen et al., 2011), especially in the absence of feedback (Izard & Dehaene, 2008). No effect of color order (blue–yellow vs. yellow–blue) was found in the double-judgment trials: For both the first and the second judgments, a Bayesian *t* test supported the hypothesis that the average amount of overestimation was the same for both orders (BF = 2.47 in both cases).

Comparison of judgment biases in single-judgment and double-judgment trials. Overestimation on single-judgment trials was on average 1.7% ± 1.0% larger and 6.3% ± 1.3% smaller than overestimation of the first and second arrays of double-judgment trials, respectively (see Figure 4b). A Bayesian paired-samples *t* test supported the null hypothesis that there was no difference in overestimation between single-judgment trials and the first judgment in double-judgment trials (BF = 1.28 in favor of the null hypothesis). In contrast, strong evidence was found for a difference between overestimation in single-judgment trials and the second judgment in double-judgment trials (BF = 631 in favor of a difference). Indeed, overestimation of the numerosity of the second array was on average 8.0% ± 1.2% larger than that of the first array. Hence, the TOE seems to be rooted in an amplification of the overestimation of the second stimulus induced by the presence of the first stimulus.

Comparison of estimation precision between first and second judgments in double-judgment trials. These findings suggest that the TOE is rooted in a bias in the judgment of the second stimulus, which would rule out any theory that explains the effect because of an impoverished precision memory of the first stimulus (e.g., due to memory decay). To further test this, we conducted a Bayesian ANOVA with estimation precision as the dependent variable, stimulus magnitude and stimulus position (first or second) as fixed factors, and subject number as a random factor. As explained in Study 2, this test quantifies the evidence for five different models: the null model, two models with a single main effect, a model with both main effects without an interaction, and a model with both main effects and an interaction. We found that the test most strongly supported the model with a main effect of only stimulus magnitude (BF = 58.2 compared to the null model), closely followed by the model with both main effects (BF = 47.8 compared to the null model) and the model with both main effects and an interaction (BF = 9.41 compared to the null model). Hence, there is strong evidence for an effect of stimulus magnitude on estimation precision but weak evidence *against* an effect of temporal stimulus position. Moreover, contrary to what a theory based on memory decay would predict, if there were a difference, then the precision of the first stimulus estimate would seem to be *higher* than that of the second stimulus not lower (see Figure 4c).

Comparison of time-order effects in the estimation task with time-order effects in the comparison tasks. At first sight, the relative overestimation of 8.0% ± 1.2% of the second stimulus compared to the first seems quantitatively consistent with the model-based effect size estimates that we obtained for the comparison tasks in Studies 1 and 2 (8.5% ± 1.5% and 8.6% ± 2.9% in Study 1 and 6.65% ± .80% in Study 2). However, these estimates are not directly comparable, because the stimuli in Studies 1 and 2 had different magnitudes, and we found earlier that for many subjects the effect size depends on stimulus magnitude. To assess more properly whether the empirical effect sizes measured in Study 3 are consistent with the model-based estimates from Studies 1 and 2, we computed for each subject in Studies 1 and 2 the average predicted bias for the trials in Study 3; that is, we computed $\delta_{1st} = \alpha + \beta \times [\log(N_{\text{blue}}) + \log(N_{\text{yellow}})]/2$ using the best fitting estimates of α and β from Studies 1 and 2 but with the N_{blue} and N_{yellow} pairs from Study 3. We found predicted average effect sizes of 10.0% ± 1.7% in the larger–smaller judgment task of Study 1, 10.4% ± 3.4% in the same–different judgment task of Study 1, and 6.0% ± 1.6% in Study 2. In all three cases, a Bayesian independent-samples *t* test supported the null hypothesis that the average predicted effect size was the same as the empirical effect size found in Study 3, with Bayes factors of 2.51, 3.04, and 2.47, respectively. This suggests that the mechanism underlying TOEs in the estimation task is the same as in the binary-decision tasks of Studies 1 and 2.

Discussion

The findings from Study 3 provide three important pieces of information about the nature of TOEs in nonsymbolic numerical judgment. First, they suggest that the TOE is rooted in an amplification of the overestimation of the second stimulus: Although both the first and second stimuli in a sequence are overestimated, the latter is overestimated more strongly than is the former. Second, the results are difficult to reconcile with any theory that explains the TOE because of an impoverished memory of the first stimulus, because there is no evidence for such an impoverishment. Third, the finding that TOEs exist even when the two numerosities do not need to be compared with each other suggests that the effect is not rooted in the comparison process but occurs already in the stimulus observations. This suggestion is further supported by our finding that the estimated effect sizes are quantitatively consistent across the larger–smaller, same–different, and estimation tasks, which had quite different decision processes but used similar stimuli.

General Discussion

Research on time-order effects (TOEs) in human judgments has a long history, dating back to the pioneering work of Gustav Fechner (1860). It is somewhat surprising that TOEs have hardly been studied in the context of nonsymbolic numerical judgments. Here, we examined the prevalence and characteristics of TOEs in three numerical judgment tasks. Although we found substantial individual differences in effect sizes, TOEs were highly prevalent in all three tasks, with a strongly consistent effect direction (see Table 1 for a summary): Regardless of task, ISI, and stimulus magnitude, subjects tended to overestimate the second stimulus relative to the first one. The average effect size was 7.91% ± 0.55% relative overestimation of the second stimulus. Although this may be perceived as a mild effect, we found that it had quite large behavioral consequences: Accuracy on trials in which the largest array was presented last was on average 16.0% ± .9% higher than in the trials in which it was presented first. Moreover, we found an interaction effect of ISI and stimulus magnitude on the TOE: At short ISI, the TOE tended to get weaker (less negative) with larger stimulus magnitudes and in some cases even reversed sign; at long ISI, on the other hand, the TOE tended to become stronger (more negative) for larger stimulus magnitudes.

Implications for Studies on Nonsymbolic Numerical Cognition

Sequential designs have been widely used in studies on nonsymbolic numerical cognition (e.g., Barth, Beckmann, & Spelke, 2008; Barth et al., 2006; Gebuis & van der Smagt, 2011; Gilmore, Attridge, & Inglis, 2011; Lindskog, Winman, & Juslin, 2014; Lyons, Ansari, & Beilock, 2012; Park & Brannon, 2013; Pica, Lemer, Izard, & Dehaene, 2004; Price et al., 2012; Tokita & Ishiguchi, 2016). However, none of these studies took the existence of TOEs into account, which may have led to wrong inferences. For example, these studies have underestimated the precision with which subjects estimated numerosities, because they may have confounded errors due to time-order effects with errors due to precision limitations. Indeed, when we refitted our own data with a model that ignores TOEs, we found that internal Weber fractions—a common measure of numerical judgment precision (Dehaene, 2007; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004)—were systematically overestimated by magnitudes up to over 100% (see Figure S3 in the online supplementary materials). Hence, future studies that aim to measure numerical judgment precision should explicitly model TOEs when using sequentially presented stimuli, and conclusions from some of the previous studies may need reconsideration.

A second implication of our results relates to the measurement of people's numerical judgment precision. There has been much debate about what is the best way to do this, because different measures of precision give different results that are not necessarily correlated (Gebuis & van der Smagt, 2011; Gilmore et al., 2011; Inglis & Gilmore, 2014; Lindskog, Winman, Juslin, & Poom, 2013; Price et al., 2012). We propose that, based on our present findings, part of the inconsistencies may be a result of the problem outlined earlier, namely that measures of numerical judgment precision may in some cases have been misestimated due to not accounting for TOEs. Our own data can again be used to illustrate this point. In Study 1, we found that raw performance measures (proportion correct) correlated rather weakly between the larger–smaller and same–different judgment tasks ($R^2 = .16$, $p = .029$). However, when we used the internal Weber fraction (i.e., the σ parameter in model M_2) as a measure of numerical judgment precision, we found a strong correlation between the tasks ($R^2 = .76$, $p = 2.79 \times 10^{-10}$). Hence, when using proportion-correct scores as a measure of judgment precision, one might conclude that the tasks do not produce equally valid measures of numerical judgment precision—as has been done in a previous study (Gebuis & van der Smagt, 2011)—but when using a measure of precision that is not contaminated by TOEs, one finds that the tasks produce highly consistent results and there is no reason to prefer one task over the other.

Although the existence of TOEs thus mostly seems to complicate matters in nonsymbolic numerical judgment research, it also provides new opportunities to validate and falsify theories about the mechanisms underlying nonsymbolic numerical judgments. For example, it has been argued that there are separate mechanisms for nonsymbolic numerical judgments in low- and high-density contexts (Anobile, Cicchini, & Burr, 2014, 2016). This theory makes the testable prediction of a possible discontinuity in time-order effects between numerical judgments in low- and high-density displays. In addition, some researchers have questioned the existence of a dedicated approximate number system and argued instead that humans may estimate numerosity indirectly by combining different visual cues (Gebuis & Reynvoet, 2012a, 2012b, 2012c; Tokita & Ishiguchi, 2013). This view predicts that the recent-is-more effect does not stand by itself but should reduce to a bias in the perception of one or more of these visual cues (e.g., the second stimulus being perceived as denser or larger than the first). Testing these two predictions would provide an opportunity to further increase understanding of the mechanisms underlying human nonsymbolic numerical judgment.

Mechanisms

Our results show that TOEs in nonsymbolic numerical judgments share many similarities with the extensively documented TOEs in other, nonnumerical judgments: The effects are largely negative (relative underestimation of the first stimulus), the effects are affected by both stimulus magnitude and ISI through an interaction, and effect sizes vary across subjects. This consistency suggests that the TOE in our data has the same underlying mechanism as do TOEs in nonnumerical judgments.

In the context of nonnumerical judgments, it has been argued that TOEs may be explained as a simple response bias (Alcalá-Quintana & García-Pérez, 2011). However, our results are inconsistent with such an explanation, because it cannot explain the time-order induced shift of psychometric curves in the same–different judgment task (see Figure 1b). Another, more comprehensive explanation of TOEs is the proposal of sensation weighting (Hellström, 1979, 1985, 2000), which states that inputs to the comparison process are not the stimulus observations themselves but rather weighted averages between each stimulus observation and a reference level. This idea has been formalized in the sensation-weighting model, which we found to account well for our data (Study 2). However, we also found that the predictions of this model are indistinguishable from those of our ideal-observer model M_2. The difference between the models is thus purely conceptual: Whereas model M_2 accounts for TOEs by simply postulating a magnitude-dependent time-order bias (while staying agnostic about the cause of this bias), the sensation-weighting model constructs the bias via a sensation-weighting mechanism. Future work could try to employ different tasks to test whether there is any quantitative evidence that supports the concept of sensation weighting in nonsymbolic numerical cognition.

Function

What could possibly be the function of time-order effects in nonsymbolic numerical judgment? One appealing proposal—that motivated the development of the sensation-weighting model—is that the TOE may be a byproduct of a mechanism that is meant to increase discriminability between stimuli (Hellström, 1986, 1989). Per this theory, discriminability of stimuli can be improved when noisy stimulus representations are weighted with a reference level, with weights depending on the noise level associated to a stimulus; when the noise levels differ for the two stimuli, then the optimal weights differ as well and cause a TOE. Although this idea is theoretically appealing, two of our findings argue against it. First, Study 3 showed that TOEs exist even when no comparison must be made. Second, and more important, there was no noticeable difference in the precision with which both stimuli are represented (see Figure 4c). In the absence of such a difference, discrimination will get worse, not better, when using unequal sensation weights.

Another possibility could be that the TOE is an artifact of a visual system that is adapted to natural environments rather than the laboratory setting. Although it is not uncommon in natural environments that visual stimuli sometimes briefly disappear (e.g., due to an eyeblink), it seems quite rare that a set of objects is replaced with an entirely new, unrelated set of objects during such an event. Consider, for example, that one were witnessing a pack of hungry wolves.[6] Suddenly an occlusion event occurs that temporarily obstructs the view, for example by a bush, a tree, or an eyeblink. In this situation, it may be safe for the brain to assume that the wolves one sees after the occlusion event are not necessarily the same ones one observed before the occlusion: A subset of the first group could be temporarily out of view. If a subset of the wolves is uncommon to both the preocclusion and postocclusion images of the pack, then it would imply that the pack must be larger than what the images in either interval indicated. Although highly speculative at this point, it could be potentially fruitful if future studies of time-order effects would take considerations of natural statistics into account.

[6] We thank an anonymous reviewer for this example.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

Alcalá-Quintana, R., & García-Pérez, M. (2011). A model for the time-order error in contrast discrimination. *Quarterly Journal of Experimental Psychology, 64,* 1221–1248. http://dx.doi.org/10.1080/17470218.2010.540018

Allan, L. G. (1977). The time-order error in judgments of duration. *Canadian Journal of Psychology, 31,* 24–31. http://dx.doi.org/10.1037/h0081647

Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological Science, 25,* 265–270. http://dx.doi.org/10.1177/0956797613501520

Anobile, G., Cicchini, G. M., & Burr, D. (2016). Number as a primary perceptual attribute: A review. *Perception, 45,* 5–31.

Barth, H., Beckmann, L., & Spelke, E. S. (2008). Nonsymbolic, approximate arithmetic in children: Abstract addition prior to instruction. *Developmental Psychology, 44,* 1466–1477. http://dx.doi.org/10.1037/a0013046

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Non-symbolic arithmetic in adults and young children. *Cognition, 98,* 199–222. http://dx.doi.org/10.1016/j.cognition.2004.09.011

Becker, E. (1957). *Mengenvergleich und Übung* [Quantitative comparison and practice]. Frankfurt am Main, Germany: Verlag Waldemar Kramer.

Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: A bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology, 58,* 39–49. http://dx.doi.org/10.1027/1618-3169/a000064

Dehaene, S. (1997). *The number sense*. Oxford, United Kingdom: Oxford University Press.

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

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Attention and Performance XXII: Sensorimotor foundations of higher cognition* (pp. 527–574). http://dx.doi.org/10.1093/acprof:oso/9780199231447.003.0024

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences, 21,* 355–361. http://dx.doi.org/10.1016/S0166-2236(98)01263-6

Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. *Quarterly Journal of Experimental Psychology, 67,* 1121–1150. http://dx.doi.org/10.1080/17470218.2013.847968

Eisler, H., Eisler, A. D., & Hellström, Å. (2008). Psychophysical issues in the study of time perception. In S. Grondin (Ed.), *Psychology of Time* (pp. 75–110). Bingley, United Kingdom: Emerald Group.

Fechner, G. T. (1860). *Elemente der psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

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

Fetterman, J. G., & MacEwen, D. (1989). Short-term memory for responses: The "choose-small" effect. *Journal of the Experimental Analysis of Behavior, 52,* 311–324. http://dx.doi.org/10.1901/jeab.1989.52-311

Gebuis, T., & Reynvoet, B. (2012a). Continuous visual properties explain neural responses to nonsymbolic number. *Psychophysiology, 49,* 1649–1659. http://dx.doi.org/10.1111/j.1469-8986.2012.01461.x

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

Gebuis, T., & Reynvoet, B. (2012c). The role of visual information in numerosity estimation. *PLoS ONE, 7*(5), e37426. http://dx.doi.org/10.1371/journal.pone.0037426

Gebuis, T., & van der Smagt, M. J. (2011). False approximations of the approximate number system? *PLoS ONE, 6*(10), e25405. http://dx.doi.org/10.1371/journal.pone.0025405

Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *Quarterly Journal of Experimental Psychology, 64,* 2099–2109. http://dx.doi.org/10.1080/17470218.2011.574710

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Society, 1,* 521. http://dx.doi.org/10.1901/jeab.1969.12-475

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008, October 2). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455,* 665–668. http://dx.doi.org/10.1038/nature07246

Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 460–477. http://dx.doi.org/10.1037/0096-1523.5.3.460

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin, 97,* 35–61. http://dx.doi.org/10.1037/0033-2909.97.1.35

Hellström, Å. (1986). Sensation weighting in comparing: A tool for optimizing discrimination. In B. Berglund, U. Berglund, & R. Teghtsoonian (Eds.), *Fechner Day 1986: Proceedings of the Second Annual Meeting of the International Society for Psychophysics* (pp. 89–94). Stockholm, Sweden: International Society for Psychophysics.

Hellström, Å. (1989). What happens when we compare two stimuli? In G. Ljunggren & S. Dornic (Eds.), *Psychophysics in action* (pp. 25–39). http://dx.doi.org/10.1007/978-3-642-74382-5_3

Hellström, A. (2000). Sensation weighting in comparison and discrimination of heaviness. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 6–17. http://dx.doi.org/10.1037/0096-1523.26.1.6

Hellström, A. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics, 65,* 1161–1177. http://dx.doi.org/10.3758/BF03194842

Inglis, M., Attridge, N., Batchelor, S., & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin & Review, 18,* 1222–1229. http://dx.doi.org/10.3758/s13423-011-0154-1

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica, 145,* 147–155. http://dx.doi.org/10.1016/j.actpsy.2013.11.009

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

Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics, 17,* 197–202. http://dx.doi.org/10.3758/BF03203886

JASP Team. (2016). JASP (Version 0.8.0.0) [Computer program].

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology, 65,* 23–35. http://dx.doi.org/10.1007/s00265-010-1029-6

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55,* 1–7. http://dx.doi.org/10.1016/j.jmp.2010.08.013

Lindskog, M., Winman, A., & Juslin, P. (2014). The association between higher education and approximate number system acuity. *Frontiers in Psychology, 5:* 462. http://dx.doi.org/10.3389/fpsyg.2014.00462

Lindskog, M., Winman, A., Juslin, P., & Poom, L. (2013). Measuring acuity of the approximate number system reliably and validly: The evaluation of an adaptive test procedure. *Frontiers in Psychology, 4:* 510. http://dx.doi.org/10.3389/fpsyg.2013.00510

Lyons, I. M., Ansari, D., & Beilock, S. L. (2012). Symbolic estrangement: Evidence against a strong association between numerical symbols and the quantities they represent. *Journal of Experimental Psychology: General, 141,* 635–641. http://dx.doi.org/10.1037/a0027248

MATLAB Optimization Toolbox. (2015). (Version 7.3) [Computer software]. Natick, MA: The Mathworks Inc.

Mechner, F. (1958). Probability relations within response sequences under ratio reinforcement. *Journal of the Experimental Analysis of Behavior, 1,* 109–121. http://dx.doi.org/10.1901/jeab.1958.1-109

Michels, W. C., & Helson, H. (1954). A quantitative theory of time-order effects. *American Journal of Psychology, 67,* 327–334. http://dx.doi.org/10.2307/1418635

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.12-2.* Available from http://cran.r-project.org/web/packages/BayesFactor/index.html

Needham, J. G. (1935). The effect of the time interval upon the time-order at different intensive levels. *Journal of Experimental Psychology, 18,* 530–543. http://dx.doi.org/10.1037/h0056775

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

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science, 24,* 2013–2019. http://dx.doi.org/10.1177/0956797613482944

Patching, G. R., Englund, M. P., & Hellström, A. (2012). Time- and space-order effects in timed discrimination of brightness and size of paired visual stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 915–940. http://dx.doi.org/10.1037/a0027593

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44,* 547–555. http://dx.doi.org/10.1016/j.neuron.2004.10.014

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004, October 15). Exact and approximate arithmetic in an Amazonian indigene group. *Science, 306,* 499–503. http://dx.doi.org/10.1126/science.1102085

Postman, L. (1946). The time-error in auditory perception. *American Journal of Psychology, 59,* 193–219. http://dx.doi.org/10.2307/1416885

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica, 140,* 50–57. http://dx.doi.org/10.1016/j.actpsy.2012.02.008

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Santi, A., Lellwitz, J., & Gagne, S. (2006). Pigeons' memory for sequences of light flashes: Reliance on temporal properties and evidence for delay interval/gap confusion. In *Behavioural processes* (Vol. 72, pp. 128–138). http://dx.doi.org/10.1016/j.beproc.2006.01.006

Tokita, M., & Ishiguchi, A. (2013). Effects of perceptual variables on numerosity comparison in 5–6-year-olds and adults. *Frontiers in Psychology, 4*(July), 431. http://dx.doi.org/10.3389/fpsyg.2013.00431

Tokita, M., & Ishiguchi, A. (2016). Precision and bias in approximate numerical judgment in auditory, tactile, and cross-modal presentation. *Perception, 45,* 56–70.

Tresselt, M. E. (1944). Time errors in successive comparison of simple visual objects. *American Journal of psychology 1,* 57, 555–558. http://dx.doi.org/10.2307/1417249

Tresselt, M. E. (1948). Time-errors in successive comparison of tonal pitch. *American Journal of Psychology, 61,* 335–342. http://dx.doi.org/10.2307/1417153

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. http://dx.doi.org/10.3758/BF03206482

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

Woodrow, H. (1935). The effect of practice upon time-order errors in the comparison of temporal intervals. *Psychological Review, 42,* 127–152. http://dx.doi.org/10.1037/h0063696

Received April 19, 2016
Revision received December 22, 2016
Accepted December 22, 2016 ■