Cognitive Psychology 120 (2020) 101288

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Modeling the interaction of numerosity and perceptual variables with the diffusion model

Inhan Kang, Roger Ratcliff*

The Ohio State University, 291 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, United States

ARTICLE INFO

Keywords:
Diffusion model
Approximate number system
Response time and accuracy
Interaction of numerosity and perceptual features
Conflict effect

ABSTRACT

Ratcliff and McKoon (2018) proposed integrated diffusion models for numerosity judgments in which a numerosity representation provides evidence used to drive the decision process. We extend this modeling framework to examine the interaction of non-numeric perceptual variables with numerosity by assuming that drift rate and non-decision time are functions of those variables. Four experiments were conducted with two different types of stimuli: a single array of intermingled blue and yellow dots in which both numerosity and dot area vary over trials and two side-by-side arrays of dots in which numerosity, dot area, and convex hull vary over trials. The tasks were to decide whether there were more blue or yellow dots (two experiments), more dots on which side, or which dots have a larger total area. Development of models started from the principled models in Ratcliff and McKoon (2018) and became somewhat ad hoc as we attempted to capture unexpected patterns induced by the conflict between numerosity and perceptual variables. In the three tasks involving numerosity judgments, the effects of the non-numeric variables were moderated by the number of dots. Under a high conflict, judgments were dominated by perceptual variables and produced an unexpected shift in the leading edge of the reaction time (RT) distributions. Although the resulting models were able to predict most of the accuracy and RT patterns, the models were not able to completely capture this shift in the RT distributions. However, when subjects judged area, numerosity affected perceptual judgments but there was no leading edge effect. Based on the results, it appears that the integrated diffusion models provide an effective framework to study the role of numerical and perceptual variables in numerosity tasks and their context-dependency.

1. Introduction

Human mathematical ability is an important part of our daily lives. We are able to calculate the time so as not to be late for an appointment, able to estimate the distance to our destination, and able to set a monthly budget. Because this ability requires symbolic and abstract manipulation, it is believed to operate in a high-level cognitive domain. However, it has been argued that the cognitive basis of this symbolic mathematical ability developed out of a primitive non-symbolic mental process, which is believed to be present not only in human adults (DeWind & Brannon, 2012; Park & Brannon, 2013, 2014), but also in young children (Gilmore, McCarthy, & Spelke, 2010; Mazzocco, Feigensen, & Halberda, 2011; Libertus, Fiegensen, & Halberda, 2011, 2013; Hyde, Khanum, & Spelke, 2014; Park, Bermudez, Roberts, & Brannon, 2016), and even in animals (Gallistel & Gelman, 1992, 2000; Halberda, Mazzocco, & Feigensen, 2008; Nieder & Miller, 2003). This view is embodied in a theory that proposes that an Approximate Number System (ANS) underlies

* Corresponding author.
E-mail addresses: kang.985@osu.edu (I. Kang), ratcliff.22@osu.edu (R. Ratcliff).

https://doi.org/10.1016/j.cogpsych.2020.101288
Received 24 February 2020; Accepted 25 February 2020

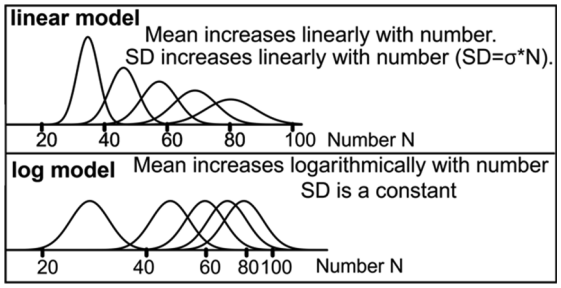

Fig. 1. Log and linear models of numerosity representation.

processing in simple non-symbolic tasks. The ANS assumes that numerosity is represented as a distribution whose central value is the true numerosity (Dehaene, 2003). Furthermore, this cognitive system is assumed to have an innate ability to approximately estimate the number of non-symbolic stimuli, which is usually referred to as 'the number sense' (Dehaene, 1997, 2003; Halberda & Feigensen, 2008; Leibovich, Katzin, Harel, & Henik, 2017).

The claim that an ability to approximate numerosity builds a cognitive foundation for symbolic mathematical proficiency has been supported by findings in numerosity studies (Feigensen, Dehaene, & Spelke, 2004; Piazza, 2010). Many studies have shown that ANS acuity, a measure of the precision of the number approximation, correlates with achievements in standardized mathematical tests (Halberda et al., 2008; Libertus et al., 2011; DeWind & Brannon, 2012). This relationship has been found at different developmental stages, from adolescent to adult years (Halberda, Ly, Wilmer, Naiman, & Germine, 2012). Furthermore, it has been shown that measures of early ANS precision are good predictors of future mathematical proficiency (Gilmore et al., 2010; Mazzocco, Feigensen, & Halberda, 2011; Libertus et al., 2013). The retrospective correlation between ANS performance and standardized test scores and prospective prediction of mathematical ability based on early ANS precision have been found significant in a recent meta-analysis (Chen & Li, 2014). Furthermore, training on tasks that engage primitive approximation of numerical quantities, such as nonsymbolic numerosity discrimination tasks, improved performance in subsequent symbolic math tasks (Hyde et al., 2014). However, this improvement did not appear in other types of numerical manipulation, such as ordering, matching, and comparison, nor in other types of tasks, such as a visuospatial short-term memory task and general knowledge training (Park & Brannon, 2013, 2014; Hyde et al., 2014; Park et al., 2016).

There has been a debate about how to represent numerosity information in the ANS. Currently, there are two competing models that represent numerosity information as a distribution over numerosity. These are a log model and a linear model (Fig. 1). The difference between these two models is the assumption about scale and variability. The log model assumes that the numerosity is represented on a logarithmic scale and variability is equal around all log numerosities. The linear model assumes that the numerosity is represented on a linear scale, and variability increases with numerosity.

It has been claimed that the log and linear models cannot be discriminated using behavioral data (Dehaene, 2003). Although the models produce predictions that are not identical, they are very similar (e.g., for values of parameters that produce predictions for accuracy similar to those in our data, the predictions match to within 1%). Both of them are able to explain two standard findings in numerosity judgment tasks, the distance effect and the size effect. As the difference (distance) between two numerosities increases, the discrimination becomes easier. For example, 30 vs 10 is easier to discriminate than 15 vs 10. The linear differences between these two pairs are 20 and 5, which allows the linear model to predict that the discrimination is easier for the first pair. Also, because log (30/10) = 1.10 is greater than log(15/10) = 0.405, the log model is also capable of predicting the distance effect. Furthermore, as the total number of items increases, performance decreases. For example, 20 vs 30 is easier to discriminate than 60 vs 70. Because log (30/20) = 0.405 is greater than log(70/60) = 0.154, the log model is able to explain the size effect. Although linear differences of these two pairs are the same (i.e., 10), the linear model can also predict the size effect based on its variability assumption. The variability in the linear model is an increasing function of the numerosities. 60 vs 70 has higher variability than 20 vs 30, which makes discrimination more difficult. As can be seen from these examples, the two models yield similar predictions.

The ANS models do not explain how non-numeric perceptual variables, such as dot area and density of dots, affect numerosity judgments. Most researchers agree that numerosity discrimination is not solely determined by the numerosity of the stimuli. There are many perceptual variables confounded in the stimuli and they affect numerosity judgments (Abreu-Mendoza & Arias-Trejo, 2015; DeWind, Adams, Platt, & Brannon, 2015; DeWind & Brannon, 2012; Feigensen, Carey, & Hauser 2002; Gevers, Cohen Kadosh, & Gebuis, 2016; Gebuis, Cohen Kadosh, & Gevers, 2016, 2017; Gebuis & Reynvoet, 2012a, 2012b, 2013; Halberda, Mazzocco, & Feigenson, 2008; Im, Zhong, & Halberda, 2016; Leibovich et al., 2017). For example, when we compare the number of dots in two side-by-side arrays, we may overestimate the number in one array if it has larger dots and the dots are more widely dispersed than the dots on the other side. Thus, an important goal of research in numerosity is to uncover the roles of numerosity and perceptual variables.

Currently, there are several competing views of how we process numerosity and perceptual variables in making judgments during numerosity discrimination tasks. First, some researchers believe that, despite the effect of non-numeric features, numerosity has a separate effect on judgments from other variables in a stimulus. This point of view is consistent with the concept of 'the number sense' proposed by Dehaene and others (Dehaene & Changeux, 1993; Dehaene, 2003; Feigensen et al., 2004; Verguts & Fias, 2004). This approach has attempted to develop computational models that extract numerosity information from visually presented stimuli based

on concepts such as the numerosity detection system (Dehaene & Changeux, 1993) and number-selective neurons (Verguts & Fias, 2004). Another way to study the pure number sense is to develop numerosity-related tasks designed to control the effects of perceptual variables. For example, suppose that a subject is asked to decide which of two side-by-side arrays contains more dots. If the dots on the left side are more dispersed, and so cover more area (i.e., a larger field area) than the dots on the right side, this will affect numerosity discrimination. In order to control for this confounding effect, the total areas of dots can be kept constant across two arrays (Abreu-Mendoza & Arias-Trejo, 2015). Similarly, other perceptual variables can be kept constant across different groups of dots to control their effects. It is also possible to control different perceptual variables on different trials. For example, on some trials, dot area can be set equal between two groups of dots, while in the other trials, the constraint can be the total areas of dots (Halberda et al., 2008).

In another perspective on how numerosity judgments are made, researchers argue that numerosity cannot be processed independently of non-numeric perceptual features and in fact, both numerosity and perceptual variables are holistically processed to produce an estimate of the number in a given stimulus (Leibovich et al., 2017). Researchers with this perspective have pointed out that it is not possible to control all the perceptual variables. For example, in the task with two side-by-side arrays, keeping the convex hull (a circular area in which dots are distributed) equal across two arrays would make dots more densely distributed for larger numerosity. That is, although the influence of the convex hull might be controlled by the manipulation, this produces a confounding effect from the density of the dots. Leibovich et al. (2017) have reviewed many of the numerosity judgment tasks designed to control or minimize the confounding influences of the perceptual variables mixed in a stimulus. Results showed that in every task examined, despite all the controls, another confounding variable is introduced (Dehaene, Izard, & Piazza, 2005).

Leibovich et al. suggested that numerosity is processed along with other perceptual variables to construct a sense of magnitude. This sense of magnitude, instead of the sense of number, plays the major role in numerosity judgments. This perspective, which is termed the Holistic Process view, contrasts with the number sense view in that it claims the numerosity cannot be processed independently of perceptual features in numerosity judgments.

The Sensory Integration Theory casts doubt on the concept of a number sense. The main claim is that humans are not endowed with an innate ability to approximate pure numerosity (Gebuis & Reynvoet, 2012a, 2012b, 2013; Gebuis et al., 2016, 2017; Gevers et al., 2016). Gebuis and Reynvoet (2012a, 2013) provided experimental and neural evidence that numerosity is not processed independently from perceptual features. Instead, they proposed that sensory inputs from a stimulus are integrated to construct a numerosity estimate. According to the sensory integration theory, the sensory-integration system, rather than an ANS, is used to process perceptual information from the stimulus. Unlike an ANS, this system does not require a pure number sense in which numerosity information is processed separately from visual cues. Instead, the sensory-integration system uses a weighted combination of visual/perceptual information as the basis for estimation of numerosity. Gebuis and colleagues argued that this parsimonious system is empirically more plausible and capable of explaining patterns observed from behavioral data (Gebuis et al., 2016, 2017).

At this point, it is not easy to say which of these perspectives is more compelling. Researchers with different perspectives have provided lines of evidence with various numerosity judgment tasks to support their perspectives. However, it is puzzling that results from similar tasks and experimental settings have been used to provide evidence supporting different points of view. One important factor is the issue of measures. Many previous studies have focused on accuracy of performance in numerosity judgment tasks (Mazzocco, Feigenson, & Halberda, 2011; Desoete, Ceulemans, De Weerdt, & Pieters, 2012; Gebuis & Reynvoet, 2012a; Praet, Titeca, Ceulemans, & Desoete, 2013), or the Weber fraction (w), an accuracy-based measure of ANS acuity (Inglis, Attridge, Batchelor, & Gilmore, 2011; DeWind & Brannon, 2012; Gilmore et al., 2013; DeWind et al., 2015). Another important measure in decision-making tasks, reaction time (RT), has been ignored in many numerosity studies. When RT was the measure examined, only the correct mean RTs were compared (Libertus, Feigenson, & Halberda, 2011; Park & Brannon, 2014). In some studies, RT was used to define the numerical distance effect (Holloway & Ansari, 2009; Lonnemann, Linkersdoerfer, Hasselhorn, & Lindberg, 2011; Ferreira, Wood, Pinheiro-Chagas, Lonnemann, Krinzinger, Willmes, & Haase, 2012; Sasanguie, De Smedt, Defever, & Reynvoet, 2012, 2013). Regression analysis has also been used to examine whether RT can be reduced by an experimental manipulation such as repetitive training of the task (DeWind & Brannon, 2012).

However, in most analyses using either accuracy or RT, it has not been fully appreciated that choice and RT emerge from the same decision process (Ratcliff, Smith, & McKoon, 2015). One way to integrate these two variables is through sequential sampling models. In these models, choice and RT are assumed to arise from a latent cognitive process in which we continuously accumulate evidence from a stimulus to make a decision. This modeling approach has been used in many psychophysical, memory, psycholinguistic, semantic, and value-based decision-making tasks. For example, in the diffusion decision model (Ratcliff, 1978; Ratcliff & McKoon, 2008), there is one accumulation process with two separate boundaries, each of which represents one of two choice options. The decision is made when the amount of evidence accumulated reaches one of the two boundaries. Decision time is the time that the accumulation process takes to arrive at the threshold and the choice made corresponds to the threshold arrived at. In other models such as the leaky competing accumulator model (Usher & McClelland, 2001) or the linear ballistic accumulator model (Brown & Heathcote, 2008), two accumulators, each with a different single boundary, correspond to different choice options. Decision time is the time when the first accumulator hits its boundary and the option corresponding to that accumulator is chosen. All these models show how RT and choice are generated from a common framework.

These models require that both accuracy and RT distributions for both correct and error responses need to be explained at the same time to understand how we process information during a numerosity judgment task. If only one of the two variables is considered, the analysis is incomplete and this sometimes can lead to misleading interpretations of processing. For example, it has been pointed out that accuracy and RT can trade off in numerosity judgments (Ratcliff & McKoon, 2018). Specifically, some subjects may

emphasize accuracy more than speed while other subjects may emphasize speed more than accuracy. This trade-off implies that it is misleading to assess an individual's acuity in a numerosity judgment solely by either of accuracy or RT. Only when both accuracy and RTs, including their distributions, are explained simultaneously can the quality of numerosity information processing be measured appropriately.

Ratcliff and McKoon (2018) suggested an alternative way to study the decision process underlying numerosity representation. They argued that the ANS models can be integrated into the diffusion model (ANS diffusion models). The idea is that the drift rate parameter (v) of the diffusion model, which represents the mean rate of information processing, is determined as a function of the numerosities in a stimulus (i.e., $v = f(N_1, N_2)$ where N_i is the value of the numerosity, $i = 1, 2$). This contrasts with a typical application of the diffusion model in which drift rates for the different conditions are freely estimated. Depending on the ANS model chosen, different drift rate expressions can be examined. Thus, the competing ANS theories can be studied by a model comparison procedure. Since this approach uses the diffusion model, RT and accuracy are examined simultaneously.

One major result from the application of the ANS diffusion model in Ratcliff and McKoon (2018) is that this approach successfully discriminated between the different ANS models. As mentioned above, the linear model with increasing variability as numerosity increases and the log model with constant variability are both capable of explaining standard findings (e.g., the distance effect and the size effect) and it has been claimed that they cannot be discriminated (Dehaene, 2003). The argument that the two models cannot be discriminated was based on accuracy measures alone in the numerosity judgment tasks. By jointly modeling accuracy and reaction time distributions, Ratcliff and McKoon (2018) showed that the log and the linear models were distinguished when integrated with the diffusion model. They predicted similar patterns of accuracy but produced different reaction time patterns and different models were preferred by different types of numerosity tasks. In typical perceptual and cognitive tasks with one stimulus, accuracy decreases and RT becomes longer as the discrimination becomes more difficult. However, Ratcliff and McKoon found that, when the task is to decide whether there are more blue or more yellow dots in a single array in which dots of the two colors are intermingled, RT followed a counterintuitive pattern in which, for a fixed difference in the number of dots, RT was shorter in more difficult conditions in which the overall number increased (see the application by Ratcliff, Voskuilen, & Teodorescu, 2018, to perceptual tasks with two stimuli). In this task, the linear model performed better in capturing the counterintuitive RT pattern. In another task in which two arrays were presented side by side, the log model performed better in explaining the typical pattern in which RT increased for a fixed difference between numbers and increasing overall number. These results show that the two representations of numerosity can be discriminated. Note that it was the behavior of RTs that made it possible to distinguish the two ANS representations. This shows the need for models of processing with joint consideration of accuracy and RT distribution when studying numerosity discrimination.

In this article, we extend the ANS diffusion model to examine how non-numeric perceptual variables interact with numerosity in numerosity discrimination tasks. As discussed earlier, both numerosity values and perceptual features affect numerosity judgments, but how they interact during tasks has not yet been examined with models that explain both accuracy and RT. The three different perspectives introduced above suggest different ways that the non-numeric variables may engage in numerosity tasks. According to the pure number sense view, the ANS provides an estimate of numerosity, but this is modulated by perceptual variables confounded in a stimulus. According to the other two perspectives from Leibovich et al. (2017) and Gebuis, Cohen Kadosh, and Gevers (2016), it is also possible that different variables (which may or may not include numerosity) interact to produce a single value of magnitude, upon which subjects base their decision in numerosity tasks. Mathematically, models based on these two views of how perceptual variables work in numerosity judgments might be the same and the two views might not be able to be discriminated.

Although many previous studies tried to control for the effects of continuous perceptual magnitudes, it is not possible to control all variables. Instead, what we can do is to explicitly model the effect of all the variables in a task. Ratcliff and McKoon (2018)'s ANS diffusion model can be extended to do this. Although the focus of their research was on the comparison between the linear and the log models across a wide range of numerosity discrimination tasks, they also investigated the effect of dot area on judgments by comparing two different conditions (in several of the experiments). In one condition, the total area was proportional to the numerosity (area-proportional), while in the other, the total area was fixed to be equal across two groups of dots (area-equal). In this study, we extend the tasks used in Ratcliff and McKoon (2018) by allowing perceptual variables (e.g., dot area) to take on a range of different values. For example, a stimulus in each trial has different numerosity pairs and also different dot areas or convex hulls. This provides a large number of combinations of values that can be used in an analysis. Thus, unlike Ratcliff and McKoon (2018)'s approach which compared the two different conditions of dot area, we will use perceptual features as variables in regression models of the drift rate. This allows us to separately estimate the effect of each independent variable on the drift rate. An estimated drift rate coefficient of numerosity represents its effect controlling for the other perceptual variables included in the analysis. Similarly, an estimated drift rate coefficient of any perceptual feature represents its effect controlling for the numerosity and the other perceptual variables. Therefore, their relative contributions to a decision process can be estimated and compared. As in Ratcliff and McKoon (2018), the value of drift rate is provided by the ANS representations. However, in our extended framework, the non-numeric variables are also included as a part of the expression for drift rate.

Hence, the model is capable of accounting for the effect of not only each variable itself but also their relationships, such as the interaction between numerosity and perceptual variables. Therefore, by developing a model of the interaction of the variables, the ANS diffusion model approach produces a measurement tool for the interaction of numerosity and perceptual variables. To the extent that the drift rate alone does not account for the interaction, it is possible to model the effects on other model parameters.

2. Preview

We conducted experiments in which numerosity and non-numeric perceptual features were manipulated in numerosity or

perceptual judgments. Results showed that perceptual features, such as dot area and convex hull, affected non-symbolic numerosity judgments (as in Ratcliff & McKoon, 2018), and numerosity affected total area judgments.

The tasks used in this article had many more combinations of the two kinds of variables including the conditions in which the two variables were in conflict. This produced a new result that the perceptual features can dominate numerosity judgments when there is both a conflict between numerosity and non-numeric features and there are many dots in a stimulus. Under these conditions, subjects consistently made incorrect choices with accuracy much lower than chance in the most extreme conditions. This conflict effect was consistent across subjects and the tasks examined. The conflict effect also produced a large shift in the leading edge of the RT distributions for the numerosity judgments but not the perceptual judgments.

We explored several models that were extensions of the models in Ratcliff and McKoon (2018) in which drift rate was assumed to be linear and log functions of numerosity and perceptual features, but these models were not able to account for the conflict effect in accuracy. Eventually, we produced a model of drift rate with an interaction term which is a function of perceptual features multiplied by the total number of dots. This interaction model was able to account for the data over all the conditions, and in particular, it was able to produce accuracy lower than chance in the conflict conditions. However, this model could not account for the shift in the leading edge of the RT distribution with conflict. To accommodate this shift, we represented nondecision time as a function of numerosity and non-numeric variables. The final model does a reasonably good job of capturing most (but not all) of the shift in the RT leading edges.

Results for Experiment 4 showed that the conflict effect was asymmetric between numerosity judgments and area judgments. When subjects were asked to judge area, not numerosity, a similar conflict effect was observed but there was almost no shift in the RT distributions for the high conflict conditions. Thus, the interference due to confounding variables exerted different effects depending on the task.

3. Modeling

3.1. The ANS integrated diffusion model

It is informative to describe the evolution of our modeling. The initial models we tried were directly based on the models in Ratcliff and McKoon (2018). The aim was to produce a compact but complete analysis of accuracy and RTs of the data similar to the analysis from the accuracy-based model of DeWind et al. (2015). However, some of the results were unexpected and could not be explained by the simple models. Two examples are a crossover in which responses had less than chance accuracy when non-numeric variables are in opposition to numerosity for large values of numerosity, and shifts in the leading edges of the RT distributions when non-numeric variables are in opposition to numerosity variables. These results exposed a new kind of conflict effect that is not handled by a standard diffusion model with only drift rate changing over conditions. This led us to develop elaborated models to account for the misses between theory and data from these earlier models. These models did a good job of accounting for almost all of the data but with minor misses in some aspects of RT distributions in conflict conditions with high conflict. The elaborated models provide estimates of numerosity information being used in making decisions in numerosity judgment tasks. However, the final models have a more ad hoc nature than the models that were the starting point of this project.

The diffusion model (see Appendix A for details) can serve to connect the ANS models to the behavior of reaction time and accuracy. In Ratcliff and McKoon (2018), drift rate and across-trial variability in drift rate were modeled as a function of numerosity (the ANS Diffusion Model). In the linear model, drift rate is a function of the difference between two numerosities. Thus, the drift rate (v) for a given numerosity pair can be computed as a drift coefficient (v_1) multiplied by the linear difference of the numerosities, $v = v_1(N_1 - N_2)$. Across-trial variability of the drift rate (η) is also modeled as a regression equation which includes an intercept (η_0) plus a slope coefficient (σ_1) multiplied by the square root of the sum of two squared numerosities, $\eta = \eta_0 + \sigma_1\sqrt{N_1^2 + N_2^2}$. For the log model, drift rate is a product of a drift coefficient and the log ratio (or log difference) between two numerosities, $v = v_1(\log(N_1) - \log(N_2))$. In this log model, across-trial variability of drift rate is often assumed to be constant (Dehaene, 2003), but we use the same expression used in the linear model to give the models the same flexibility.

Ratcliff and McKoon (2018) showed that the linear and log models can be discriminated based on their RT results. In the task in which a subject is asked to decide which of the two side-by-side arrays contains more dots (termed the L/R task), a typical pattern is observed: for a constant difference in the number of dots in the two displays, as task difficulty increases with increasing overall number of dots, accuracy decreases and RT increases. In contrast, in the task in which a subject is asked to decide if there are more blue dots or more yellow dots in a single array (termed the B/Y task), a counterintuitive pattern is observed: for a constant difference in the number of dots in the single display, as task difficulty increases with increasing overall number of dots, accuracy decreases but RT decreases instead of increasing. Although the linear and log models produced very similar predictions for accuracy, their RT predictions showed qualitatively different patterns. The log model was able to reproduce the typical RT pattern which is observed in the L/R task, while the linear model was able to reproduce the unexpected RT pattern observed in the B/Y task. The conclusion was that 1) different processes are used depending on how a stimulus is processed and 2) the log representation is preferred when two separate objects are compared in a whole display whereas the linear representation is preferred when the relative number of two sets of objects intermingled in a single array is evaluated.

Results from Ratcliff and McKoon (2018)'s experiments showed that perceptual features, such as dot areas, affect numerosity judgments and the ANS diffusion model provided a way to measure the size of the effect. Ratcliff and McKoon compared two different conditions. On some trials, the total area of the dots was equal across two groups of dots regardless of their numbers. On other trials,

the areas were proportional to the numerosities. They fitted the ANS diffusion models allowing drift rate coefficients to differ between conditions which allowed the two conditions to be compared. In the B/Y task, the drift rate coefficient was more than two times larger in the area-proportional condition than the area-equal condition. In the L/R task, the drift rate coefficient was about 35% higher in the proportional-area condition. These results provided strong evidence that numerosity judgments are not determined solely by numeric properties of the stimuli in these tasks, but are also affected by non-numeric perceptual characteristics, and the contribution of the perceptual features is modulated by the type of task.

The modeling presented here aims to extend the modeling of Ratcliff and McKoon (2018) by using numerosity and perceptual variables as independent variables in the drift rate equation in the ANS diffusion model. The regression model of the drift rate includes coefficients for each of the variables and so the model is compact with relatively few parameters compared to the number of degrees of freedom in the data (all the combinations of numerosity and the perceptual variables). We present a number of models with different drift rate representations and evaluate them. The models are fit to binned data (response proportions and RT quantiles) obtained by grouping data based on numerosity and perceptual variables. A G^2 multinomial likelihood method is used to fit models to the data and the G^2 values averaged over subjects are provided to describe the model fits to data (see Appendix A for details). The G^2 method is robust to contaminants (Ratcliff & Tuerlinckx, 2002) but it does not use the information contained in all choices and RTs. In contrast, Maximum likelihood estimation (MLE) uses all the individual data points (Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002) and if there are no contaminants or their effects have been minimized, then the MLE can produce smaller standard errors. We fit the data from one of the experiments with the MLE to show that the two fitting methods produce similar results in parameter estimation and model comparison and show advantages of the MLE methods in modeling the effects of many perceptual variables on numerosity judgments (Section 5).

Recently, DeWind et al. (2015) proposed a regression modeling approach for numerosity tasks but applied only to accuracy data. The model can be understood as an extension of Piazza, Izard, Pinel, Le Bihan, and Dehaene (2004), Piazza et al. (2010)'s model. The model is based on Dehaene and Changeux (1993)'s view that numerosity is encoded by "numerosity detectors" responsive to a specific numerosity which is normally distributed on an internal logarithmic scale. DeWind et al. (2015) extended Piazza et al.'s model by modeling the contribution of non-numeric features in numerosity tasks. They first proposed that all features, including numeric and non-numeric ones, can be represented on a three-dimensional space. Then, they defined *Size* and *Spacing* variables which were functions of numerosity and perceptual variables. These variables were designed to be orthogonal to numerosity and to each other, and to represent the axes of the three-dimensional feature space along with an axis for numerosity. This resulted in a choice probability model that represents the probability as a function of numerosity, size, and spacing (Appendix B).

DeWind et al. (2015)'s modeling is similar to our ANS diffusion models because they both take a regression-based approach with numerosity and non-numeric features as covariates. DeWind et al.'s model does not take RTs and their distributions into account. However, its choice predictions can be compared and contrasted with predictions from the ANS diffusion models. Following the experimental results, a comparison of these two approaches based on the choice probabilities will be provided (Section 6). More details of DeWind et al.'s model and a theoretical comparison of it and the ANS diffusion model can be found in Appendix B.

4. Experiments

4.1. Experimental methods

Fig. 2 shows examples of stimuli that were used in the following four experiments. These stimuli were adapted from Experiments 1 and 2 in Ratcliff and McKoon (2018). Fig. 2A shows a stimulus used in the B/Y task. There are blue and yellow dots intermingled in a single array. The task is to decide whether there are more blue dots or more yellow dots, as accurately and quickly as possible. Fig. 2B shows a stimulus used in the L/R task. There are two side-by-side arrays, each of which contains many yellow dots. The task is to decide whether there are more dots on the left side or the right side.

In Ratcliff and McKoon (2018), the main focus was on the numerosity variable. For the perceptual variable, for the blue and yellow dots or the left and right dots, dot areas were either equal to each other or dot sizes were selected randomly so that the area was proportional to the numerosity. In the experiments presented below, dot areas were varied in the B/Y task and both dot areas and convex hulls were varied in the L/R task. The values of these variables and numerosity variables were selected randomly (from a restricted range of possibilities) to produce a large number of combinations in the experiment.

In Experiments 1, 3 and 4, subjects participated in tasks using stimuli in Fig. 2A. In Experiment 1, stimuli were presented for 300 ms and then removed (to prevent counting or other slower strategies). However, the presentation duration of 300 ms might be too short and have some unexpected effect on performance, and so in Experiment 3, we allowed the dots to remain on the screen until the decision was made. Experiment 4 used the same stimuli and method as Experiment 1, but the task was to judge the total area covered by dots, not the numerosity. This experiment was performed to examine the effect of numerosity on perceptual area judgments.

In Experiments 1, 3, and 4, two variables were manipulated; the number of dots (numerosity) and the area of the dots. By doing so, we aimed to extend the log and linear models for numerosity judgment tasks to the non-numeric variables. There were 14 numerosity pairs used: 15/10, 25/20, 40/35 (for a difference of 5), 20/10, 40/30 (for a difference of 10), 30/10, 40/20 (for a difference of 20) and their opposites. With these pairs, we can manipulate numerosities of two groups of dots and the difference between them. Six dot radius values, 6, 8, 10, 12, 14, and 16 pixels, were used. In each trial, a dot radius for each color was randomly selected and all the dots with the same color had the same dot radius. Because dot radii and numerosity pairs were randomly selected, each combination did not have equal numbers of trials in a block or across the experiment.

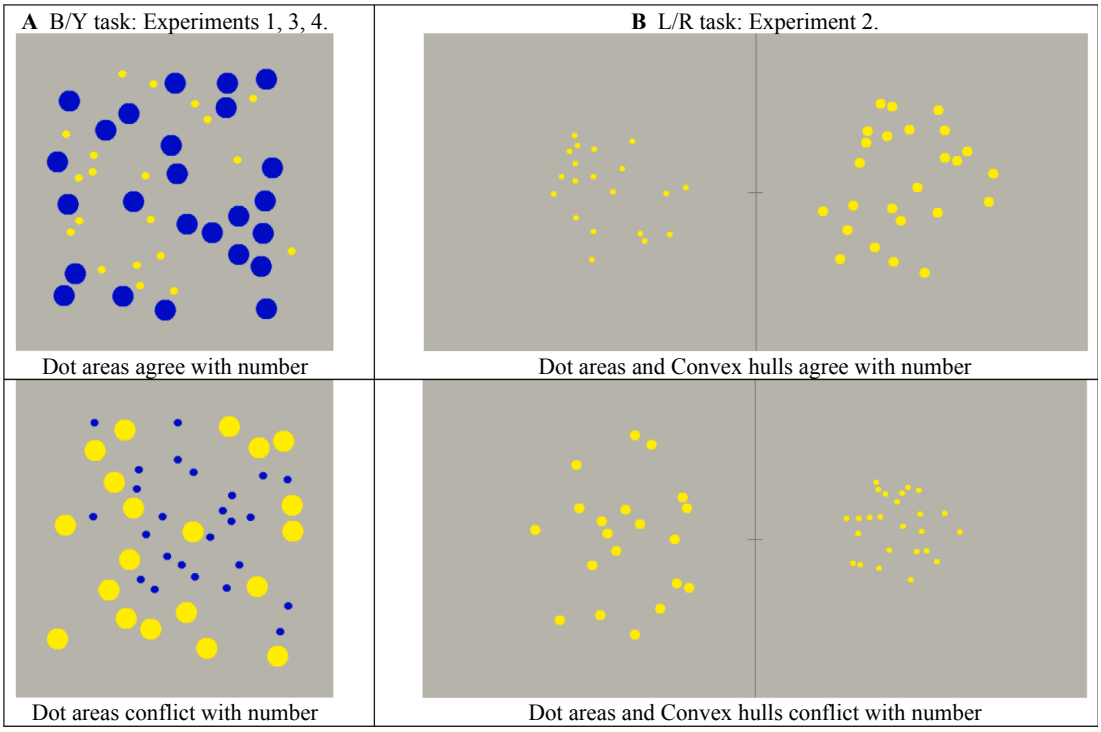

Fig. 2. Examples of stimuli for the two numerosity discrimination tasks used in experiments.

The L/R task was used for Experiment 2. This task used the same numerosity pairs as the B/Y task in Experiment 1. Dot radii were also selected in the same way as in the B/Y task, except that the values of dot radii in this task were 3, 4, 5, 6, 7, and 8 pixels. All dots on the same side of the stimulus display had the same dot radius and the radii of the left and the right arrays were selected randomly and hence were independent of each other. In this task, one more perceptual feature, the convex hull, was varied along with the numerosity pairs and the dot radii. The radius of the convex hull was uniformly selected from 85 to 160 pixels in increments of 5 pixels and the radius of the left side was determined independently from that of the right side. The centers of the convex hulls were equally spaced from the vertical center-line and they were 400 pixels apart.

For all the tasks, subjects were instructed to make a choice by pressing one of two keyboard buttons; ‘z’ for one choice and ‘/’ for the other choice. At the beginning of each experiment, there were 4 practice trials. Each of these trials showed the correct answer before displaying a stimulus (e.g., in the B/Y task, it would say “An example of more yellow dots” and then showed a stimulus with more yellow than blue dots in a single array). This ensured that subjects could understand the instructions. During the task, a 'TOO FAST' message was presented for 1500 ms if subjects responded before 280 ms. Also, a 'TOO SLOW' message was shown for 500 ms when subjects answered with an RT greater than 1500 ms. Each subject was tested on 20 blocks of 98 trials, providing close to 2000 responses. After each block, subjects began the next block by pressing the space-bar on the keyboard. Subjects were instructed that they could take a break before initiating each block if they needed to do so.

The experiments used the same monitor screens as Ratcliff and McKoon (2018). They were 17-inch diagonal CRT monitors, which were 32 cm wide and 24 cm high. The 4 × 3 screen was set to 1,280 × 960 pixels with 256 colors. In the B/Y task, blue and yellow dots were distributed in a 360 × 360 playground on a 640 × 640 gray background. The playground was centered on the background located at the center of the screen. The minimum spacing between dot edges was 5 pixels and the maximum horizontal/vertical distance that dot centers could be separated by was 360 pixels. In the L/R task, a stimulus for each trial was given on the gray background with 640 × 1160 pixels. There was a thin vertical line at the center of the background separating the left and the right sides. At the center of this line, there was a small cross at which subjects were instructed to focus during the task. Dots were displayed within each array. The minimum spacing between dot edges was 5 pixels. The minimum and maximum horizontal/vertical distance by which dot centers could be separated between each patch were 80 and 540 pixels, respectively. The maximum horizontal/vertical distance that dot centers could be separated by in each patch was 230 pixels.

The subjects for the experiments were students in an introductory psychology class at the Ohio State University and they participated in the experiments for class credit. Informed consent was obtained from all subjects. Some of the subjects were eliminated before the main analysis because they were uncooperative and responded either with chance accuracy or many fast guesses. We diagnosed fast-guessing subjects by setting RT cutoffs of 0 ms (lower bound) and 300 ms (upper bound). If the proportion of responses within this RT range was high (e.g., 5% or higher) and their accuracy was about chance, the subject was excluded from further analysis. For Experiment 1, 29 subjects participated and all were used in the analyses. For Experiment 2, 38 subjects participated in the task but 4 of them were removed, one subject due to chance accuracy, and the other three subjects due to fast guessing, leaving 34

subjects. Experiments 3 and 4 had 40 and 15 subjects, respectively, with none eliminated. For the main analysis, we set lower and upper bounds for RT, which were 300 ms and 2000 ms, respectively. Data points with RT below the lower bound or above the upper bound were excluded. This eliminated less than 2% of the trials in Experiments 1 and 3 and less than 5% in Experiments 2 and 4.

4.2. Experiment 1

In Experiment 1, subjects were asked to decide if there were more blue dots or yellow dots in a single array (the B/Y task). Correct responses for blue dots and for yellow dots were aggregated and error responses for blue and yellow dots were aggregated because the choice data for blue and yellow responses were symmetric in both accuracy and RT. The different numerosity pairs and dot areas produced 252 conditions which means that each condition had only a small number of observations. Thus, to examine the data and to produce groups for the G^2 fitting method (using RT bins), we defined four conditions based on dot areas. With 7 numerosity pairs we used, this yielded 28 combinations for the G^2 method. The conditions we used were 'Agree', 'Conflict', 'Large', and 'Small'. Dot radii varied from 6 to 16 pixels by a difference of 2 and so their mean value was 11. The 'Agree' condition included trials in which a color with the larger numerosity had a dot radius larger than the mean (i.e., 12, 14, or 16) and the other color with the smaller numerosity had a dot radius smaller than the mean (i.e., 6, 8, or 10). For example, if there were more blue dots in an array and their common radius was 14 while the radius of yellow dots was 6, this trial was included in the agree condition. We called the opposite case the 'Conflict' condition and this corresponded to conditions in which the color with more dots had dots with a smaller dot radius and conditions in which the color with smaller numbers of dots had dots with a larger radius. Thus, dot areas conflicted with numerosity. In the 'Large' and the 'Small' conditions, both of the dot radii were larger or smaller than the mean. Because there was little difference in performance between the large and small conditions, we sometimes refer to these two as the 'Equal' condition. This grouping of the data also helped with the presentation of model fits to the data.

4.2.1. Results

Table 1 shows the accuracy and mean correct RTs by different area conditions, averaged over numerosity pairs. The results are consistent with those from previous research; accuracy was lower and RT was longer when the areas were equal across two groups of dots than when the areas were proportional to the numerosities (Ratcliff & McKoon, 2018). In the current task, the agree condition is similar to the area-proportional condition, and the equal conditions are similar to the area-equal condition in the previous study. The conflict condition did not appear in Ratcliff and McKoon's experiments. We expected that this condition would be more difficult than the other conditions, and thus, would have the lowest accuracy and the longest RT. This result was obtained as is shown in Table 1. Among the equal conditions, accuracy was lower and RT was slightly longer in the small condition than in the large condition. But the difference is relatively small compared to the differences between other pairs of conditions.

Table 2 shows accuracy and mean RTs for correct responses as a function of numerosity for the agree, the conflict, and the large and small conditions combined (equal). The accuracy values showed an interesting pattern that has not been observed in previous numerosity tasks: in conflict conditions with large numerosity values, decisions were dominated by the dot areas, which led accuracy to fall below chance (the bolded entry in Table 2). Ratcliff and McKoon (2018) found that accuracy tends to fall as the total number of dots increases for constant differences in the number of blue and yellow dots. Accuracy in Table 2 confirmed this pattern. However, there was no condition in which there were more errors than correct responses in Ratcliff and McKoon's study. In Ratcliff and McKoon's experiments, there were only two conditions, area-equal and area-proportional, and so there were no conflict conditions. In contrast, in Experiment 1, dots were of the same size for a given color and were allowed to have a range of different radii from 6 to 16. Thus, for example, when there were 40 yellow dots and 35 blue dots in a single array, the radius of yellow dots could be 6 while that of blue dots could be 16, and then the total area covered is $40(6)^2\pi$ for yellow dots and $35(16)^2\pi$ for blue dots, which is a ratio of 1 to 6.22. In conflict conditions like this (on average less extreme than this example), accuracy was 0.379 which was below chance. This pattern was observed when there were many dots in the array and the difference between the two numbers of the dots was small ($\Delta = 5$). The pattern occurred consistently across subjects; 24 of 29 subjects chose correct answers only for 20–30% of the trials and the other 5 also showed low accuracy (0.548–0.651). This result shows that under some conditions, numerosity judgments are dominated by non-numeric features such as dot areas. Specifically, when there are many dots in a stimulus display and the difference in numerosity is small, the signal from numerosity is small. When there is conflict between numerosity and perceptual features and the numerosity signal is small, the non-numeric contribution dominates and this produces lower than chance performance in numerosity judgments. In contrast, given a constant difference in numerosity, as the number of dots in a display decreases, the signal from numerosity increases in size and dominates non-numeric information and decisions are above chance.

Table 1
Experiment 1 (B/Y task): Accuracy and correct mean RT. Data are divided into different conditions by the values of the dot areas.

Experiment and task	Conditions	Accuracy	Mean RT
1, B/Y	Agree	0.867	607
	Large	0.800	636
	Small	0.781	647
	Conflict	0.640	668

Table 2
Experiment 1: Accuracy and mean RTs for correct responses as a function of numerosity pairs and the three area conditions (Agree, Equal, and Conflict).

Expt, Task	Measure	Condition	Δ = 5			Δ = 10		Δ = 20	
			(15, 10)	(25, 20)	(40, 35)	(20, 10)	(40, 30)	(30, 10)	(40, 20)
1, B/Y	Accuracy	Agree	0.839	0.820	0.832	0.889	0.867	0.905	0.916
		Equal	0.769	0.679	0.652	0.868	0.754	0.916	0.889
		Conflict	0.659	0.506	**0.379**	0.791	0.532	0.858	0.769
	Mean RT	Agree	641	625	612	591	590	572	576
		Equal	681	677	653	627	644	585	603
		Conflict	747	718	704	680	711	617	660

When there were few dots (e.g., 15 vs 10), or the difference was large (e.g., 40 vs 20), numerosity dominated area and the effect of the conflict was reduced. In these cases, accuracy was higher than chance despite the conflict between numerosity pairs and dot areas. This finding implies that there is an interaction between numerosity information and perceptual features in processing a stimulus.

Mean RTs for correct responses in Table 2 showed two important patterns. In general, the mean RTs decreased as numerosity difference increased. For each of the numerosity difference values, different patterns were obtained: when the difference was 5, mean RTs decreased as the total number of dots increased. When the difference was 10 or 20, mean RTs increased or remained constant as the total number of dots increased. The speed-up pattern obtained when the numerosity difference was 5 is counterintuitive because RTs usually get longer as the task difficulty increases and accuracy decreases. This finding is consistent with the results from the B/Y task in Ratcliff and McKoon (2018).

4.2.2. Model development

It is a challenge for any mathematical model of numerosity judgments, including the ANS diffusion models, to explain the behavioral patterns we observed in the Experiment 1. The model needs to capture accuracy and reaction time patterns across all the experimental conditions. In our experiments, the combination of different numerosity pairs and dot radii produced a large number of conditions. Even when we aggregated data by the conditions defined above, there were 28 combinations to be fit (7 numerosity pairs with 4 congruency conditions: agree, conflict, small, and large).

The aim of the ANS diffusion models in the current study is to account for accuracy and RT distributions by modeling the joint effects of the numerosity and the perceptual/non-numeric variables. The drift rate equation should produce positive drift rates in most of the conditions, but in some of the conflict conditions, the contribution from the non-numeric perceptual variable should dominate to produce negative drift rates. With this purpose in mind, several variants of the integrated diffusion models were built and examined. We started model development with early models that are not capable of producing negative drift rates for the conflict conditions and then progressed to models with interaction terms that do produce negative drift rates (for an early suggestion of this way of integrating different signals including conflict signals, see Logan, 1980).

First of all, we fitted the same models used in Ratcliff and McKoon (2018). In that study, Ratcliff and McKoon compared the linear and log models of the drift rate and found that the linear model is preferred in the B/Y task while the log model is preferred in the L/R task. Also, by allowing the drift rate coefficients to vary by different area conditions (area-equal and area-proportional), they found

Table 3
Models used to fit data from Experiment 1, their G^2 values, and degrees of freedom (*df*). N_1 and N_2 are two numerosities and A_1 and A_2 are two dot areas. Dot area is computed as $A_i = \pi * r_i^2$, where r_i is a dot radius. For all of the models, across-trial variability of drift rate was modeled by $\eta = \eta_0 + \sigma_1\sqrt{N_1^2 + N_2^2}$.

Model			G^2	*df*
1	Linear	$v = v_{1j}(N_1 - N_2)$	433.4	222
2	Log	$v = v_{1j}(\log(N_1) - \log(N_2))$	451.7	
3	Linear - Linear	$v = v_1(N_1 - N_2) + v_2(A_1 - A_2)$	404.7	300
4	Linear - Log	$v = v_1(N_1 - N_2) + v_2(\log(A_1/A_2))$	400.1	
5	Log - Linear	$v = v_1(\log(N_1/N_2)) + v_2(A_1 - A_2)$	390.1	
6	Log - Log	$v = v_1(\log(N_1/N_2)) + v_2(\log(A_1/A_2))$	386.4	
7	Linear - Linear Interaction	$v = v_1(N_1 - N_2) + v_2(A_1 - A_2) * (N_1 + N_2)$	381.1	
8	Linear - Log Interaction	$v = v_1(N_1 - N_2) + v_2(\log(A_1/A_2)) * (N_1 + N_2)$	381.0	
9	Log - Linear Interaction	$v = v_1(\log(N_1/N_2)) + v_2(A_1 - A_2) * (N_1 + N_2)$	386.9	
10	Log - Log Interaction	$v = v_1(\log(N_1/N_2)) + v_2(\log(A_1/A_2)) * (N_1 + N_2)$	384.2	
11	Linear - Log Interaction + Moderated Numerosity	$v = v_1(N_1 - N_2) + v_2(\log\left(\frac{A_1}{A_2}\right)) * (N_1 + N_2) + v_3\frac{(N_1 - N_2)}{N_1 + N_2}$	358.7	299
12	Linear - Log Interaction + Moderated Numerosity + T_{er} Model	Model 11 for the drift rate v with $T_{er} = t_0 + t_1\frac{\log(A_1) - \log(A_2)}{(N_1 + N_2) * (N_1 - N_2)}$	353.0	298

that the area manipulation affected the difficulty of the task and the quality of information processing during the task. In these linear and log models, drift rates are determined by regression equations given in Table 3 (Model 1 and 2). For these models, we allowed the drift rate coefficient to vary by the three conditions (agree, conflict, and equal) following Ratcliff and McKoon's application. Across-trial variability (η) of drift rate was also modeled in the same way as in Ratcliff and McKoon ($\eta = \eta_0 + \sigma_1\sqrt{N_1^2 + N_2^2}$, see Panel C in Fig. A1).

These models cannot fit the conflict conditions and so we augment these models with terms that are designed to account for the aspects of the data that the numerosity-only models fail to predict (see the discussion in Ratcliff & McKoon, 2018, p. 212). The initial models are motivated by similar principles as the numerosity-only models, but these fail to account for aspects of the data. Extensions of these models begin to become somewhat ad hoc and are not motivated by the principles on which the ANS models were developed. Of the models we have examined, the models shown in Table 3 represent the evolution of our best-fitting models. As we will see later, there are still some misses between the model prediction and data and so the modeling presented here should be seen as a progress report rather than the final word.

Models 3–6 assumed that drift rate was composed of components based on numerosity and perceptual features (i.e., dot area). Because the linear model was preferred in the B/Y task in Ratcliff and McKoon (2018), these new models were defined by adding a linear difference or a log ratio of dot areas into the drift rate equation of the linear model. Model 3 (the linear–linear model) has drift rate a function of the linear numerosity difference and the linear area difference ($A_1 - A_2$, where A_j is the dot area of blue/yellow dots). Model 4 (the linear-log model) has the linear numerosity difference and the log area difference (the log ratio of areas, $\log(A_1) - \log(A_2) = \log(A_1/A_2)$). These straightforward extensions of the ANS diffusion model assume that the dot areas have the same kind of contribution (log or linear) to drift rate as numerosity. The reason for this is that non-numeric features should behave according to Weber's and Fechner's laws as does numerosity in the ANS theory. This choice should not be taken as comprehensive, but the linear and log functions are consistent with theory in psychophysics and so provide two plausible alternatives.

In these extended models, the drift rate coefficients were not allowed to vary by conditions. Thus, the new models were more restricted than the numerosity-only models because they should capture the behavioral patterns of accuracy and RT distributions from different conditions with the same drift-rate coefficients for numerosity and area in the drift rate equation. In contrast, the numerosity-only models were allowed to have different drift rate coefficients for each condition (agree, conflict, and equal). In comparing the two new models, we speculated that the linear-log model would perform better because comparing two different dot areas (between blue and yellow dots) is similar to comparing two separate objects. That the log model was preferred in a task with two separate arrays may imply that the log representation would be more appropriate for the dot area.

However, it was expected that the linear–linear model and the linear-log model would not be able to explain the result that accuracy fell below chance in some conflict conditions. To produce this pattern, both of the drift rate coefficients, one for numerosity and the other for the area, should be positive and the coefficient for the area should have a moderately large magnitude. In this case, the models produce a positive and high drift rate in the agree condition while they produce a much smaller, or even negative drift rate in the conflict condition. This occurs because $A_1 - A_2$ or $\log(A_1/A_2)$ is negative while $N_1 - N_2$ is positive in the conflict condition. Within the conflict condition, however, the models should be able to predict large differences in accuracy induced by the numerosity values. Specifically, the accuracy of the (15, 10) pair in the conflict condition was 0.659 which would be produced from a positive drift rate, whereas the accuracy of the (40, 35) pair was 0.379 which would be produced from a negative drift rate. Because $N_1 - N_2 = 5$ in both pairs and the area difference remains the same within the conflict condition, Models 3 and 4 predict the same drift rate for any pairs in the conflict condition. This implies that the models should predict the large difference between the (15, 10) and (40, 35) pairs only by the across-trial variability of the drift rate, but the effect of across trial variability in drift rate was too small to capture this difference in accuracy values.

An alternative way to model the drift rate is to take the log ratio of the numerosity pairs. The log representation of the numerosity might be able to improve the model's performance because $\log\left(\frac{15}{10}\right)$, $\log\left(\frac{25}{20}\right)$, and $\log\left(\frac{40}{35}\right)$ are about 0.405, 0.223, and 0.134, respectively. Thus, the log model of numerosity can produce different drift rates within the conflict condition even when the numerosity pairs have the same linear difference. Therefore, we also tested the log-linear model and the log–log model, which have the log ratio of the numerosities in common, and the linear difference or the log ratio of the dot areas, respectively (Models 5 and 6 in Table 3).

A better way to model the behavioral pattern we observed, especially in the conflict condition, is to consider an interaction effect of the numerosity pairs and the dot areas. The intuition is that, given the constant difference in two numerosities, the signal from numerosity information is smaller when there is a large number of dots than when there are fewer dots. Thus, when the total number of dots in a display is large, judgments can be more affected by non-numeric perceptual features in stimuli. The interaction term needs to produce a small effect of the area when there are a few dots, but a large effect when there is a large number of dots. To accomplish this, we defined interaction models by adding an interaction between the area effect and the total sum of numerosities to the linear model of numerosity (Models 7 and 8 in Table 3). As in the previous models, both the log and the linear representations of dot areas were examined and the area components of drift rate were multiplied by the total number of dots. The models with the interaction term were expected to appropriately capture the accuracy pattern across different conditions. If drift rate coefficients of the numerosity effect and the interaction effect are both positive, the model would be able to predict lower accuracy in the conflict condition because the interaction term becomes negative and it is large enough so the whole drift rate becomes negative. For example, for the numerosity pair (40, 35) in the agree condition, the linear–log interaction model (Model 8) predicts the drift rate to be $v = v_1(40 - 35) + v_2\left(\log\left(\frac{14^2\pi}{8^2\pi}\right)\right) * (40 + 35) = 5v_1 + 83.942v_2$. In contrast, in the conflict condition, the drift rate would be

$v = v_1(40 - 35) + v_2\left(\log\left(\frac{8^2\pi}{14^2\pi}\right)\right) * (40 + 35) = 5v_1 - 83.942v_2$ (for Model 8 average estimates of $v_1 = 0.027$ and $v_2 = 0.003$ for Experiment 1, so the drift rates would be 0.387 for the agree condition and -0.117 for the conflict condition). This shows that the model can predict a positive drift rate in the agree condition and a negative drift rate in the conflict condition. Also, the interaction enables the models to capture the differences among the numerosity pairs within the same area condition because the interaction effect becomes larger as the total numerosity increases. For example, in the conflict condition, average areas remain the same, producing the same conflict effect over all the numerosity pairs. However, the numerosity pair (40, 35) has a larger negative contribution by the interaction term than the pair (15, 10) because $\log(A_1/A_2)$ (or $A_1 - A_2$) would be negative in the conflict condition and so $v_1(40 - 35) + v_2\log\left(\frac{A_1}{A_2}\right) * (40 + 35) < v_1(15 - 10) + v_2\log\left(\frac{A_1}{A_2}\right) * (15 + 10)$ (for Model 8 average estimates of $v_1 = 0.027$ and $v_2 = 0.003$ for Experiment 1, so the drift rates would be -0.117 for (40, 35) and 0.051 for (15, 10) in the conflict condition). This enables the model to predict lower accuracy for the (40, 35) pair unless v_2 is negative or very close to zero. In this way, it would be possible for the models with an interaction term to predict different behavioral patterns over different conditions. Along with the model introduced above, we also examined the interaction model with the log representation of numerosity (Models 9 and 10 in Table 3).

Another possible source of interaction comes from the numerosity difference and the total numerosity. As it does for the area representation, the total numerosity may also moderate the effect of the numerosity difference (Table 2; Ratcliff & McKoon, 2018). Given the same numerosity difference and the same perceptual feature magnitude, accuracy decreases as the total numerosity increases (the size effect). In the linear representation of numerosity, although N_1 and N_2 increase, the difference ($N_1 - N_2$) would be the same (e.g., 15 vs 10, 25 vs 20, and 40 vs 35), missing the decrease in accuracy. The interaction effect addresses this to some degree, but not in the equal conditions because the two areas are nearly the same and so the area effect would be canceled out (both $A_1 - A_2$ and $log(\frac{A_1}{A_2})$ are zero). This pattern is accounted for by changes in across-trial variability in Ratcliff and McKoon (2018). However, in the current experiment, because of the larger range of values of dot radii over trials, the size effect cannot be modeled by across-trial variability in drift rate alone and needs to be modeled explicitly in the equations for drift rate. Thus, we examine a model in which the interaction between the numerosity difference and the total numerosity is explicitly represented with an additional term $(N_1 - N_2)/(N_1 + N_2)$. We call this term the 'moderated numerosity difference' and it is added to the interaction model to examine whether this improves the fit to data (Model 11 in Table 3).

4.2.3. Fitting the ANS diffusion models to the data from Experiment 1

All the models proposed above were fitted to the data from Experiment 1 using the G^2 method. As correct responses for blue dots and correct responses for yellow dots were combined and produced a single measure of accuracy and correct RT distribution per condition (errors were combined in a similar manner), the two boundaries of the ANS diffusion model represent correct and error responses. Thus, the starting point parameter is fixed as $z = a/2$ in the model fits. The models and their G^2 values averaged across subjects are presented in Table 3 and the parameter estimates are in Table 4. For the numerosity-only models (Models 1 and 2), the results agreed with the findings from the previous study. First, the linear model fit the data better than the log model (G^2 were 433.4 and 451.7, respectively). This was reasonably consistent across subjects as 21 of 29 subjects favored the linear model. The area manipulation also affected the drift rate; the drift-rate coefficients were 0.036, 0.027, and 0.015 in the linear model, and 0.672, 0.522, and 0.310 in the log model, for the agree, equal, and conflict conditions, respectively. However, both models failed to capture the behavioral pattern of the results observed in the data (Figure S1 and S2 in the Supplementary Materials). Specifically, they fitted the equal conditions quite well, but 1) in the agree condition, they predicted large differences in accuracy across numerosity pairs with the same difference (e.g., 15/10, 25/20, and 40/35) while there were only small differences in the data, and, 2) in the conflict condition, they failed to predict the below-chance accuracy in the highest conflict condition. The *df* for the numerosity-only models

Table 4
Parameter estimates for models in Experiment 1. a is the boundary separation, T_{er} is the nondecision time, and s_z and s_t are the across-trial variability in starting point and nondecision time, respectively. η_0 and σ_1 are the intercept and slope of the across-trial variability in drift rate. t_0 and t_1 are the intercept and slope of the nondecision time regression used in Model 12. For the linear and log models, v_1, v_2, and v_3 are the separate drift rate coefficients for the agree, equal, and conflict conditions, respectively. For the other models, v_i's are the drift rate coefficients in the drift rate equations in Table 3. The starting point parameter z is fixed as $z = a/2$.

	Model	a	T_{er}	η_0	$10\sigma_1$	s_z	s_t	v_1	v_2	v_3	t_0	$t_1/10$
1	Linear	0.122	0.455	0.023	0.058	0.098	0.259	0.036	0.027	0.015		
2	Log	0.121	0.452	0.165	0.010	0.097	0.253	0.672	0.522	0.310		
3	Linear - Linear	0.127	0.456	0.092	0.046	0.101	0.251	0.027	0.001			
4	Linear - Log	0.112	0.432	0.031	0.034	0.050	0.246	0.019	0.092			
5	Log - Linear	0.112	0.432	0.127	0.003	0.050	0.245	0.384	0.001			
6	Log - Log	0.127	0.451	0.189	0.016	0.102	0.253	0.536	0.003			
7	Linear - Linear Interaction	0.126	0.452	0.051	0.056	0.101	0.254	0.027	0.001			
8	Linear - Log Interaction	0.126	0.453	0.048	0.057	0.101	0.255	0.027	0.003			
9	Log - Linear Interaction	0.127	0.452	0.186	0.016	0.101	0.255	0.536	0.001			
10	Log - Log Interaction	0.121	0.441	0.155	0.015	0.077	0.245	0.472	0.002			
11	Linear - Log Interaction + Moderated Numerosity	0.115	0.431	0.063	0.029	0.051	0.244	0.011	0.002	0.398		
12	Linear - Log Interaction + Moderated Numerosity + T_{er} model	0.125		0.110	0.034	0.095	0.250	0.013	0.003	0.542	0.450	−0.540

was 222: the total number of conditions was 21 from the three area conditions (agree, equal, and conflict) and seven numerosity pairs and this was multiplied by 11 *df* from the proportions of correct and error responses from the RT bins (12 proportions minus 1 since the proportions should sum to 1). Then the number of parameters (9) was subtracted.

Next, we examined the extended ANS diffusion model which included the area effect in the drift rate equation. The first thing to note is that the models without an interaction effect (Models 3–6) failed to explain the accuracy pattern. They cannot predict high accuracy in the agree and equal conditions and accuracy lower than 0.5 in the conflict condition. Unlike the numerosity-only models, when area was included in the drift rate model, the log models (Models 5 and 6) performed better than the linear models (Models 3 and 4; $G^2 = 386.4$ and 390.1 for the log models and $G^2 = 400.1$ and 404.7 for the linear models). This was consistent with what we expected from the behavior of the models: the log representation can produce differences in drift rate for a constant difference in numerosity while the linear model cannot produce large enough differences. However, even the log models failed to fit below chance accuracy in the conflict condition.

In contrast, the models with the interaction effect (Models 7–10) produced reasonably good fits, capturing benchmarks in the data. Fig. 3 shows the quantile-probability plot of the linear-log interaction model (Model 8), which was the best among the interaction models. This plot fully describes the observed and the predicted accuracy and RT distributions for correct and error responses across different conditions. In the plot, x's indicate the observed data and circles ('o') represent the predictions. The proportions of correct and error responses are plotted on the x-axis and the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles of the RT distribution are plotted vertically on the y-axis. Predictions were generated from all the numerosity pairs used in the experiment. The pairs corresponding to each panel are shown on the top-right side of the panel.

To produce model predictions for the agree, large, small, and conflict conditions, we group data based on the radii or areas of dots for the conditions. We grouped data based on the smaller and the larger dot radii: the smaller were 6, 8, 10 with a mean of 8 and the larger were 12, 14, and 16, with a mean of 14. We used these means to produce the predictions. Specifically, we used 14/8 for dot radii of large/small numerosities in the agree condition. Similarly, we used 8/14, 14/14, and 8/8 for dot radii of large/small numerosities in the conflict, large, and small conditions, respectively.

In the majority of perceptual decision-making tasks, accuracy is higher than 0.5. Therefore, in quantile probability plots, points on the right-hand side of 0.5 on the x-axis usually represent correct responses and points on the left-hand side represent error responses. In the conflict condition, however, accuracy was sometimes less than chance. This produced a crossover pattern in the conflict condition (the x's and dotted lines cross over the 0.5 point in the conflict condition in Fig. 3). Note that this matched the values reported in Table 2.

In general, the linear-log interaction model (Model 8) was able to explain the accuracy and RT patterns for the different conditions, especially the lower than chance accuracy in the conflict condition. It also explained high accuracy with a little difference across numerosity pairs in the agree condition. It should be noted that this was possible only when the interaction term was included in the drift rate equation. Therefore, these results support the idea that area may dominate numerosity when there is a large conflict between the variables and the task is difficult.

There were misfits in a few of the conditions. Specifically, the model tended to predict accuracy much lower than the observed proportion of correct responses for some numerosity pairs in the conflict and equal conditions. Also, the model missed some of the RT patterns. For example, the observed RTs tended to be shorter for the (20, 10) pair than for the (40, 30) pair for the conflict condition, but the model predicted that RTs would be only slightly shorter for the (40, 30) pair. Error RTs were captured well in most of the conditions except for the misses at the leading edges and tails in the equal conditions. Despite these misses, the model predictions generally matched the observed patterns with a high degree of consistency and provided support for the interaction model.

Another interaction model, the linear–linear interaction model (Model 7), differed to only a small degree from the linear-log interaction model both quantitatively ($G^2 = 381.1$) and qualitatively (Figure S3 in the Supplementary Materials). This implies that an important factor in numerosity judgments is the interactive relationship between the numerosity and the perceptual variable, rather than the specific form of the perceptual dot area representation. Once the interaction term was added, there was little difference between the log and the linear representations of the dot area. We also examined the log-linear interaction model (Model 9) and the log–log interaction model (Model 10). However, both of them performed worse than the interaction models with the linear difference of numerosities (G^2=386.9 and 384.2, respectively). Both of the models predicted accuracy much lower than the data when there are many dots (e.g., 40 vs 30 and 40 vs 20) in the conflict and the equal conditions (Figure S4 and S5).

For Models 3–10, the number of *df* was 300. The data were divided by four conditions (agree, conflict, large, and small) and seven numerosity pairs which yields 28 combinations and this was multiplied by 11 *df* of response proportions from RT bins. Then the number of parameters, 8, was subtracted.

Because the linear-log interaction model (Model 8) performed quantitatively and qualitatively the best, we added the moderated numerosity difference $(N_1 - N_2)/(N_1 + N_2)$ to this model (Model 11). Adding this single term produced a noticeable improvement in model fit ($\Delta G^2 = 22.352$). The quantile-probability plot of this model is shown in Fig. 4. The general pattern of the prediction was similar to that of the linear-log interaction model. However, the new model generally predicted higher accuracy for the numerosity pairs in the conflict and equal conditions in which the previous model predicted accuracy much lower than the data. This improved the fit to accuracy data. The model also improved fits for RT distributions. In the conflict and equal conditions, the model had some misfits in the leading edges for correct and error RTs. This new conflict effect is addressed in the next section. Error RT distributions were also fit quite well except for some misses at the leading edges and tails in the large and small conditions. Note that this model fits all the patterns from the 28 different conditions with only 9 parameters (with both correct and error RT distributions represented by 5 quantiles, this gives 299 degrees of freedom). This provides a compact account of a complicated interaction between numerosity and perceptual features.

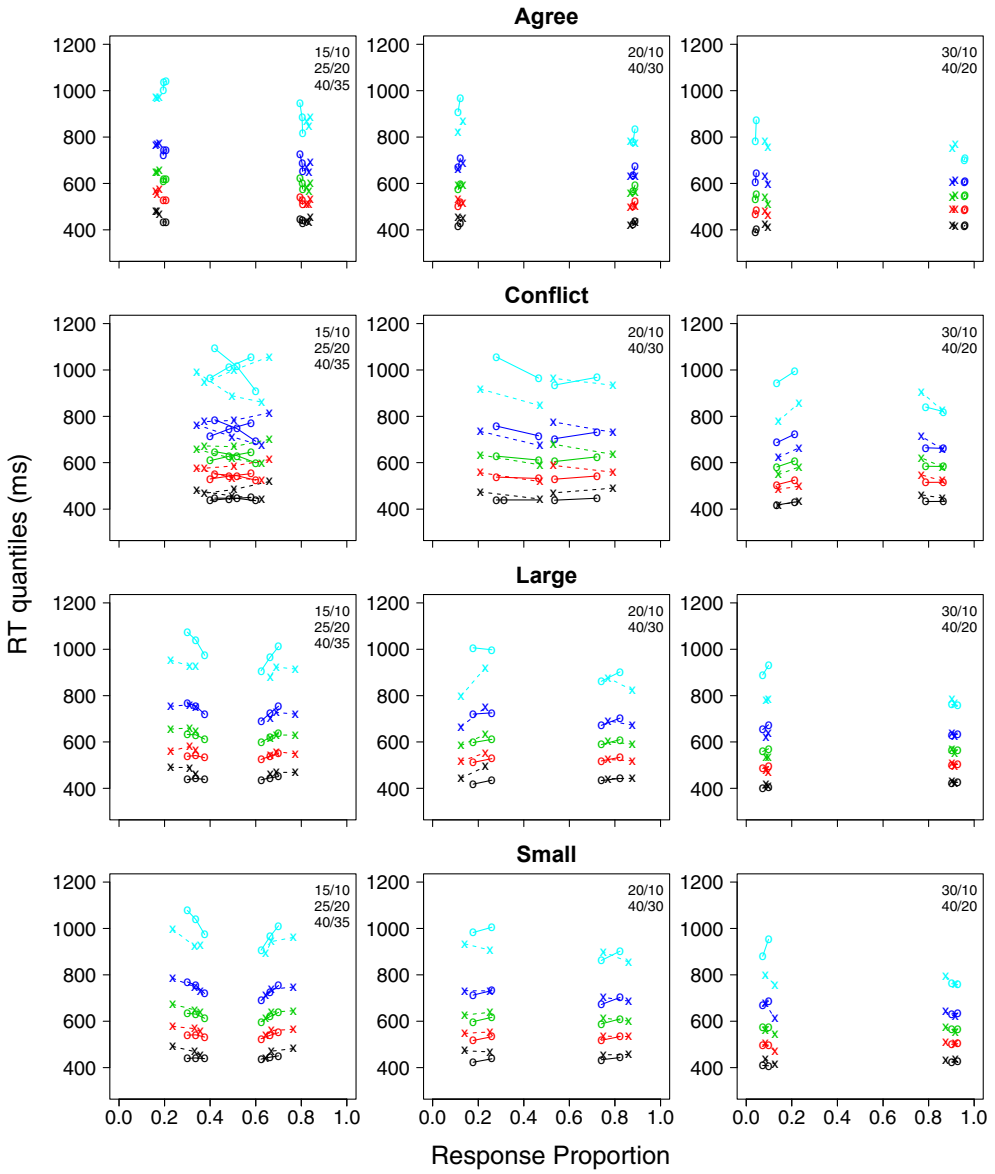

Fig. 3. Quantile-probability plot for Experiment 1 (Model 8 in Table 3: the linear-log interaction model). In the plot, RT quantiles are plotted against response proportions. The four conditions (agree, conflict, large, and small) are by dot radii: the agree condition has dot radius larger than its average for the larger number of dots and smaller for the smaller number of dots. The conflict condition has dot radius smaller than its average for the larger number of dots and larger for the smaller number of dots. The large (small) condition had both dot radii larger (smaller) than its average regardless of the number of dots. The numerosity pairs are presented in the top right corner of each panel. Typically, correct responses are on the right and error responses are on the left. However, in the conflict conditions, sometimes accuracy decreased to less than chance which means correct responses were plotted on the left (the lines joining the conditions show the cross-over in the conflict condition).

4.2.4. Estimating relative contributions of variables

Parameter estimates of the models are presented in Table 4. Among the estimates, v_1, v_2, and v_3 are slope coefficients in the drift rate models: v_1 is a coefficient for the numerosity difference effect, v_2 is for the interaction term, and v_3 for the moderated numerosity difference. Each of these estimates indicates the amount of change in the drift rate by a unit change in the covariate multiplied to that estimate. Given these estimates, the drift rate and the across-trial variability in drift rate can be estimated using the regression equations. For the linear-log interaction model with the moderated numerosity difference (Model 11), the drift rate and its variance can be computed as follows.

$$v = 0.011 * (N_1 - N_2) + 0.002 * (\log(A_1/A_2)) * (N_1 + N_2) + 0.398 * (N_1 - N_2)/(N_1 + N_2)$$

$$\eta = 0.063 + 0.0029\sqrt{N_1^2 + N_2^2}$$

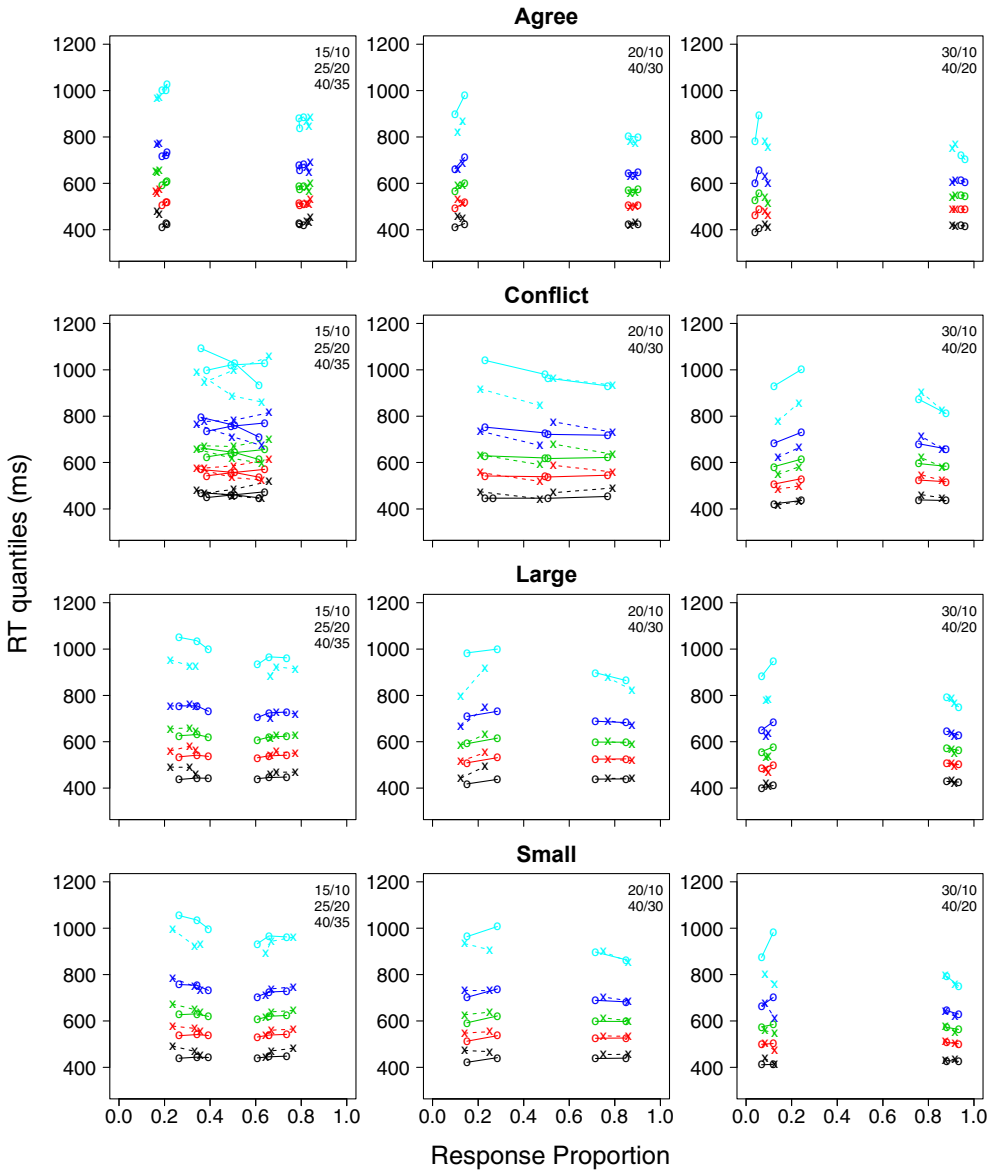

Fig. 4. Quantile-probability plot for Experiment 1 (Model 11 in Table 3: the linear-log interaction model with the moderated numerosity difference effect).

Parameter estimates can also be used to separately estimate and compare the effects from the numerosity and the other terms on the drift rate and the across-trial variability. In Ratcliff and McKoon (2018), the effects of different area conditions were examined by comparing their drift rate coefficients. Their finding that the coefficient was much larger in the area-proportional condition compared to the area-equal condition clearly showed how much dot area affected numerosity judgments. In our modeling, the number of dots and the dot areas have different scales, and so a comparison between the drift rate coefficients does not show the relative sizes of the contributions. To examine the contributions, the range of each independent variable multiplied by the corresponding drift rate coefficient (e.g., $v_1 * (N_1 - N_2)$) was computed and shown in Table 5.

In the B/Y task, the current model predicted the drift rate range of (-0.242, 0.610) and the across-trial variability range of (0.116, 0.219). These are in the typical range of values in perceptual decision-making tasks (Ratcliff, 2014). For the components of the drift rate equation, the linear numerosity difference term ranged between (0.055, 0.218), the log-linear interaction term ranged between (-0.323, 0.323), and the moderated numerosity difference term ranged between (0.027, 0.200). Overall, it can be concluded that, although the numerosity had a major effect on numerosity judgments, the interaction between the dot area and the total number of dots also had a large effect. In the conflict condition, this made the non-numeric perceptual variable dominate judgments by producing a negative drift rate. Although the value of the drift rate coefficient was the largest for the moderated numerosity term, its estimated effect was relatively small. This shows why it is necessary to examine the ranges of the terms to understand the relative

Table 5
Ranges of drift rate, across-trial variability of drift rate, and their components calculated from the linear-log interaction model with moderated numerosity effect (Model 11). N_i's are the numbers of dots and A_i's are dot areas. v_i's are the drift rate coefficients and η_0 and σ_1 are the intercept and slope of the across-trial variability model.

Terms	Drift Rate (v)	Numerosity $v_1(N_1 - N_2)$	Area Interaction $v_2\left(\log\left(\frac{A_1}{A_2}\right)\right)(N_1 + N_2)$	Moderated Numerosity $v_3\frac{N_1 - N_2}{N_1 + N_2}$	Across-trial Variability (η)	η_0	$\sigma_1\sqrt{N_1^2 + N_2^2}$
Range	(−0.242, 0.610)	(0.055, 0.218)	(−0.323, 0.323)	(0.027, 0.200)	(0.116, 0.219)	0.063	(0.053, 0.156)

contributions of the variables to the numerosity judgments.

4.2.5. Another effect of conflict: A shift in the leading edge of the RT distribution for correct responses

The models examined in the previous analysis all assumed that the conflict between numerosity and perceptual variable affected only the quality of information processing during the numerosity judgment task. Although the best-fitting model (Model 11) was able to capture most of the behavioral patterns of accuracy and RT, there were some misfits in the RT distributions. In particular, the prediction for the leading edge of the RT distribution did not match the data in the conflict condition. There was a large shift induced by the conflict between numerosity and dot area. This is another unexpected pattern, which has been obtained occasionally in conflict tasks such as Stroop and Eriksen-Flanker tasks. Usually in most perceptual tasks, the leading edge of the RT distribution (e.g., the 10% quantile) does not change much over conditions that vary in difficulty compared to the tail of the RT distribution. The diffusion model with only drift rate changing over conditions can capture this pattern well (e.g., the leading edge and the tail typically change in a ratio of about 1:4 with a change in drift rate, cf. Ratcliff & McKoon, 2008).

To illustrate the miss between theory and data in the leading edge, Fig. 5 shows accuracy, the leading edge (10% quantile), and the median RT (50% quantile) against the four conditions used in model fitting. Because the different conditions represent different degrees of conflict, this plot shows RT and accuracy by different levels of conflict and different numerosity pairs.

The top three panels in Fig. 5 show the data (the dashed lines) and the corresponding predictions (the solid lines) from the interaction model with the moderated numerosity difference (Model 11). As shown in Table 2, accuracy decreased as the degree of conflict increased. Also, the median RT was longer when the degree of conflict was larger. Importantly, 10% quantile RTs were also affected by the conflict. Averaging across the five numerosity pairs presented in Fig. 5, the leading edge was about 51 ms longer in the conflict condition than in the agree condition. Unlike the effect of conflict on accuracy, its effect on the leading edge was larger when

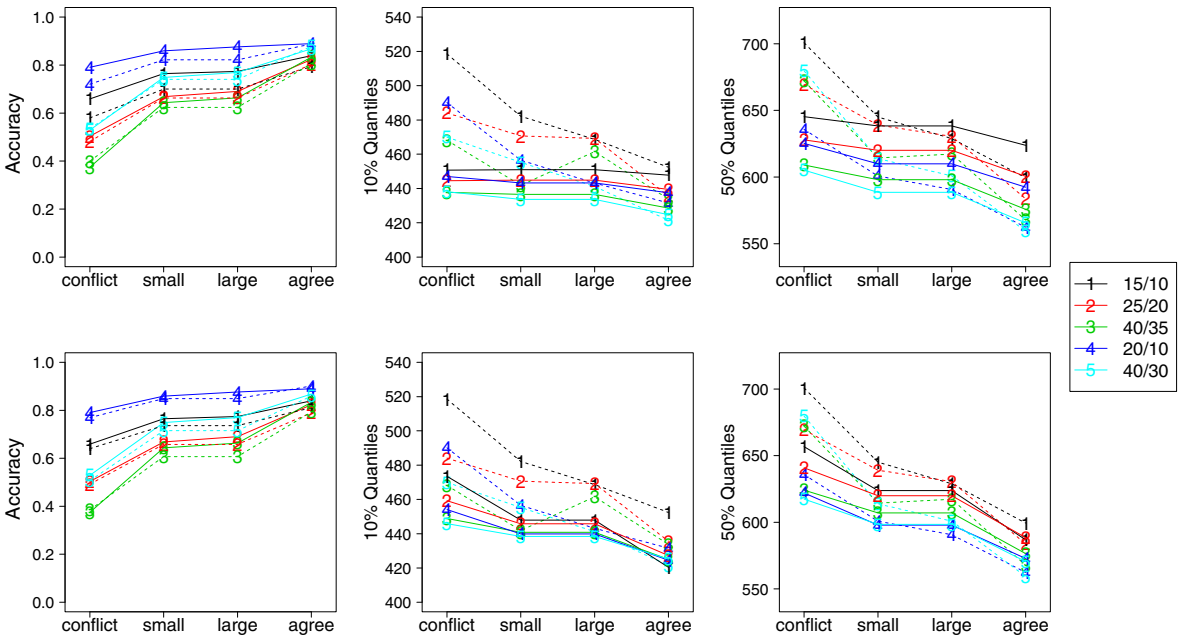

Fig. 5. Accuracy, 10% quantiles, and 50% quantile RTs plotted against the conditions with different degrees of conflict for Experiment 1. Dashed lines represent the observed result while solid lines represent predictions. The top three panels show the predictions from the linear-log interaction model with the moderated numerosity difference (Model 11 in Table 3). The bottom three panels show the predictions from the same drift rate model but with the nondecision time regression model (Model 12). For visual clarity, the plots show only five numerosity pairs. In the other two pairs in which the numerosity difference is large (Δ = 20), the effect of conflict was smaller.

there were fewer dots: the increasing pattern of the 10% quantiles was larger for numerosity pairs (15, 10), (25, 20) and (20, 10) than for (40, 35) and (40, 30). It is important to note that this cannot be produced by a simple probability mixture of processes. If, for example, on half the trials, subjects noticed a conflict and delayed their response by encoding the stimulus again, then the new 10% quantile RT would be at the old 20% quantile which is not enough change to account for the results in Fig. 5. Such a simple strategy would also produce a much larger 0.9 quantile RT, longer than would be consistent with the data.

The current model (Model 11) was not able to capture this shift at the leading edge. The model prediction for the leading edge was about 450 ms for all the conditions and predictions missed the increasing pattern in the data. This miss was about 30–40 ms in the conflict condition and this miss was consistent across all the models in the comparison above. The interaction effect and the moderated numerosity difference in the drift rate equation improved the model fit by capturing most of the accuracy and RT patterns, but it was not able to provide a full description of the RT leading edge in the conflict effect.

It is possible that modeling changes in other parameters of the diffusion model with conflict, such as the nondecision time (T_{er}) or the boundary separation (a) may help the model to explain the pattern observed for the leading edge of the RT distribution (Ratcliff & Frank, 2012; Ratcliff & Smith, 2010). Following these approaches, we examined the effect of allowing the conflict to affect nondecision time, for example, under high conflict, subjects may require more time to encode the stimulus. We examined several models of nondecision time and boundary separation. These models were built with the effects used in the drift rate model. For example, the nondecision time can be modeled by equations such as $T_{er} = t_0 + t_1(A_1 - A_2)$ or $T_{er} = t_0 + t_1(N_1 - N_2) + t_2\left(\log\frac{A_1}{A_2}\right) * (N_1 + N_2)$. If the slope coefficient of the area difference is negative and has a moderately large magnitude, the T_{er} models can predict that subjects take more time for nondecision processes when they are presented with a stimulus with conflict. Thus, the models might be able to reduce the vertical misfit in Fig. 5. However, none of these simple models improved the model fit. The estimate of the area difference was too small to reduce the mismatch between theory and data and thus the models still predicted the almost flat 10% quantile RTs across the conditions. The value of the intercept was almost equal to the T_{er} estimates in the previous models and thus the vertical misfit in Fig. 5 was not reduced at all. We also attempted to model boundary separation in the same way but were not able to find a model that improved predictions.

The result above implies that the way in which the conflict affects the RT distribution, especially the leading edge, might be different from the way in which the conflict affects the drift rate. Therefore we explored other possible expressions for the effect of conflict on nondecision time. Among the models we examined, the regression equation $T_{er} = t_0 + t_1\frac{\log(A_1) - \log(A_2)}{(N_1 + N_2) * (N_1 - N_2)}$ gave the best description of the RT pattern under the conflict. The covariate term in the equation represents the area effect moderated by the total number of dots and the numerosity difference. The bottom three panels in Fig. 5 show the predictions from this model (Model 12 in Table 3). Unlike the previous model, adding this nondecision time model provided some improvement by capturing the increasing pattern in the 10% quantiles.

The intercept and the slope of the nondecision time model were $t_0 = 0.450$ and $t_1 = -0.054$. With these parameter estimates and the equations above, we can calculate the range of T_{er}. The second term of the equation had a range of (-0.085, 0.085), which made T_{er} vary between (0.365, 0.534) with 80% of the values falling in the range (0.430, 0.470). The difference across conditions produced by the nondecision time regression allowed the model to capture the longer leading edge in the presence of the conflict. Although the negative slope coefficient produced some improvement, it did not reduce the G^2 value that much (ΔG^2=5.6). Also, even with the nondecision time regression, there still was a vertical misfit showing that the predicted 10% quantiles were still shorter than the observed RT leading edge across the conditions. However, without modeling nondecision time, the leading edge prediction is flat across conditions, which is inconsistent with the observed behavior. Although the expression ($\frac{\log(A_1) - \log(A_2)}{(N_1 + N_2) * (N_1 - N_2)}$) has a limited contribution, this analysis shows that some part of the regularity in the RT leading edge can be explained by modeling nondecision time with values of numerical and perceptual variables.

Taken altogether, the last variant of the ANS diffusion model provided the best explanation of the numerosity judgment data that we could find. Initially, we believed that we could find a principled model based on the modeling approach proposed in the previous study, but we ended up with a model that has several somewhat ad hoc assumptions (assumptions that are justified in terms of the behavior they have in explaining data). Despite these somewhat ad hoc assumptions, the model was able to capture most of the patterns from a large number of conditions with only 10 parameters. Considering this parsimony of the model, the result indicates that the ANS diffusion model provides a promising framework to study how we make a choice in numerosity judgment tasks.

One of the reviewers asked whether extreme flexibility in drift rate can account for the reaction time pattern induced by the conflict without the nondecision time regression. We examined this possibility by fitting a model with a different independent drift rate for each condition (i.e., 28 drift rates for 28 conditions). The model fit the accuracy pattern well but not the increasing leading edge of the RT distribution (G^2 = 354.6; Figure S6 and S7). Thus, the conflict effect on the RT leading edge cannot be explained by the drift rate alone. Also, Model 11 (the linear – log interaction model with the moderated numerosity difference, but without the nondecision time modeling) performed as well (G^2 = 358.7; Fig. 4 and Fig. 5) as this independent drift rate model, but with only three drift rate coefficients instead of the 28 separate drift rates. Of course, the 28 drift rate parameters are based on our grouping of the data and a model of all the combinations of variables would require 256 drift rates.

The inability of the standard model to fully explain the behavior of the leading edge of the RT distributions points to an important result from this research. This is the finding of a new conflict effect that occurs in a numerosity judgment task when perceptual variables conflict with numerosity. This is a new conflict effect that standard models cannot explain.

We now examine whether the same modeling approach can be applied to the other experiments. We fitted the same (or similar) ANS diffusion models to the data from the other three experiments. In Experiment 2, the model was generalized to a different type of numerosity judgment task (the L/R task, Fig. 2). In this task, convex hull was the perceptual variable of main interest (it had a much

larger effect than the area variable). Thus, drift rate and nondecision time were modeled as a function of the numerosity pairs and the convex hull areas. In Experiment 3, we fitted the models to the B/Y task, but in this task, the stimuli remained on the screen until subjects made their decision. This was to examine whether the conflict effect, especially the slowdown in the RT leading edge, was due to the short presentation time. If so, the shift in the leading edge could be the result of limited viewing time, but if not, the large shift can be attributed to the conflict between numerical and perceptual variables.

4.3. Experiment 2

In Experiment 2, subjects were presented with two side-by-side arrays of dots and asked to decide if there were more dots on the left side or on the right side (i.e., the L/R task). Correct responses for the left and the right side were aggregated and error responses for the left and the right side were aggregated as in Experiment 1. In the L/R task, we manipulated numerosity pairs, dot radii, and convex hull radii. For the numerosity conditions, we used the same pairs as Experiment 1. The dot radii took 6 different values from 3 to 8 pixels while the convex hull radii took 16 different values from 85 to 160 pixels with a difference of 5. This produced a much larger number of combinations (64512 = 7*6*6*16*16) than in the B/Y task (252). As in Experiment 1, we grouped conditions over perceptual features of a stimulus and computed RT quantiles for these conditions. Because we manipulated two different perceptual features, the conditions can be defined for either or both of them. In Experiment 2, we focused on the convex hull variable because it had a larger effect on performance than did dot area. Unlike what was found in Experiment 1, dot areas had a relatively small contribution to judgments in the L/R task, compared to numerosity pairs and convex hull areas. Also, if we defined conditions by both of the two perceptual features, this produced too many conditions (e.g., 4*4 multiplied by 7 numerosity conditions if grouped as for Experiment 1) and made the average sample size for each condition too small to produce RT distributions. Therefore, the conditions in Experiment 2 were defined only by the convex hull for the initial analyses. However, joint analysis of dot areas and convex hulls can be done more easily by MLE which does not require binned data. The result from MLE will be discussed later.

In the 'Agree' condition, the convex hull radius was larger than its mean (122.5) in the array with more dots and it was smaller than the mean in the array with fewer dots. In the 'Conflict' condition, the convex hull radius was smaller than the mean in the array with more dots while it was larger in the array with fewer dots. In the 'Large' and 'Small' conditions, the two arrays had similar convex hulls, both large or both small, regardless of the numerosities. With 7 numerosity pairs, these conditions yielded 28 conditions. The dot areas in each condition were uniformly distributed and their effect was averaged out.

4.3.1. Results

Accuracy and mean RTs for correct responses presented in Table 6 confirmed the patterns we found in Experiment 1. First of all, the behavioral pattern differed over conditions defined by the convex hull. In the agree condition, accuracy was very high, around 0.8–1.0, and mean RT was shortest. In the conflict condition, accuracy was low and mean RT was longest. Accuracy and RT in the equal (the large and small) conditions were between those in the agree and conflict conditions. This agrees with the general findings in Ratcliff and McKoon (2018) and Experiment 1 that show that perceptual variables affect numerosity judgments; the same occurs in the L/R task. Dot radius or dot area had a much smaller effect on accuracy and RT than the convex hull (Table 6) and the effect was somewhat inconsistent on RT. Thus, convex hull is a more important perceptual variable in this task and we used only convex hull rather than dot radius in the following analysis. In a later analysis using the MLE method, both these variables along with numerosity were included in modeling.

Even though the general pattern over conditions was the same as in Experiment 1, the size of the effects of the convex hull perceptual variable was different. The mean RT difference was smaller in the L/R task. Specifically, the mean RT difference between the agree and conflict conditions was about 60 ms (607 ms and 668 ms) in Experiment 1 while it was about 40 ms (504 ms and 545 ms) in Experiment 2. In contrast, the difference in mean accuracy was much larger in Experiment 2 than Experiment 1. In Experiment 1, mean accuracy values in the agree and conflict conditions were 0.867 and 0.640, respectively, and the difference was about 0.227. In Experiment 2, mean accuracy values were 0.908 and 0.589 in those conditions, so the difference was 0.319, which was about 30% larger than in Experiment 1. Therefore, on average, the L/R task was much easier than the B/Y task in the agree condition, but it was more difficult in the conflict condition.

Table 7 shows accuracy values and mean RTs for correct responses as a function of numerosity pairs and the three conditions defined by convex hull (the large and small conditions are combined in the equal condition). The accuracy values showed a similar pattern as obtained in Experiment 1. Generally, accuracy increased as the numerosity difference increased. For a constant numerosity difference, accuracy decreased as the total number of dots increased. In the conflict condition, numerosity judgments were dominated

Table 6
Experiment 2 (L/R task): Accuracy and correct mean RT. Data are divided into different conditions by values of the convex hull or the dot radius.

Experiment and task		By Convex Hull		By Dot Radius	
	Conditions	Accuracy	Mean RT	Accuracy	Mean RT
2, L/R	Agree	0.908	504	0.808	525
	Large	0.794	529	0.788	520
	Small	0.777	534	0.760	525
	Conflict	0.589	545	0.722	532

Table 7
Experiment 2: Accuracy and mean RTs for correct responses as a function of numerosity pairs and the three area conditions (Agree, Equal, and Conflict).

Expt, Task	Measure	Condition	Δ = 5			Δ = 10		Δ = 20	
			(15, 10)	(25, 20)	(40, 35)	(20, 10)	(40, 30)	(30, 10)	(40, 20)
2, L/R	Accuracy	Agree	0.904	0.861	0.827	0.955	0.892	0.973	0.950
		Equal	0.792	0.669	0.603	0.906	0.698	0.950	0.883
		Conflict	0.624	**0.363**	**0.313**	0.787	**0.418**	0.895	0.723
	Mean RT	Agree	511	511	511	502	501	484	496
		Equal	541	542	543	522	532	496	512
		Conflict	566	558	551	548	557	512	539

by the convex hull variable for some numerosity pairs (the bolded entries in Table 7). When the task was difficult, the subjects relied more on the convex hull to make a judgment. For example, when the difference between two numerosities was small (Δ = 5 or 10) and there was a large number of dots in a stimulus (e.g., 40 vs 35), accuracy was lower than chance. For the numerosity pair (40, 35) under the conflict condition, accuracy was 0.313. This was consistent across subjects: only 2 of 34 subjects chose the correct answer on more than half of the trials (their accuracy was 0.593 and 0.500 and the others had accuracy that ranged between 0.122 and 0.439). As before, this provides strong support that numerosity decisions can be dominated by perceptual variables.

Mean RTs for correct responses in Table 7 decreased as numerosity difference increased. For a constant numerosity difference, mean RTs varied a little or increased as a function of the total numerosity in the agree and conflict conditions, which is consistent with the results of the log representation of numerosity fitted to the L/R task in Ratcliff and McKoon (2018). However, in the conflict condition, a different pattern was obtained: as the total number of dots increased, mean RTs decreased when the numerosity difference was 5, but increased when the numerosity difference was 10 or 20. This pattern was more similar to the pattern obtained in Experiment 1.

4.3.2. ANS diffusion models for Experiment 2

Several versions of the integrated diffusion models were derived from the models used in Experiment 1 and fitted to the data from Experiment 2. There were two main differences. The first one was that we used convex hull areas instead of dot areas and so the models represent how numerosity and convex hull areas interact in numerosity discrimination. The second difference was that the log representation of numerosity was used in these models for Experiment 2. This was based on the results from modeling in Ratcliff and McKoon (2018) that showed that the log model fit data better in the L/R task.

The equations for the models we examined are presented in Table 8. The first model was the log–log interaction model. This model was similar to the linear-log interaction model in Experiment 1. The model had terms representing the main effect of the log numerosity difference and the interaction effect of the convex hull and the total number of dots. The second model was constructed by adding the moderated numerosity difference to the first model and this term used the log ratio of numerosities for the numerator and the total number of dots for the denominator. The third model had the nondecision time equation which is the same as the model used in the previous analysis. In addition, another pair of models was constructed by replacing the total number of dots by the sum of the log-transformed numerosities, i.e., $\log(N_1) + \log(N_2)$ (Model 2–1 and 3–1 in Table 8). These modified models were examined to see if the log representation can explain the effect of the total numerosities better than the sum of the raw numerosities.

We also examined other models such as models with the linear representation of numerosity and models without interaction effect but we do not report the results. They performed more poorly than those with the log representation and we did not investigate them further.

Table 8
Models used to fit data from Experiment 2, their G^2 values, and degrees of freedom (*df*). N_1 and N_2 are two numerosities and C_1 and C_2 are two convex hull areas. Convex hull area is computed as $C_i = \pi * s_i^2$ where s_i is a convex hull radius. For all of the models, across-trial variability of drift rate was modeled by $\eta = \eta_0 + \sigma_1\sqrt{N_1^2 + N_2^2}$.

Model			G^2	*df*
1	Log - Log Interaction	$v_1\left(\log(\frac{N_1}{N_2})\right) + v_2\left(\log\left(\frac{C_1}{C_2}\right)\right)(N_1 + N_2)$	342.6	300
2	Log - Log Interaction + Moderated Numerosity	$v_1\left(\log(\frac{N_1}{N_2})\right) + v_2\left(\log\left(\frac{C_1}{C_2}\right)\right)(N_1 + N_2) + v_3\left(\log(\frac{N_1}{N_2})\right)/(N_1 + N_2)$	340.4	297
2–1		$v_1\left(\log\left(\frac{N_1}{N_2}\right)\right) + v_2\left(\log\left(\frac{C_1}{C_2}\right)\right)(\log(N_1) + \log(N_2)) + v_3\left(\log\left(\frac{N_1}{N_2}\right)\right)/(\log(N_1) + \log(N_2))$	342.8	
3	Log - Log Interaction + Moderated Numerosity + T_{er} Model	Model 2 for the drift rate with $T_{er} = t_0 + t_1\frac{\log(C_1) - \log(C_2)}{(N_1 + N_2) * (N_1 - N_2)}$	335.0	298
3–1		Model 2–1 for the drift rate with $T_{er} = t_0 + t_1\frac{\log(C_1) - \log(C_2)}{(\log(N_1) + \log(N_2)) * (N_1 - N_2)}$	338.4	

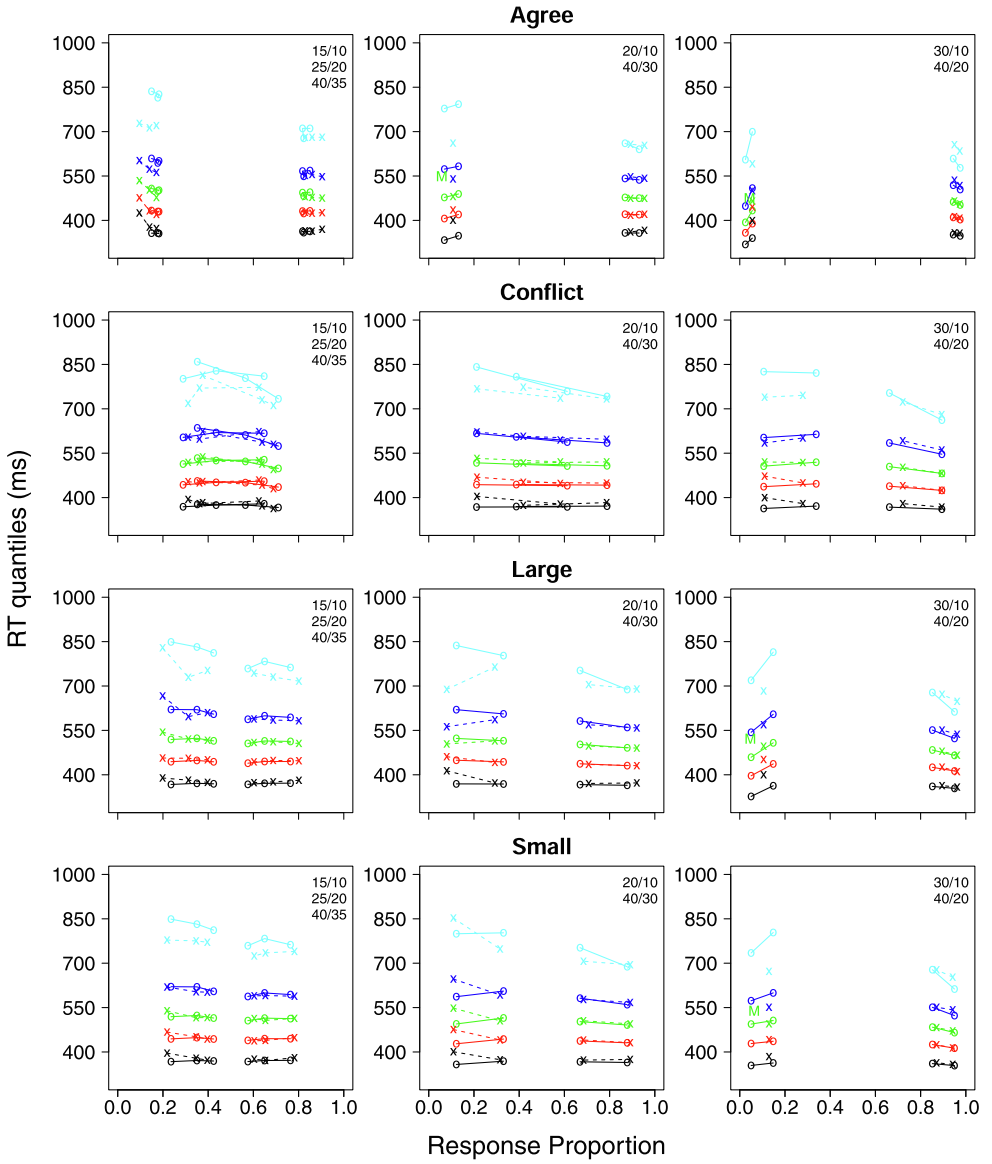

Fig. 6. Quantile-probability plot for Experiment 2 (Model 3 in Table 8: the log–log interaction model with the moderated numerosity difference effect and the nondecision time regression).

4.3.3. Fitting the ANS diffusion model to Experiment 2 data

All the models were fitted using the G^2 method as in Experiment 1 and Table 8 shows the G^2 values for the models. The models with the sum of log-transformed numerosities (Models 2-1 and 3-1) did not fit better than the models with the raw total numerosity (Models 2 and 3) and so were not considered further. As in Experiment 1, the interaction model with the moderated numerosity difference and the nondecision time equation (Model 3) produced the best fit in terms of G^2. However, the difference between the models was small which suggests that the two terms added to the log–log interaction model are probably not necessary. Accuracy was captured well by the interaction effect alone. Also, there was a smaller shift of the RT leading edge in the conflict condition than in Experiment 1. These two features made the moderated numerosity difference and the representation model of the nondecision time component produce only a modest contribution to goodness of fit.

The quantile-probability plot for the log–log interaction model with the moderated numerosity difference and the nondecision time regression (Model 3 in Table 8) is shown in Fig. 6. Predictions were obtained in the same way as in Experiment 1, except that the convex hull was used as the major perceptual variable in this experiment. In general, the model captured behavioral patterns of the observed data well. The model predicted the highest accuracy in the agree condition, with only small differences between the numerosity pairs with a constant numerosity difference. Also, the predicted reaction time was the shortest in this condition. In the conflict condition, the model predicted the lowest accuracy, even lower than chance in some of the conflict conditions, and also

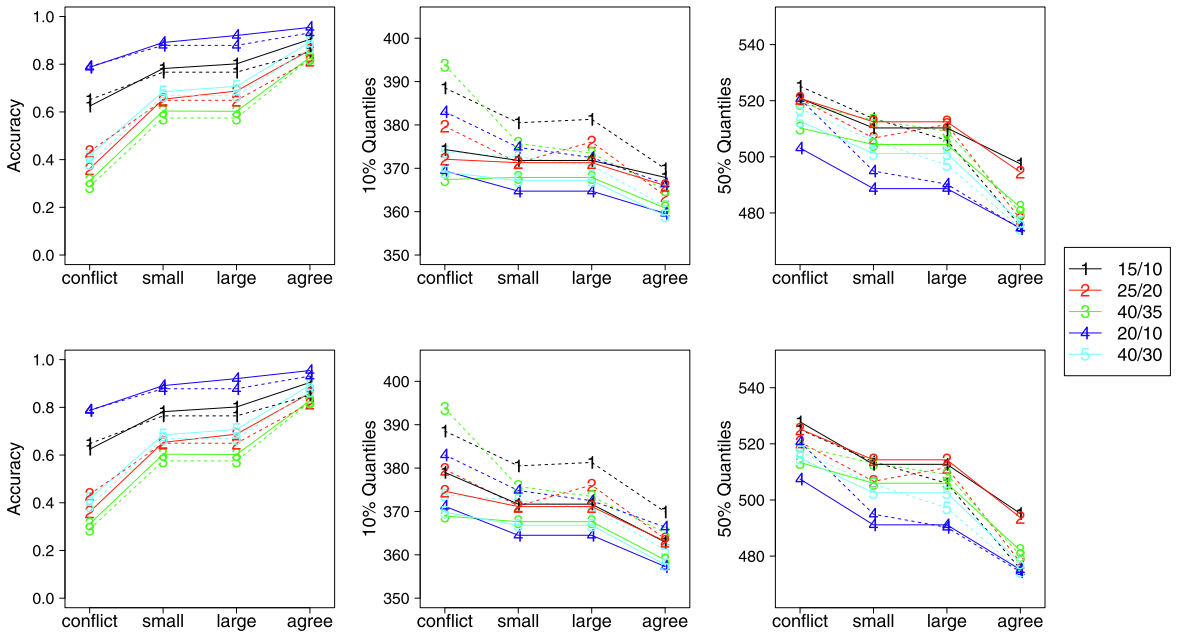

Fig. 7. Accuracy, 10% quantiles, and 50% quantile RTs plotted against the conditions with different degrees of conflict for Experiment 2. Dashed lines represent the observed result while solid lines represent predictions. The top three panels show the predictions from the log–log interaction model with the moderated numerosity difference (Model 2 in Table 8). The bottom three panels show the predictions from the full model with the nondecision time regression model (Model 3). For visual clarity, the plots show only five numerosity pairs. In the other two pairs in which the numerosity difference is large ($\Delta = 20$), the effect of conflict was smaller.

predicted the longest reaction time in this condition. Predictions for the equal conditions were between the agree and conflict conditions, which matched the observed results.

Predictions from the other models were very similar to those in Fig. 6 because the moderated numerosity difference and the nondecision time regression had only small effects on accuracy patterns. The ability of the model to capture most of the trends in the data for the L/R task suggests that the ANS diffusion model approach can be generalized to different types of numerosity judgment tasks.

The model also accounted for error RTs quite well. Some misfits in error RTs were mainly due to the fact that accuracy was very high in some conditions. In these cases, there were only a few error responses and so the error reaction time distributions were poorly estimated. This occurred in most of the agree conditions and some of the equal conditions in which accuracy was around or higher than 0.9. For the cases with accuracy higher than 0.95, the error RT distributions cannot be obtained properly (because many subjects had less than 5 observations) and so only the median RTs were plotted for these cases ('M' in the figure). Except for these cases, predicted error RTs matched the corresponding data well.

The nondecision time regression model produced a qualitative improvement in fits to the increasing pattern of the leading edge of RT distributions from the agree to conflict conditions. Fig. 7 shows accuracy, the 10% quantile RT, and the median RT for the grouped conditions. In each panel, dashed lines represent observations and the solid lines represent predictions. For the top three panels, the predictions were generated from the log–log interaction model with the moderated numerosity difference (Model 2 in Table 8). For the bottom three panels, the predictions were generated from the full model which included the nondecision time equation (Model 3 in Table 8). Consistent with what we found in Experiment 1, the RT leading edge and median RT increased and accuracy decreased as the degree of conflict increased. However, the increase in the leading edge was smaller than Experiment 1, and, for example, the leading edge was close to flat across conditions for the numerosity pair of (40, 30). The log–log interaction model was able to capture the behavior of RT medians but failed to explain all of the leading edge shift. Adding the nondecision time regression improved the fit to the leading edge, and the mismatch between theory and data was about 15 ms. While this is not large, it is quite systematic.

4.3.4. Estimating the relative contributions of variables

Parameter estimates of the models are in Table 9. Given these estimates, the drift rate and the across-trial variability in drift rate can be estimated using their regression equations. For the log–log interaction model with the moderated numerosity difference and the nondecision time regression, we obtained the estimates shown in the equations below.

$$v = 0.777*(\log(N_1/N_2)) + 0.008*(\log(C_1/C_2))*(N_1 + N_2) - 3.40*(\log(N_1/N_2))/(N_1 + N_2)$$

$$\eta = 0.165 + 0.0046\sqrt{N_1^2 + N_2^2}$$

$$T_{er}=0.388 - 1.80 * (\log(C_1/C_2))/((N_1 - N_2)(N_1 + N_2))$$

Table 9
Parameter estimates of some selected models for Experiment 2. a is the boundary separation, T_{er} is the nondecision time, and s_z and s_t are the across-trial variability in starting point and nondecision time, respectively. v_i's are the drift rate coefficients in the drift rate equations. η_0 and σ_1 are the intercept and slope of the across-trial variability in drift rate. t_0 and t_1 are the intercept and slope of the nondecision time regression used in Model 3. The starting point parameter z is fixed as $z = a/2$.

Model	a	T_{er}	η_0	$10\sigma_1$	s_z	s_t	v_1	v_2	$v_3/10$	t_0	$t_1/10$
Model 1	0.113	0.383	0.196	0.036	0.090	0.200	0.685	0.007			
Model 2	0.113	0.383	0.170	0.044	0.090	0.200	0.757	0.008	−0.270		
Model 3	0.111		0.165	0.046	0.089	0.212	0.777	0.008	−0.340	0.388	−0.180

Model 1: Log - Log Interaction
Model 2: Log - Log Interaction + Moderated Numerosity
Model 3: Log - Log Interaction + Moderated Numerosity + Nondecision time regression
* Some parameter estimates are very similar for Model 1 and 2 and this is consistent over subjects. The actual values are different at the 4th/5th decimal place. This is because the added term (moderated numerosity difference) in Model 2 has a very small effect and so parameter estimates of other parameters such as a and T_{er} are not affected by the additional term.

The drift rate coefficients in the drift rate model represent the effects of the variables given that the other covariates are controlled. Therefore, the estimates of the coefficients provide a way to evaluate the relative contributions of the variables in the experiment. Because the variables have different scales, however, the magnitudes of the parameter values cannot be directly compared. Instead, we calculated the range of each variable multiplied by the corresponding coefficient estimate, as we did in Experiment 1. This range shows how much the variable contributed to drift rate.

We carried out this calculation for Model 3 and the result is shown in Table 10. The range of the drift rate was (-0.618, 1.143) and that of the across-trial variability in drift rate was (0.248, 0.410). Among the components in the drift rate equation, the numerosity effect had the range of (0.104, 0.853), the convex hull interaction had the range (−0.715, 0.715), and the moderated numerosity difference had the range (−0.092, −0.006). This shows that the contribution from numerosity to drift rates was strongly affected by the interaction with the convex hull, but the moderated numerosity difference had only a small effect on the drift rate. In particular, when there was a conflict between numerosity and convex hull and the discrimination was difficult, the convex hull interaction term produced a negative drift rate which enabled the model to predict the below-chance accuracy. Taken together with the findings in Experiment 1, this result confirmed that the perceptual variables can produce a large effect on numerosity discrimination, sometimes even dominating the judgment. This effect is modulated by the total numerosity.

The different results for the moderated numerosity difference between Experiments 1 and 2 might be attributed to the difference between the linear representation and the log representation of numerosity. For a fixed numerosity difference, the linear model is not able to produce a difference in drift rate. That is, (15, 10) and (40, 35) produce the same contribution to the drift rate because the difference is the same (i.e., $\Delta = 5$). Thus, without an additional term, the across-trial variability in drift rate is the only factor that can explain the size effect in the linear-log interaction model in Experiment 1. However, it turned out that the variability model was not sufficient to fit observed response proportions in Experiment 1. Therefore, the moderated numerosity difference was able to make a significant contribution to the models with the linear representation of numerosity. In contrast, the log model naturally produces the size effect for numerosity pairs with the same difference by the log-ratio of numerosities: for example, $\log(15/10) = 0.405$ and $\log(40/35) = 0.134$. Therefore, when the log model is used, the moderated numerosity difference might not be required. This implies another task-dependency in numerosity judgment tasks because the linear model accounts for data better in the B/Y task but it needs the moderated numerosity difference to capture the size effect properly. In contrast, the log model fits data better in the L/R task and it was able to explain the size effect with the log contribution of numerosity to drift rate alone.

Nondecision time had a range of (0.369, 0.407), which was smaller than the range obtained in Experiment 1. Although the nondecision time equation enabled the full model to capture the increasing pattern in the leading edge better than the log–log interaction model, the small range of nondecision time was obtained because the conflict had a smaller effect on the RT leading edge in the L/R task.

Table 10
Range of drift rate, across-trial variability of drift rate, nondecision time, and their components. For Experiment 2, these values were estimated using the log–log interaction model with the moderated numerosity difference and the nondecision time regression. N_i's are the numbers of dots and C_i's are convex hull areas. v_i's are the drift rate coefficients, η_0 and σ_1 are the intercept and slope of the across-trial variability model, and t_0 and t_1 are the intercept and slope of the nondecision time regression.

Drift Rate (v) (-0.618, 1.143)			Across-trial Variability (η) (0.248, 0.410)		Nondecision Time (T_{er}) (0.369, 0.407)	
Numerosity $v_1(\log N_1/N_2)$	Convex hull Interaction $v_2\log\left(\frac{C_1}{C_2}\right)(N_1+N_2)$	Moderated Numerosity $v_3\frac{N_1-N_2}{N_1+N_2}$	η_0	$\sigma_0\sqrt{N_1^2+N_2^2}$	t_0	$\frac{\log(C_1/C_2)}{(N_1-N_2)(N_1+N_2)}$
(0.104, 0.853)	(-0.715, 0.715)	(-0.092, −0.006)	0.165	(0.083, 0.245)	0.388	(-0.019, 0.019)

Table 11
Experiment 3: Accuracy and correct mean RT. The conditions were defined by dot areas.

Experiment and task	Conditions	Accuracy	Mean RT
3, B/Y	Agree	0.842	656
	Large	0.788	695
	Small	0.776	712
	Conflict	0.644	731

4.4. Experiment 3

Experiment 3 was carried out to examine whether limited stimulus presentation time resulting in reduced stimulus processing was responsible for the conflict effects found in Experiment 1. Experiment 3 used the B/Y task and the experimental conditions were the same as in Experiment 1, but the stimulus remained on the screen until a response was made.

Because Experiment 3 shared the design of Experiment 1, we grouped the data and fitted the models in the same way. We defined the conditions by numerosity pairs and dot areas, fitted the ANS diffusion models, and generated predictions. However, we restricted our interest to the three models we applied in Experiment 1 and 2. The first one was the full model which had the linear difference of numerosity pairs ($N_1 - N_2$), the interaction between numerosity pairs and dot areas $\log(A_1/A_2)(N_1 + N_2)$, the moderated numerosity difference $\frac{N_1-N_2}{N_1+N_2}$, and the nondecision time regression $t_0 + t_1\frac{\log(A_1)-\log(A_2)}{(N_1-N_2)(N_1+N_2)}$ (Model 12 in Table 3). This was the best-fitting model in Experiment 1, capturing all the patterns of results across the different conditions. This full model was compared to two simpler models. The first one was the interaction model without the other two terms (Model 8 in Table 3) and the second one was the interaction model with the moderated numerosity difference but without the nondecision time regression (Model 11 in Table 3).

4.4.1. Results

Table 11 shows accuracy and mean correct RTs in Experiment 3 for the agree, large, small, and conflict conditions. The general pattern was consistent with the results from the previous experiments and showed that numerosity discrimination was affected by perceptual variables. Comparing the results with those from Experiment 1, accuracy was slightly lower, but the patterns in the accuracy data were similar. The mean RTs were longer by 50 ms or so which can be attributed to the longer presentation time.

Table 12 presents accuracy values and mean RTs for correct responses as a function of numerosity and agree, equal, and conflict conditions as in Table 2. As in Experiment 1, accuracy decreased as the total number of dots increased, but this change was much larger in the conflict condition than the other conditions. In the conflict condition, accuracy fell below chance when the total number of dots was large and the numerosity difference was small (e.g., 40 vs 35). This result showed that the effect of perceptual variables was moderated by the numerosity and that area sometimes dominated judgments. Mean RTs also showed a similar pattern as shown in Experiment 1. In general, mean RTs decreased as the numerosity difference decreased. As a function of the total number of dots, mean RTs decreased when the numerosity difference was 5 but increased when the numerosity difference was 10 or 20. This replicates the results from Experiment 1 and Ratcliff and McKoon (2018).

4.4.2. Fitting the ANS diffusion models

Parameter estimates and G^2 values of the fitted models are presented in Table 13. Generally, the boundary separation and nondecision time components were larger in Experiment 3 than in Experiments 1 and 2. Except for this, there were no sizable differences between the parameter values for Experiment 3 and those for Experiment 1. The models with the moderated numerosity difference and the nondecision time regression improved the fit over the model without these components. These results were consistent with those from Experiment 1.

The quantile probability plots for Experiment 3 are shown in Fig. 8. All the behavioral patterns were similar to those from Experiment 1. The predictions matched the observations about as well as those in Fig. 4 for Experiment 1.

Fig. 9 shows the accuracy, the leading edge, and the median RT in Experiment 3 by conditions. In both figures, the top three panels display the predictions generated from the linear-log interaction model with the moderated numerosity difference (Model 2 in Table 13) while the bottom three panels present predictions from the full model (Model 3 in Table 13) with the nondecision time

Table 12
Experiment 3: Accuracy and mean RTs for correct responses as a function of numerosity pairs and the three area conditions (Agree, Equal, and Conflict).

Expt, task	Measure	Condition	Δ = 5			Δ = 10		Δ = 20	
			(15, 10)	(25, 20)	(40, 35)	(20, 10)	(40, 30)	(30, 10)	(40, 20)
3, B/Y	Accuracy	Agree	0.819	0.800	0.800	0.862	0.856	0.877	0.887
		Equal	0.775	0.697	0.662	0.840	0.746	0.882	0.866
		Conflict	0.669	0.514	**0.416**	0.786	0.536	0.840	0.758
	Mean RT	Agree	697	668	667	649	637	606	614
		Equal	758	756	732	688	709	620	642
		Conflict	813	809	779	732	769	667	709

Table 13
Parameter estimates, G^2 values, and degrees of freedom (*df*) of the models used in Experiment 3. The models are defined in Experiment 1, Table 3. a is the boundary separation, T_{er} is the nondecision time, and s_z and s_t are the across-trial variability in starting point and nondecision time, respectively. v_i's are the drift rate coefficients in the drift rate equations. η_0 and σ_1 are the intercept and slope of the across-trial variability in drift rate. t_0 and t_1 are the intercept and slope of the nondecision time regression used in Model 3. The starting point parameter z is fixed as $z = a/2$.

Model	a	T_{er}	η_0	$10\sigma_1$	s_z	s_t	v_1	v_2	v_3	t_0	$t_1/10$	G^2	*df*
Model 1	0.153	0.438	0.082	0.047	0.122	0.210	0.023	0.002				393.8	300
Model 2	0.152	0.433	0.133	0.029	0.115	0.210	0.011	0.002	0.462			378.2	299
Model 3	0.146		0.107	0.029	0.100	0.199	0.011	0.002	0.425	0.427	−0.069	362.0	298

Model 1: The linear-log interaction model (Model 8 in Table 3)
Model 2: The linear-log interaction model + Moderated numerosity difference (Model 11 in Table 3)
Model 3: The linear-log interaction model + Moderated numerosity difference + Nondecision time regression (Model 12 in Table 3)

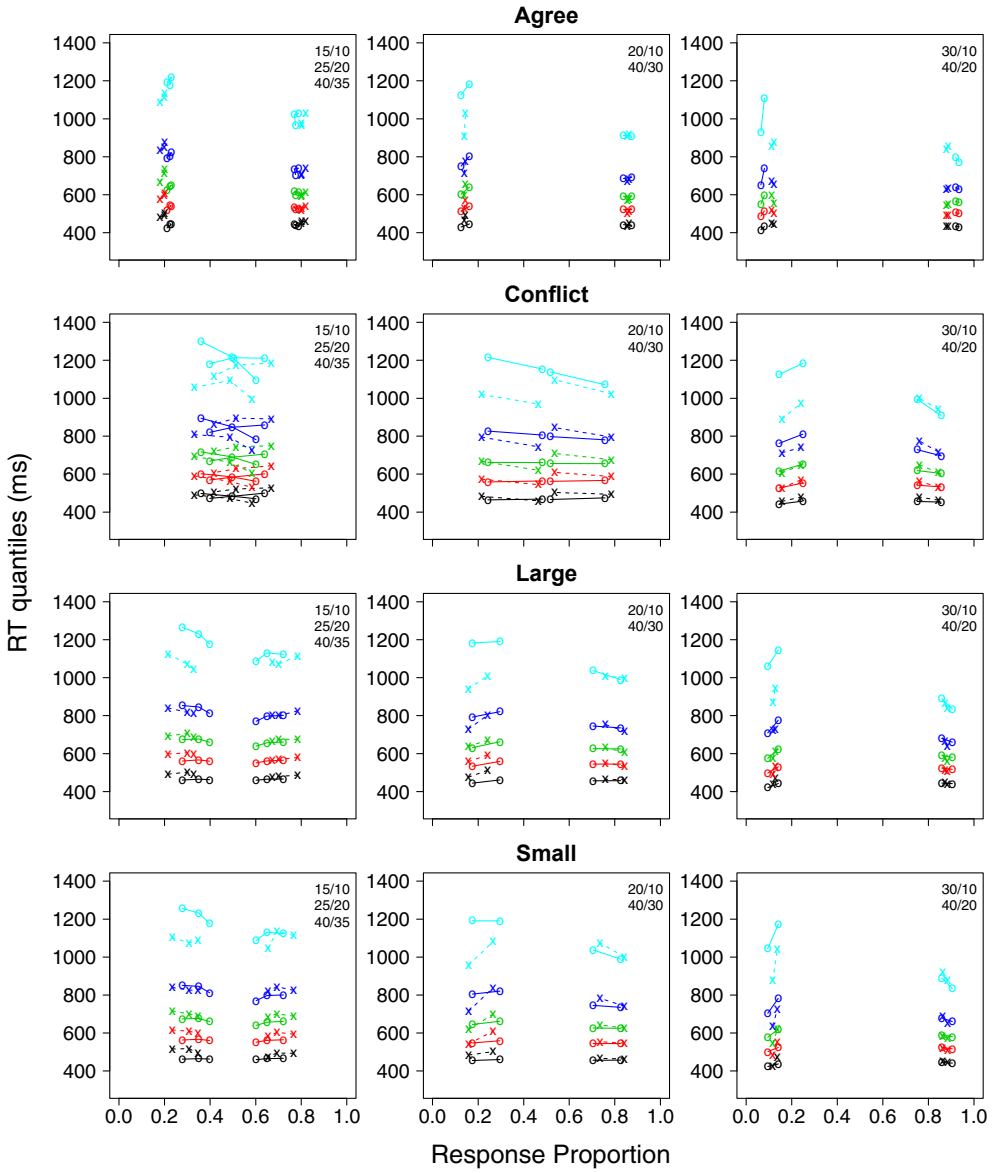

Fig. 8. Quantile-probability plot for Experiment 3 for the linear-log interaction model (Model 3 in Table 13) with the moderated numerosity difference effect and nondecision time regression.

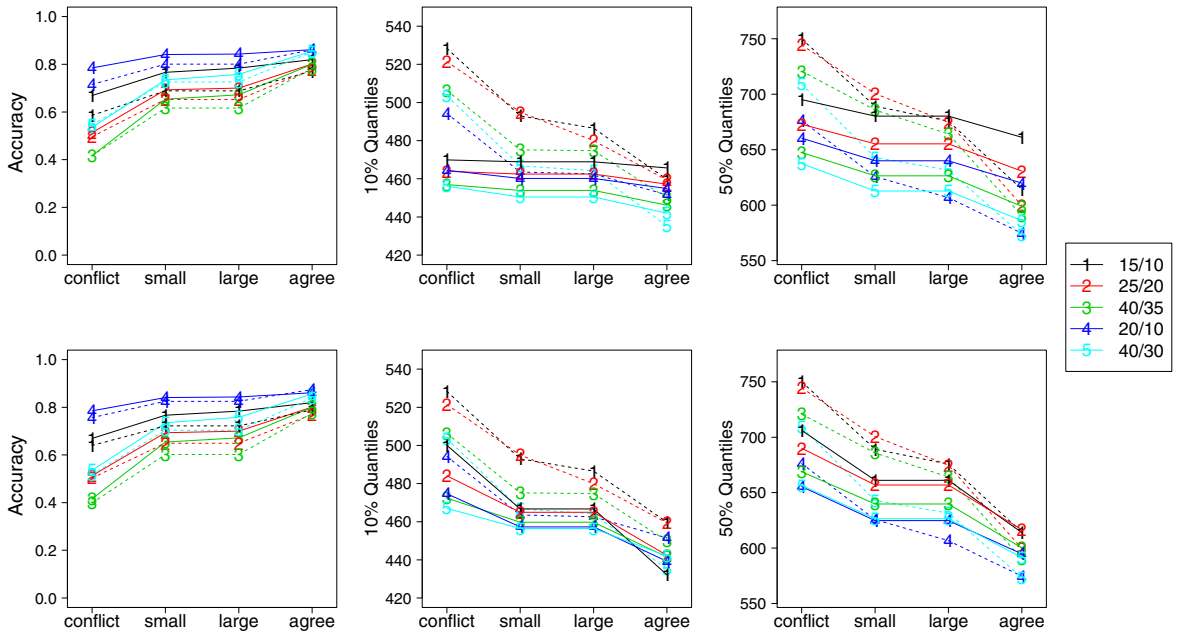

Fig. 9. Accuracy, 10% quantiles, and 50% quantile RTs plotted against the conditions with different degrees of conflict for Experiment 3. Dashed lines represent the observed result while solid lines represent predictions. The top three panels show the predictions from the linear-log interaction model with the moderated numerosity difference (Model 2 in Table 13). The bottom three panels show the predictions from the full model with the nondecision time regression model (Model 3 in Table 13). For visual clarity, the plots show only five numerosity pairs. In the other two pairs in which the numerosity difference is large ($\Delta = 20$), the effect of conflict was smaller.

regression. The shift of the RT distribution as a function of the conflict, found in both Experiments 1 and 2, was replicated in the data. The linear-log interaction model with the moderated numerosity difference was not able to explain this shift, but the model with the nondecision time regression captured all except about 20 ms of the vertical misfit as it did in the previous experiments.

Experiment 3 replicated most of the behavioral patterns observed in Experiments 1 and 2. The ANS diffusion model fit about as well as it did in Experiments 1 and 2. In particular, the model fit accuracy patterns such as below chance accuracy, the effect of the numerosity difference moderated by the total number of dots, and most of the effect of conflict on the RT distribution. The results show that the conflict effect we observed in the previous experiments is not due to a short presentation duration.

4.5. Experiment 4: Area judgment task

In the first three experiments, we examined the effects of non-numeric perceptual variables on numerosity judgments. We found a new conflict effect on RT which produced a slowdown in the leading edge of RT distributions when perceptual variables were in conflict with numerosity. In conflict conditions, we found that accuracy could go below chance when numerosity values were larger.

In this section, we ask the opposite question: does numerosity influence judgments about a non-numeric perceptual variable and does it do it in the same way as in Experiments 1–3. To answer this question, we conducted an experiment using the same stimuli as in Experiment 1. The task was to judge whether blue dots or yellow dots in a single array had a larger total area. In this 'area judgment' task, the main variable upon which subjects should base their judgment was a perceptual feature of the stimulus, and dot numerosity is the confounding variable.

4.5.1. Results

In order to explore how numerosity affects area judgment, accuracy was calculated for different conditions. First, the total area values used in the task were divided into four conditions: a difference of 0–1000, 1000–2000, 2000–4000, and 4000 or more. These are computed from the radius of each dot and the number of dots. For these divisions, in Tables 14 and 15, we present accuracy values and 10% quantile RTs for, first, whether numerosity is proportional to or conflicting with the total area, second, as a function of differences between two numbers, and third, as a function of the sum of the number of dots. As expected, accuracy was largely affected by the total area differences. Within the same total area conditions, accuracy fell when there was a higher degree of conflict between number and total area or when the total numerosity was larger. This interaction became more salient when there was a smaller difference between the two total areas, that is, when the task was more difficult. This interaction pattern was the same as was obtained in Experiments 1–3.

However, the 10% quantile RTs in Table 15 showed an effect of the total area difference but only a small effect of the numerosity variables: proportional versus conflicting, difference, or total number. As area difference increased, the 10% quantile became shorter. Within the same area difference, 10% quantiles did not vary much across different numerosity conditions (congruency, difference, or sum). This result demonstrates that the interference effect from the conflict between numerosity and perceptual features is

Table 14
Accuracy as a function of total area differences and numerosity conditions.

Accuracy	4000–10000	2000–4000	1000–2000	0–1000
Congruency				
Number – proportional	0.903	0.881	0.855	0.712
Number - conflicting	0.902	0.861	0.760	0.601
Difference ($N_1 - N_2$)				
20	0.891	0.887	0.853	0.707
10	0.912	0.883	0.844	0.741
5	0.917	0.875	0.863	0.693
−5	0.891	0.874	0.739	0.617
−10	0.922	0.831	0.803	0.591
−20	–	0.854	0.740	0.579
Sum ($N_1 + N_2$)				
25–30	0.917	0.892	0.848	0.693
40–45	0.900	0.887	0.830	0.650
60	0.891	0.849	0.801	0.648
70–75	0.908	0.859	0.754	0.602

Table 15
Leading edges of RT distributions (10% quantiles) as a function of total areas and numerosity conditions.

Leading edge (10% Quantiles)	4000–10000	2000–4000	1000–2000	0–1000
Congruency				
Number – proportional	350	358	369	386
Number – conflicting	356	360	372	386
Difference ($N_1 - N_2$)				
20	349	357	368	385
10	353	357	372	392
5	355	361	378	388
−5	355	361	374	389
−10	357	360	378	394
−20	–	370	376	382
Sum ($N_1 + N_2$)				
25–30	357	356	375	391
40–45	356	359	369	380
60	352	358	370	388
70–75	353	360	371	384

asymmetric. In the numerosity tasks, confounding variables can have large effects on task performance and reaction times, but in the perceptual task, had a moderate effect on accuracy but little effect on the leading edge of the RT distribution.

4.5.2. Model development

The next question is if the ANS diffusion model can fit the interaction between numerical and perceptual variables in the area judgment task. In the previous experiments, it was shown that the ANS diffusion model can fit most of the behavioral patterns but with misses in the RT leading edge. Although the model with nondecision time regression explained the increasing pattern in the RT leading edge induced by the conflict, the predicted 10% quantiles were somewhat shorter than the experimental values across the conditions. Unlike the results from the numerosity judgment tasks, numerosity did not show a noticeable effect on the RT leading edge in the area judgment task. Therefore we might expect that an integrated diffusion model will fit the data without misfits in the leading edge of the RT distributions.

An integrated diffusion model for the area judgment task (as in Ratcliff et al., 2018) was developed and fitted in the same way as in the previous experiments. First, experimental conditions were defined by dot areas as in Experiment 1 and the large and the small conditions were combined which produced three conditions (agree, conflict, and equal). The model was constructed in the same way as the models used in the numerosity judgment tasks, except that the major perceptual variable here is the area. Because there is an interaction effect of the area difference and numerosity on accuracy, the model included an interaction term. However, because there was little effect of numerosity on the RT leading edge, we did not include a model of nondecision time. The model evaluated was:

$$v = v_1(\log A_1 - \log A_2) + v_2(\log A_1 - \log A_2) * (N_1 + N_2)$$

$$\eta = \eta_0 + \sigma_1\sqrt{A_1 + A_2}$$

$$A_i = \pi * r_i^2 (i = 1, 2)$$

Table 16
Parameter estimates, G^2 value, and the degree of freedom (*df*) of the ANS diffusion model for the area judgment task used in Experiment 4. a is the boundary separation, T_{er} is the nondecision time, and s_z and s_t are the across-trial variability in starting point and nondecision time, respectively. v_i's are the drift rate coefficients in the drift rate equations. η_0 and σ_1 are the intercept and slope of the across-trial variability in drift rate. The starting point parameter z is fixed as $z = a/2$.

Expt	a	T_{er}	η_0	$10\sigma_1$	s_z	s_t	v_1	v_2	G^2	*df*
4	0.114	0.375	0.042	0.029	0.091	0.193	0.186	0.0033	290.8	223

4.5.3. Fitting the ANS diffusion model

The model was fit using G^2 method and the parameter estimates and the G^2 value are presented in Table 16. The quantile probability plots for the model are shown in Fig. 10. The results show that the model fits the patterns of results from the area judgment task well. Across all the conditions, predicted RT distributions matched the observed distributions, and in particular, the model was able to fit the RT leading edges well (without the need for any nondecision time model). For the response proportions, there were some misses in the equal conditions, but overall, predicted accuracy was close to the observed response proportions.

Overall, the area judgment task and its integrated diffusion modeling presented here demonstrate that the conflict effect is not symmetric across the tasks. The way in which area affects numerosity judgments is not the same as the way in which numerosity affects area judgments. While area produces the conflict effect on numerosity judgments in terms of accuracy and RT leading edges, numerosity produces the conflict effect only on accuracy in area judgments. The result that the model fits the data from the area judgment task without a nondecision time component supports this finding.

5. Maximum likelihood estimation

For the analysis presented above, we employed the G^2 fitting method to fit the various ANS diffusion models using grouped data. We defined four conditions based on the numerosity pairs and perceptual features of interest. Then, we aggregated data by these

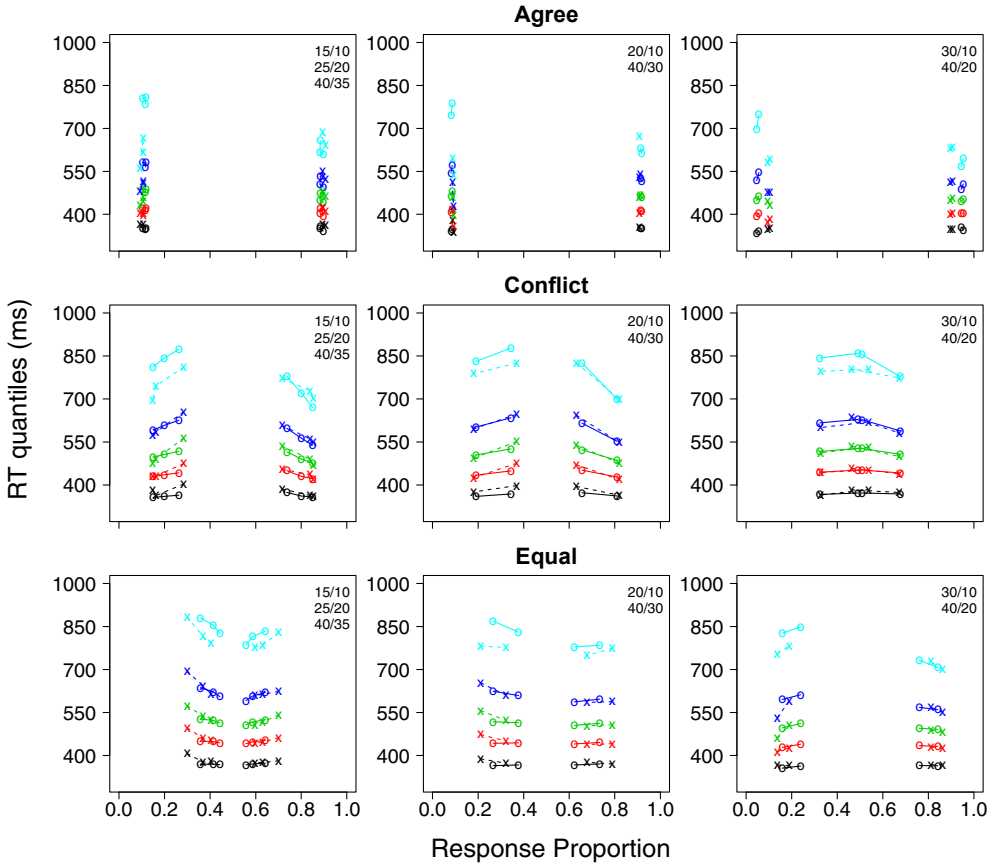

Fig. 10. Quantile-probability plot of the ANS diffusion model for the area judgment task used in Experiment 4.

Table 17
Comparison between G^2 and MLE results (parameter estimates). The MLE result in the last line is for the full model, which includes both dot area and convex hull in the modeling. a is the boundary separation and s_z and s_t are the across-trial variability in starting point and nondecision time, respectively. v_i's are the drift rate coefficients in the drift rate equations. η_0 and σ_1 are the intercept and slope of the across-trial variability in drift rate. t_0 and t_1 are the intercept and slope of the nondecision time regression. The starting point parameter z is fixed as $z = a/2$.

Expt	Method	a	η_0	$10\sigma_1$	s_z	s_t	v_1	v_2	v_3	v_4	t_0	$t_1/10$
1	G^2	0.125	0.110	0.034	0.095	0.250	0.0129	0.0025	0.542		0.450	−0.540
1	MLE	0.125	0.073	0.037	0.093	0.280	0.0123	0.0019	0.496		0.436	−0.501
2	G^2	0.111	0.165	0.046	0.089	0.212	0.777	0.008	−3.351		0.388	−0.180
2	MLE	0.117	0.197	0.035	0.085	0.132	0.544	0.005	−1.167		0.351	−0.073
2	MLE, full	0.118	0.246	0.026	0.089	0.155	0.614	0.006	−1.559	0.00088	0.361	−0.191

conditions to produce the RT quantile data used in the G^2 fitting method. However, as discussed earlier, the G^2 method does not consider the full range of the individual data points, inevitably losing some information. Despite grouping the data like this, it has been shown that the χ^2 and G^2 methods are able to produce valid and reliable results (Ratcliff & Childers, 2015). To avoid this data aggregation and to check for biases and distortions from the G^2 method, we used the MLE method which uses each of the data points in estimation. Because MLE uses all the individual data points to estimate parameters, there is no need to define conditions and aggregate data by these conditions to fit the ANS diffusion models. However, the MLE method is sensitive to outliers and if cutoffs are used, nondecision time will vary as a function of the lower cutoff. Fortunately, this was not a problem in the analyses presented here.

We fitted the interaction model with the moderated numerosity difference and the nondecision time regression models as used in Experiments 1 and 2 (Model 12 in Table 3 and Model 3 in Table 8), using MLE. The parameter estimates are presented in Table 17 with the corresponding parameters from the G^2 method (from Tables 3 and 8). The values in the table show that the two methods produced similar parameter estimates. Some mismatches, especially for v_3 (the slope coefficient corresponding to the moderated numerosity difference) and t_1 (the slope coefficient in the nondecision time regression) for Experiment 2 occurred because the moderated numerosity difference and the slope of the nondecision time had only small effects on performance in this experiment. Specifically, the estimated ranges of these two terms were (−0.032, −0.002) for the moderated numerosity difference and (−0.007, 0.007) for the nondecision time from the MLE method, and (−0.092, −0.006) and (−0.019, 0.019) from the G^2 method. Although there were some differences in the coefficients, the differences in their ranges were much smaller.

The agreement between MLE and G^2 implies that both methods can represent the information and behavioral patterns underlying the data equally well. Even though the G^2 method lost some information due to data aggregation, it was able to produce fits comparable to MLE. This agrees with the simulation results from Ratcliff and Childers (2015).

An advantage of MLE over G^2 is that MLE is capable of considering all the variables at once. If we fit the model with the G^2 method, we have to aggregate data in some way. The more conditions we use, the smaller the number of observations per condition that can be used to compute quantile RTs (often too few to provide quantiles in error conditions with high accuracy). For this reason and because the area had little effect on performance, we considered the convex hull variable and did not examine the area in model fitting with the G^2 method for Experiment 2.

To examine both the area and convex hull variables in Experiment 2, we used the MLE method to fit the full model for Experiment 2 with the drift rate defined as:

$$v = v_1\left(\log\left(\frac{N_1}{N_2}\right)\right) + v_2(\log(\frac{C_1}{C_2})) * (N_1 + N_2) + v_3\frac{(\log(N_1) - \log(N_2))}{N_1 + N_2} + v_4(\log(\frac{A_1}{A_2})) * (N_1 + N_2)$$

This equation is different from the model we fit with G^2 method (Model 3 in Table 8) because it includes the dot area variable which was not included in the previous analysis. The parameter estimates and G^2 value are presented in the last line of Table 17. The dot area interaction effect had a regression coefficient of 0.00088 which produces a range in drift rate of (−0.129, 0.129). This was smaller than the main effect of the log ratio of numerosities (0.082, 0.574) and the convex hull interaction (−0.530, 0.530) which shows that the effect of area was small. This analysis demonstrates that MLE has an advantage when the model involves relationships among multiple variables because it can include all the variables into analysis at once. So long as there is no problem with outlier RTs, the MLE method makes models with multiple variables more tractable.

One of reviewers suggested a model comparison study with MLE fits of the models used for the data from Experiment 1 to confirm that the aggregation used in performing G^2 analyses does not distort model selection results. Because the interaction term between area and numerosity was necessary to explain the observed accuracy patterns across conditions, we only compared models with this interaction term (models 7–12 in Table 3). Table 18 shows the negative log likelihood values for the models. Out of Models 7–10 that have main effects and interaction effects, Model 8 (Linear–Log Interaction) performed the best. This agrees with the G^2 result presented in Experiment 1. Among the other models, Model 7 performed the worst and Model 10 performed better than it did in the G^2 analysis. However, this difference in the order of Models 7–10 did not change the overall conclusions because Model 8 performed

Table 18
MLE model comparison of the models used for Experiment 1 data. N_1 and N_2 are two numerosities and A_1 and A_2 are two dot areas. Dot area is computed as $A_i = \pi * r_i^2$, where r_i is a dot radii. For all of the models, across-trial variability of drift rate was modeled by $\eta = \eta_0 + \sigma_1\sqrt{N_1^2 + N_2^2}$. $-ll$: negative log likelihood.

Model			$-ll$
7	Linear - Linear Interaction	$v = v_1(N_1 - N_2) + v_2(A_1 - A_2) * (N_1 + N_2)$	122.4
8	Linear - Log Interaction	$v = v_1(N_1 - N_2) + v_2(\log(A_1/A_2))*(N_1 + N_2)$	108.6
9	Log - Linear Interaction	$v = v_1(\log(N_1/N_2))+v_2(A_1 - A_2) * (N_1 + N_2)$	121.7
10	Log - Log Interaction	$v = v_1(\log(N_1/N_2))+v_2(\log(A_1/A_2))*(N_1 + N_2)$	113.0
11	Linear - Log Interaction + Moderated Numerosity	$v = v_1(N_1 - N_2) + v_2(\log\left(\frac{A_1}{A_2}\right)) * (N_1 + N_2) + v_3\frac{(N_1 - N_2)}{N_1 + N_2}$	103.2
12	Linear - Log Interaction + Moderated Numerosity + T_{er} Model	Model 11 for the drift rate v with $T_{er} = t_0 + t_1\frac{\log(A_1) - \log(A_2)}{(N_1 + N_2) * (N_1 - N_2)}$	97.8

the best in both analyses. Next, we added the two effects, the moderated numerosity difference and the nondecision time regression, to Model 8. The negative log likelihood values showed that those effects improved the model fit, which is consistent with the G^2 result. In sum, the MLE model comparison led us to the same conclusions as the G^2 model comparison.

6. Comparison with a previous regression-based approach: DeWind et al. (2015)

DeWind et al. (2015) proposed a generalized linear model for the choice probability that uses a similar approach to the ANS diffusion model presented here. In DeWind et al.'s approach, the probability is represented as a function of numerosity and two perceptual variables, size and spacing. These variables represent the effects of non-numeric features and they were argued to be orthogonal to numerosity so that numerosity can be studied separately from perceptual variables. The model is presented in detail in Appendix B.

Unlike the ANS diffusion model, DeWind et al.'s model accounts for accuracy data only. However, we can compare the two models on the basis of their accuracy predictions. To do this, we fitted their model to the data from Experiment 2 and compared its predictions for accuracy with those from the ANS diffusion model. Fig. 11 shows the observed accuracy in the L/R task in Experiment 2 (the black solid lines with the dots) and the corresponding prediction from DeWind et al.'s model (the red dotted lines with the triangles). Both observations and predictions were computed for each subject and then averaged across all the subjects. This was done for each combination of numerosity pairs and conditions. Fig. 11 also shows accuracy predictions from the ANS diffusion model (the blue dotted lines with the squares) computed in the same way (fitting individual subjects and averaging over the predictions). Both models fit accuracy data well, and there were no large differences between the two models. The root mean squared error of prediction

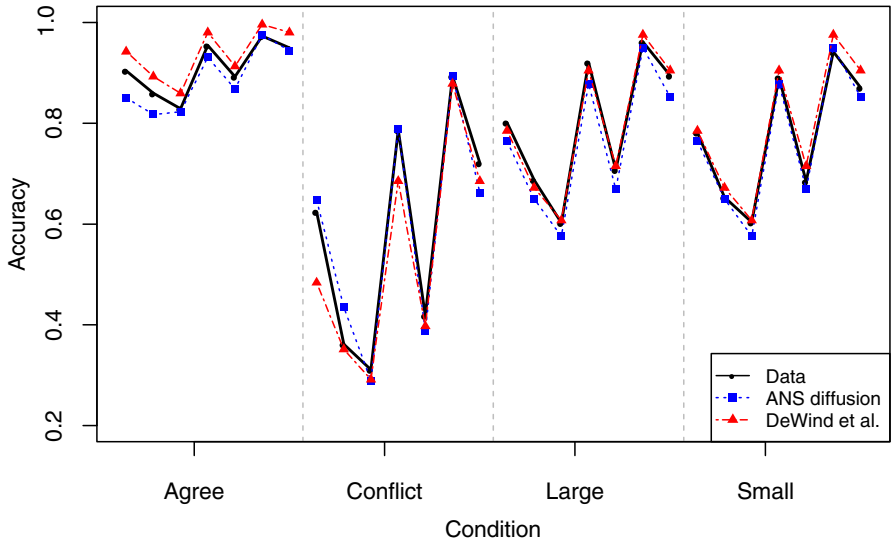

Fig. 11. Accuracy prediction from the ANS diffusion model and DeWind et al.'s model. Data is from the task with two spatially separated arrays (Experiment 2). The four panels divided by the gray dashed lines indicate different conditions of convex hulls (Agree, Conflict, Large, and Small). In each panel, the seven points represent different numerosity pairs (15/10, 25/20, 40/35, 20/10, 40/30, 30/10, and 40/20). The black solid lines with circle points indicate observed accuracy, the blue dashed lines with squares show predictions from the ANS diffusion model, and red dashed lines with triangles indicate predictions from DeWind et al.'s model (for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

was 0.039 for DeWind et al.'s model and 0.031 for the ANS diffusion model. As the ANS diffusion model did, DeWind et al.'s model was able to account for accuracy lower than 0.5 in the conflict condition with high numerosity.

In sum, the ANS diffusion model accounts for accuracy as well as DeWind et al.'s accuracy-only model but the ANS diffusion model also accounts for the behavior of RT distributions for correct and error responses.

7. Discussion

The modeling approach presented here aims at examining the relationship between numerosity and perceptual variables in two numerosity discrimination tasks and has attempted to model the effects of numerosity and non-numeric perceptual variables on both accuracy and RT. Results presented here have shown that non-numeric characteristics of a stimulus affect numerosity judgments (as in many other published studies), and our results show that they even dominate the decision-making process under some conditions. When visual cues of stimuli, such as the dot area and the convex hull surrounding the dots, agree with numerosity (are congruent with numerosity), performance is improved, but when the perceptual features are incongruent with numerosity, performance is decreased. We also have shown that the size of the effect is modulated by the total numerosity.

The conflict effect from non-numeric variables that is obtained here is much larger than that obtained in the previous literature. Our results show that when the number of dots is large, and there is a conflict between numerosity and non-numeric variables, the numerosity signal (component of drift rate) is weak and non-numeric features dominate the decision. This results in lower than chance accuracy for judgments of numerosity. When the number of dots is small, numerosity information dominates non-numeric information and accuracy is above chance. In Ratcliff and McKoon (2018), these conflict conditions were not part of the design and non-numeric variables were either consistent with numerosity information or were equated across conditions (note, as discussed earlier, all non-numeric variables cannot be controlled, for example, equating area means on average, a smaller numerosity has larger dots).

The framework provided by the ANS diffusion model provides a way of understanding the relationship between perceptual and numerical variables. The integration of the two kinds of variables allows models of representation to be evaluated within the context of all aspects of accuracy and RT data. This provides a much more powerful tool for testing models of representation and decision than using either variable alone. In the modeling, drift rate is determined by both perceptual and numerical variables through drift rate coefficients multiplied by functions of those variables. The role of each variable was represented by its coefficient in the drift rate model. Because the numerosity and non-numeric variables had different scales, the coefficients themselves cannot be used to directly compare the effects of the variables. Instead, the range of each variable multiplied by the coefficient showed the size of contributions of the variables to the drift rate. In this way, the effects of the variables could be compared on a scale-free basis.

The model fits showed that in numerosity judgment tasks, perceptual variables had different effects depending on the task. In the B/Y task, dot area had a large effect on numerosity judgments, but in the L/R task, in which both area and convex hull were manipulated, convex hull was the perceptual variable that had the largest effect on accuracy and RT. In this L/R task, area had only a small effect on accuracy and almost no effect on RT.

In the modeling, we found it necessary to make nondecision time a function of perceptual variables in order to account for large changes in the leading edges of RT distributions in numerosity tasks when perceptual and numerosity variables were in conflict. It has been found that modeling nondecision time is needed to understand how the conflict affects information processing in different contexts (reinforcement learning, Ratcliff & Frank, 2012; the Stroop task, Fennell & Ratcliff, 2019; identifying categorical stimuli in dynamic noise, Ratcliff & Smith, 2010). In the numerosity studies presented in this article, the regression model for the nondecision time component accounted for most (but not all) aspects of the data. Because we still do not have a complete explanation of all details of the results, this account should be viewed as a progress report on this modeling approach rather than a complete solution.

The ANS diffusion model presented here is highly constrained. There are 28 conditions in each experiment when the data are grouped by the values of the perceptual variables. A traditional diffusion model with separate drift rates for the different conditions would require 28 drift rates and 28 across-trial variability in drift rate parameters to fit these data. If all the possible combinations are considered (as in maximum likelihood estimation), then for Experiments 1 and 3 there are 7 combinations of numeracy and 36 dot area conditions (252 combinations) which might lead to 504 different drift rates and across-trial variability in drift rate parameters. For Experiment 2, there are even more combinations. In the ANS diffusion model, only 10 parameters are needed for Experiment 1 and 2 and this provides a spectacular reduction in degrees of freedom in the model relative to data. The cost associated with this is the inability to modify the model to fit any single or small number of aberrant data points. Again, it is important to stress that the number of degrees of freedom in the data is large, with accuracy and correct and error RT distributions for each experimental condition (we summarize distributions with 5 quantile RTs which means there are 11 degrees of freedom per condition). So if we had data for all 252 combinations of the variables, there would be 11 times 252 for 2772 degrees of freedom for Experiment 1, and many more for Experiment 2.

The ANS diffusion model was also compared to the recent regression-based model of choice probability in numerosity judgment tasks (DeWind et al., 2015). The model predictions for accuracy showed that both of the models produce good predictions for accuracy with only a little difference. The ANS diffusion model has an advantage over the DeWind et al.'s model in that it provides a full account for choices and RT distributions from numerosity judgment tasks. This difference is important in that the RT distributions can further constrain models and lead to important findings that cannot be found from accuracy data only.

In the numeracy literature, there has been a vigorous debate about the roles of perceptual variables on numerosity judgments. It has long been understood that controlling one perceptual variable (e.g., equating it across conditions) will introduce differences in another variable. This means it is impossible to control all the variables. This issue becomes critical in the context of recent arguments in which the case has been made that the effects of numerosity are actually mediated by perceptual variables (Abreu-Mendoza & Arias-Trejo, 2015; DeWind et al., 2015; DeWind & Brannon, 2012; Feigensen et al., 2002; Gevers et al., 2016; Gebuis et al., 2016, 2017; Gebuis & Reynvoet,

2012a, 2012b, 2013; Halberda et al., 2008; Im et al., 2016; Leibovich et al., 2017). The solution that the ANS diffusion models provide is to jointly model the effects of both numerosity and perceptual variables on performance as in DeWind et al. (2015). In this way, the ANS diffusion model is used to measure the sizes of the relative effects of perceptual and numerosity variables on performance in the task.

However, there is another solution to the problem of perceptual variables. Ratcliff and McKoon (2018) showed that the impact of perceptual variables is quite different for different numerosity discrimination tasks. Our results showed that when the task is to decide which of two colors of dots is more numerous when the dots are intermingled in one array (B/Y task), the perceptual variable area has a large effect on performance (Experiment 1). When two arrays are side by side and spatially separated (L/R task), convex hull has a large effect on performance and area has a much smaller effect (Experiment 2). When there is a single array and the task is to judge whether there are more or fewer dots than a criterion number (e.g., 25), area has almost no effect on performance (Ratcliff & McKoon, 2018). In Ratcliff and McKoon's experiment, the drift rate coefficients were correlated across tasks, which shows that if individuals are good at one task, they are good at the others. This suggests that if perceptual variables are a serious concern, then a task should be chosen (e.g., the latter single-array task) that minimizes the effects of perceptual variables.

The ANS diffusion model was not able to fully account for the effects of numerosity and perceptual variables in conflict by modeling their effects on drift rate alone in our experiments. This failure represents a new conflict effect. Conflict effects are obtained in a variety of paradigms in which two variables are set in conflict with each other. In the Stroop task, a stimulus is a color word in a colored font and the task is to name the color or to name the word (Fennell & Ratcliff, 2019; MacLeod, 1991; MacLeod & Dunbar, 1988; Spieler, Balota, & Faust, 2000; Steinhauser & Hübner, 2009; Stroop, 1935). There are two dimensions in the stimulus, the word name and the word font color. These can agree with each other (e.g., BLUE in blue font color) or conflict with each other (e.g., BLUE in red font color). In the latter case, the conflict leads to less accurate and slower performance when the task is color naming. Also, Fennell and Ratcliff (2019) found that the conflict effect was associated with a shift in the RT leading edge in the Stoop task when the task was color identification and subjects responded vocally but not manually (see also Spieler et al., 2000). Similarly, in the Ericksen-Flanker task, a subject is asked to discriminate a single item flanked by items that imply the same or opposite response (Eriksen & Eriksen, 1974; Hübner, Steinhauser, & Lehle, 2010; Servant, Montagnini, & Burle, 2014; Servant, White, Montagnini, & Burle, 2015; White, Ratcliff, & Starns, 2011). For example, if a stimulus consists of arrows, the task is to decide if the central item is a right or a left arrow while the surrounding items face the same direction (e.g., > > > > >) or the opposite direction (e.g., < < > < <). As in the Stroop task, responses become slower and less accurate when the flankers are in the opposite direction.

Conflict effects have also been observed in other paradigms. For instance, Ratcliff and Frank (2012) examined the effect of high response conflict, in which two infrequently reinforced choice options are tested together. It was found that RTs were slowed and shifted in this condition relative to conditions in which two frequently reinforced options were tested together. A difference between reinforcement learning and conflict paradigms discussed above (our numerosity judgment tasks and the conflict tasks) is that it was other choice options that induced the conflict in reinforcement learning, while it was task-irrelevant variables that caused the interference effect in the conflict tasks. For example, a color word in the Stroop task, a direction of flankers in the Eriksen-Flanker task, and dot area or convex hull in a numerosity task are not directly related to targets in those tasks.

The slowdown in responding, one of the hallmarks of conflict paradigms, is often greater than would be expected by a drift-rate effect alone in a standard diffusion model. Often the empirical effect is a large change in the leading edge of the RT distribution corresponding to a delay in processing (e.g., Stroop task, Fennell & Ratcliff, 2019, and Spieler et al., 2000; Flanker task, White et al., 2011; reinforcement learning, Ratcliff & Frank, 2012). The conflict between perceptual and numerosity variables shows exactly this pattern when the task is numerosity discrimination. However, when the task is area discrimination (Experiment 4), numerosity produces no additional shift in the RT leading edge. This asymmetry is similar to the effects seen in the Stroop task in which a color word affects color identification but not vice versa (Cohen, Dunbar, & McClelland, 1990; Fennell & Ratcliff, 2019; MacLeod & Dunbar, 1988; MacLeod, 1991; Mewhort, Braun, & Heathcote, 1992). The conflict effect in a numerosity task may be worth studying further and it may be useful as another paradigm for studying conflict in processing.

The research reported in this article demonstrates the strength of integrating a cognitive theory and a mathematical model of choice and RT as in Ratcliff and McKoon (2018). By combining the ANS model for numerosity with the diffusion decision model, the integrated model was able to account for complicated behavioral patterns and the roles of numerical and perceptual variables in numerosity judgments. Similar approaches have been taken in different literature such as perceptual matching (Ratcliff, 1981), reinforcement learning (Frank et al., 2015; Pedersen, Frank, & Biele, 2017) and neural data analysis (Cavanagh, Wiecki, Kochar, & Frank, 2014; Ratcliff, Sederberg, Smith, & Childers, 2016; Turner, van Maanen, & Forstmann, 2015). Integrated models can lead to a valuable understanding of cognitive processes and provide a way to analyze behavioral data along with other variables of interest.

Acknowledgments

This work was supported by the National Institutes of Health grants R01-AG041176 and R56-AG057841-01.

Appendix A:. The diffusion decision model and fitting it to data

1.1 The diffusion decision model

The diffusion decision model is a mathematical model of the cognitive processes underlying simple two-choice decisions (Ratcliff, 1978; Ratcliff & McKoon, 2008). It simultaneously accounts for choice accuracy and reaction time distributions for both correct and error responses. A fundamental assumption of the model is that information is extracted from the stimulus and accumulated over time. A decision between

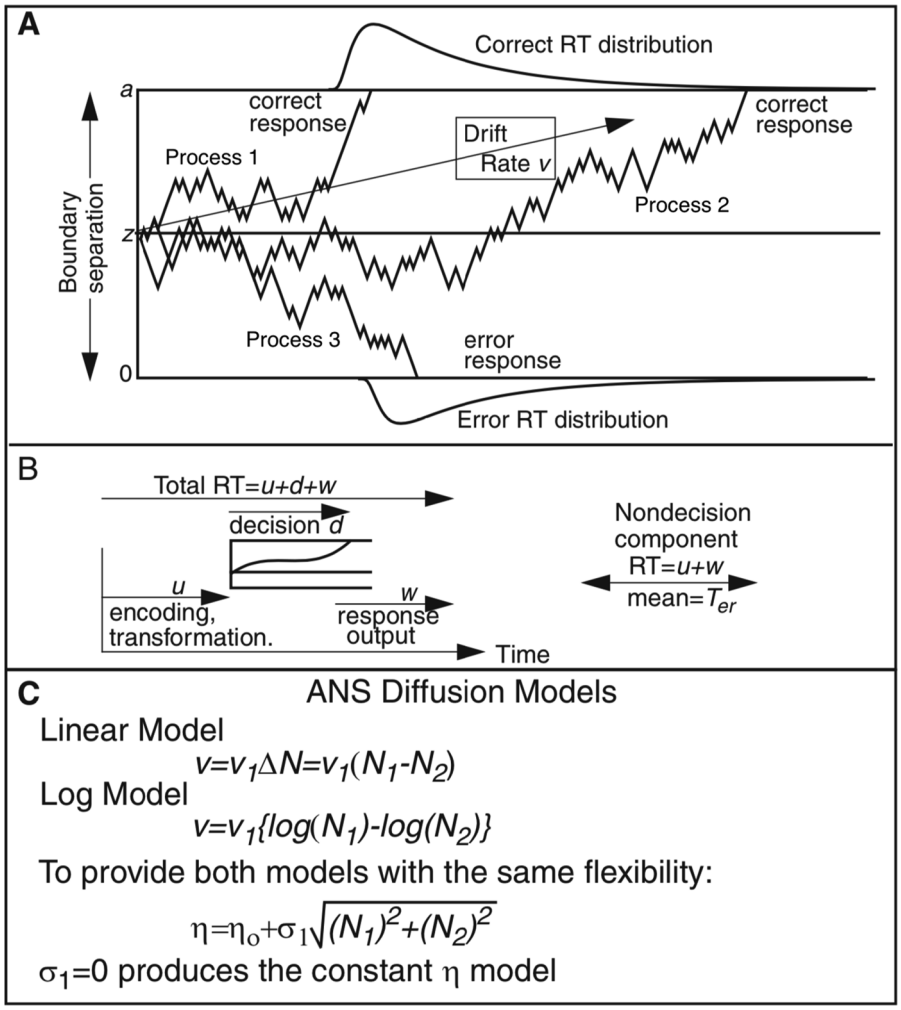

Fig. A1. Panel (A) illustrates the basic components of the diffusion model. Panel (B) shows how the reaction time is constructed in the diffusion model analysis. Panel (C) shows regression equations of the drift rate in the ANS diffusion model.

two choices is made when the accumulation process reaches one of two decision boundaries (Fig. A1). The model assumes that this accumulation process begins at a starting point z with two decision boundaries at a and 0. When the accumulation process arrives at one of those boundaries, the process is terminated and the choice corresponding to the boundary is made. The accumulation process is governed by the mean drift rate v, which represents the quality of information accumulation. This process is noisy so that decision processes with the same mean drift rate may terminate at different times, sometimes even at different boundaries, which produces reaction time distributions of both correct and error responses. Three lines in Fig. A1 (A) indicate accumulating processes with the same drift rate (Processes 1–3). All the lines have zig-zag fluctuations which represent that the information accumulation is noisy. Reaction times for decision making consist not only of the time for the decision process, but also of the time for external processes such as encoding stimuli, transforming the stimulus representation to produce the decision variable, and producing output. The time for these processes outside the decision process is modeled as the nondecision time component T_{er}. Thus, RT in a decision making task is determined as a sum of the decision time (DT) and the nondecision time $(RT = DT + T_{er})$.

The diffusion model also assumes that the value of each component underlying processing varies across trials (Ratcliff, 1981; Ratcliff & McKoon, 2008; Ratcliff & Tuerlinckx, 2002; Ratcliff, Van Zandt, & McKoon, 1999). Drift rate, starting point, and nondecision time are all assumed to vary across trials which implements the assumption that processing cannot be set back to exactly the same value on identical trials (cf. Signal Detection Theory). The drift rate is assumed to be normally distributed with standard deviation η, while the starting point and the nondecision time are uniformly distributed with ranges s_z and s_t, respectively. In fact, it is this assumption of the across-trial variability in model components that gives the model the ability to explain the relative behaviors of correct and error RTs. Across-trial variability in drift rate enables the model to predict error reaction times slower than correct reaction times and across-trial variability in the starting point allows the model to produce error reaction times faster than correct reaction times (Ratcliff & Rouder, 1998; Ratcliff & McKoon, 2008).

The diffusion model decomposes the processing in two-choice decision tasks into separate components of cognitive processing. These components are represented by the model parameters introduced above: the drift rate represents the quality of information processing, boundary separation describes the amount of information required to make a decision, the starting point represents the initial bias in a decision-making process, and nondecision time is the time consumed for the processes outside the decision process.

1.2 Fitting the ANS diffusion model

There are several methods to fit the diffusion model to data from two simple choice tasks, and these can be directly used to fit ANS diffusion models (Ratcliff & Tuerlinckx, 2002; Tuerlinckx, 2004; Ratcliff & Childers, 2015). All the estimation methods require the response probability $P(\cdot)$ and the cumulative distribution function $G(\cdot)$ of the diffusion process described below (Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002).

$$P(X=0;\,\xi,\,\zeta)=\left(e^{-2\xi a/s^2}-e^{-2\xi\zeta/s^2}\right)/\left(e^{-2\xi a/s^2}-1\right)$$

$$G_{X,T}(0,\,t)=P(X=0,\,T\le t;\,\xi,\,\zeta)$$

$$=P(X=0;\,\xi,\,\zeta)-\frac{\pi s^2}{a^2}e^{-(\xi\zeta/s^2)}\times\Sigma_{n=1}^{\infty}\frac{2nsin(n\pi\zeta/a)e^{-\frac{1}{2}(\xi^2/s^2+n^2\pi^2s^2/a^2)(t-\tau)}}{(\xi^2/s^2+n^2\pi^2s^2/a^2)}$$

Here random variables X and T represent a choice and RT, respectively. For the choice variable X, $X=0$ if the accumulation process terminates at the lower boundary and $X=1$ if the accumulation process terminates at the upper boundary. Also, ξ, ζ, and τ indicate the drift rate, the starting point, and the nondecision time for a trial. a is the boundary separation and s is within-trial variability in the evidence accumulation. The within-trial variability parameter s is a scaling parameter and usually fixed at some constant (e.g., 0.1). Note that $P(X=0;\,\xi,\,\zeta)$ above represents a probability of choosing the lower boundary. Similarly, $G_{X,T}(0,\,t)=P(X=0,\,T\le t;\,\xi,\,\zeta)$ is the cumulative probability function, which produces a probability of choosing a response corresponding to the lower boundary before time t. From these two formulae, we can also obtain $P(X=1;\,\xi,\,\zeta)$ and $G_{X,T}(1,\,t)=P(X=1,\,T\le t;\,\xi,\,\zeta)$, each of which indicates a probability of choosing the upper boundary at the end of the accumulation process and a probability of choosing a response corresponding to the upper boundary before time t, respectively, by changing ξ to $-\xi$ and ζ to $a-\zeta$.

To fit the full diffusion model with across-trial variabilities, distributions of these variables need to be taken into account as part of the likelihood function. Therefore, $G_{X,T}(\cdot,\cdot)$ is integrated over these distributions as follows. The three distributions can be denoted as $\xi\sim N(v,\,\eta)$, $\zeta\sim U\left(z-\frac{s_z}{2},\,z+\frac{s_z}{2}\right)$, and $\tau\sim U(T_{er}-\frac{s_t}{2},\,T_{er}+\frac{s_t}{2})$ where v, z, and T_{er} are the means of the drift rate, the starting point, and the nondecision time, respectively, η is the variance of the drift rate, and s_z and s_t are the ranges of the starting point and the nondecision time, respectively. Then, the cumulative distribution function of the error reaction time in the diffusion model is calculated as follows (Tuerlinckx, 2004).

$$F_{X,T}(0,\,t)=\int_{T_{er}-\frac{s_t}{2}}^{T_{er}+\frac{s_t}{2}}\int_{z-\frac{s_z}{2}}^{z+\frac{s_z}{2}}\int_{-\infty}^{\infty}G_{X,T}(0,\,t;\,\xi,\,\zeta)N(v,\,\eta)U\left(z-\frac{s_z}{2},\,z+\frac{s_z}{2}\right)U\left(T_{er}-\frac{s_t}{2},\,T_{er}+\frac{s_t}{2}\right)d\xi d\zeta d\tau$$

In many cases, the diffusion model is fitted to binned data which consists of response proportions and RT quantiles. In order to obtain binned data, RT data is divided into quantiles, by correct and error responses separately, and by experimental conditions. The number of quantiles can be defined by users, but five quantiles, 0.1, 0.3, 0.5, 0.7, and 0.9 are often used. This produces 6 bins, each having fixed proportions between the quantiles (0.1, 0.2, 0.2, 0.2, 0.2, and 0.1). Across correct and error responses, there are 12 bins and these bins are obtained separately for each condition.

Two different estimation methods can be exploited to fit the model to the binned data. The first is the chi-square (χ^2) method, which finds a set of parameters that minimizes the following quantity (Ratcliff & Childers, 2015; Ratcliff & Smith, 2004; Ratcliff & Tuerlinckx, 2002; Tuerlinckx, 2004).

$$\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$$

Here the summation is taken over all bins, correct and error responses, and all the experimental conditions used. To calculate this value, first, the diffusion model prediction of the response proportions in the bins is computed by taking sequential differences in the cumulative distribution functions evaluated at the RT quantile values. Then, the number of observations in each bin is multiplied by the observed and the predicted proportions. This produces observed frequencies (O_i) and predicted frequencies (E_i), which are used to calculate χ^2. Minimizing the χ^2 statistic by an optimization algorithm such as SIMPLEX (Nelder & Mead, 1965) will produce the best parameter estimates.

Another method that uses binned data is the multinomial likelihood (G^2) estimation (Ratcliff & Smith, 2004). This method

estimates parameters by minimizing the G^2 statistic $2\Sigma Np_i\log(p_i/\pi_i)$. Note that the observed proportions (p_i) and their corresponding predictions (π_i) can be obtained as in the χ^2 method. Also, the sum is over all bins from both correct and error responses of all conditions. It is widely known that the G^2 statistic asymptotically follows the chi-square distribution. For this reason, when the sample size is not too small, the G^2 method produces a very similar result to that of the chi-square method.

For the binning methods, the degree of freedom (*df*) can be calculated from the number of conditions and the number of RT bins. For example, if response proportions are obtained between and outside of 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles, there would be 6 proportions of correct responses and 6 proportions of errors. Because the response proportions should sum to 1, the number of *df* from these RT bins from this one condition is 12–1 = 11. Then, this is multiplied by the number of conditions to give the number of *df* in the data. A model's *df* can be calculated by subtracting the number of parameters from the *df* from data. For example, if there are 28 conditions and the number of parameters is 11, the number of *df* from data is 28*11 = 308 and the model *df* is 308–11 = 297.

The chi-square and G^2 methods use binned RTs and this has advantages and disadvantages. The advantage is that it is robust to contaminants (Ratcliff & Tuerlinckx, 2002) and the disadvantage is that it does not use the information contained in all RTs. Maximum likelihood estimation (MLE) uses all the individual data points (Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002) and if there are no contaminants or their effects have been minimized, then the method is more efficient and produces smaller standard errors in model parameters than the binned methods. In order to compute the likelihood, a probability density at each data point (x_i, t_i) is evaluated. To do so, $F_{X,T}(x_i, t_i)$ and $F_{X,T}(x_i, t_i + \Delta t)$ are computed for each RT t_i with Δt set to a small value (e.g., 0.00001 ms). Then the probability density $f(x_i, t_i)$ can be approximated by $f(x_i, t_i) = (F_{X,T}(x_i, t_i + \Delta t) - F_{X,T}(x_i, t_i))/\Delta t$. Multiplying this value for each RT produces the likelihood. The model parameters are adjusted using a minimization method to find the model parameters that produce the maximum likelihood (by minimizing the minus log likelihood) and hence the best fit to the data by this criterion.

Often, binning methods have been preferred over the MLE. First of all, using quantile data has a benefit with respect to computing time. When the sample size is large, the MLE is time-consuming because it requires the evaluation of densities at all the trials separately. However, the binning methods simply use (10) quantile RTs computed per each condition regardless of the sample size. In spite of this data-reduction, the chi-square method performs nearly as well as the MLE method (Ratcliff & Childers, 2015). As noted above, the binning methods are more robust to extreme values and outliers. This is because the quantiles are less affected by several single points. For example, suppose we increase all the RT values greater than the 0.9 quantiles by three times. This would produce a large change in a likelihood value, however, the 0.9 quantiles and the estimation result from the binning methods would remain unchanged. If data do not contain extreme contaminants, MLE will produce better estimates than binned methods with smaller standard errors in the model parameters (Ratcliff & Tuerlinckx, 2002).

There is an important limitation of the binning method when the number of observations per condition is small. Specifically, when there is a large number of conditions and each condition has only a few observations, quantiles cannot be calculated reliably and this can produce unreliable parameter estimation. In particular, error quantiles may not be able to be computed when accuracy is high (e.g., 0.9–0.95) and the number of observations is small. For example, if there is a condition in which accuracy is 0.9 and the number of observations is 20, there are only two error observations and so the error quantiles cannot represent the target error RT distribution. In typical experiments, the number of conditions is not that large and so there is no problem. However, in the experiments in this study, we manipulate not only numerosity but also several perceptual variables. This produces a large number of conditions. For example, in Experiment 1, there are 7 numerosity pairs and 6 dot area values (for each of the two dot groups) considered in stimuli, which produces 7 * 6 * 6 = 252 different conditions. Given that the total number of observations in perceptual decision-making tasks for a single session is usually 500–2000, modeling this large number of conditions cannot avoid problems that result from a small sample size for each condition. For binning methods, it is likely that the quantiles in some conditions cannot be computed (especially for errors). In order to circumvent this, it is possible to aggregate data across conditions to produce a larger number of observations for correct and error responses in each condition. In the main analysis of the current study, we group some conditions carefully by their similarity in congruency between numerosity pairs and a major perceptual feature of a task so as not to average over important differences among conditions in the data. This provides a reasonable sample size for each condition for the binning method.

For the current study, we mainly employ the G^2 method following Ratcliff and McKoon (2018). After the presentation of the G^2 results for each of the experiments, we also present results from MLE and find similar model parameters as those from the G^2 method. This way of performing analyses has the advantage that the G^2 method can produce fits to (say 30) individual subjects in 10–30 min and so models can be explored. Once a model has been evaluated, the MLE method can be applied with G^2 parameter estimates as initial values. Usually, MLE (to fit the same number of subjects) takes several hours or even days to find the best set of parameters. However, implementing G^2 parameter estimates as initial values for MLE reduces the computation time dramatically to 2–3 h.

For each analysis in this article, G^2 values (or minus log likelihood for MLE results) averaged across subjects are provided as a goodness of fit measure, but only for descriptive purposes. Quantile-probability plots are also reported to compare observed data and model predictions (Ratcliff & McKoon, 2018). To generate this plot, observed correct and error response proportions and quantile RTs are calculated for each subject for each condition and these are averaged across subjects. The predictions are computed by producing each individual's predicted correct and error response proportions and quantile RTs and then averaging them over subjects. Then, for

both observations and predictions, RT quantiles are plotted against response proportions. This shows how well the predictions can reproduce behavioral patterns observed from the data. Along with these, individual fits of accuracy and 0.1, 0.5, and 0.9 quantile RTs are reported in the Appendix C (Fig. C1).

Appendix B. DeWind et al. (2015)'s model and its theoretical comparison to the ANS diffusion model

In this appendix, DeWind et al. (2015)'s model, which was briefly introduced in the body of the article, is reviewed and compared to the ANS diffusion model. In the development of their model, DeWind et al. first claimed that, even though numeric and perceptual features in a visual stimulus are all confounded, the variables affecting numerosity judgment have three degrees of freedom. Thus, it is possible to find a three-dimensional space in which all the features in the stimulus can be defined. This space is represented by three axes of *numerosity*, *size,* and *spacing*. The numerosity axis indicates the true number of objects and the other two variables are defined as follows (DeWind et al., 2015; Starr, DeWind, & Brannon, 2017).

$$\log_2(size) = \log_2(total\ surface\ area) + \log_2(item\ surface\ area)$$

$$\log_2(spacing) = \log_2(field\ area) + \log_2(sparsity)$$

Item surface area represents the area covered by each item in a stimulus and the total surface area is the number of items multiplied by the item surface area. Sparsity is the inverse of the density or the average convex hull per each item (DeWind et al. used the term field area to refer to the convex hull area). DeWind et al. claimed that size and spacing are orthogonal to each other and also to numerosity, meaning that the three dimensions defined by these three variables are independent.

According to DeWind et al., numerosity judgments can be modeled as a function of these three variables because all the features of a stimulus can be represented on the three dimensional feature space. For the numerosity judgment task with two side-by-side arrays of dots (i.e., the L/R task), the following probability model of choosing the right side was proposed.

$$P(ChooseRight) = (1-\gamma)\left(\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\beta_{side} + \beta_{num}\log_2(r_{num}) + \beta_{size}\log_2(r_{size}) + \beta_{spacing}\log_2(r_{spacing})}{\sqrt{2}}\right)\right] - \frac{1}{2}\right) + \frac{1}{2}$$

This can be rewritten as follows.

$$P(ChooseRight) = (1-\gamma)\left(\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log_2(r_{num}) - \left(\frac{-\beta_{side} - \beta_{size}\log_2(r_{size}) - \beta_{spacing}\log_2(r_{spacing})}{\beta_{num}}\right)}{\sqrt{2}\frac{1}{\beta_{num}}}\right)\right] - \frac{1}{2}\right) + \frac{1}{2}$$

Here r_{num}, r_{size}, and $r_{spacing}$ indicate ratios of numerosity, size, and spacing of the stimulus on the right side to the stimulus on the left side. β_{side} is a bias for the two response options and the other β's are slope coefficients corresponding to the numerosity, spacing, and size variables. γ is a guessing term representing the probability that a subject is distracted momentarily during a task and makes a choice based on guessing. The erf function is the error function: $\mathrm{erf}(x) = 2\Phi(\sqrt{2}x) - 1 = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

The model is an extension of the previous model proposed by Piazza et al. (2004, 2010):

$$P(ChooseRight) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\log_2(r_{num})}{\sqrt{2}(\sqrt{2}w)}\right)\right)$$

Compared to Piazza et al.'s model, DeWind et al.'s model is able to account for the effects of non-numeric features on the choice probability by adding the size and spacing variables into the regression equation. Thus, contributions from numerosity information on numerosity judgments can be evaluated independently of those from other non-numeric features in a stimulus.

Piazza et al.'s model is based on the assumption that numerosity information can be represented with a normal distribution on the log scale. Also, numerical acuity (ANS acuity) can be estimated from the standard deviation of the distribution, which they called the internal Weber fraction, w. In the L/R task, as two numerosities are compared, the distribution of the difference between two numbers should be processed. This distribution has the standard deviation $\sigma = \sqrt{2}w$ and so the internal Weber fraction is estimated as $w = \frac{\sigma}{\sqrt{2}}$. In the same way, in DeWind et al.'s model, $\sigma = \frac{1}{\beta_{num}}$ and the acuity in estimating numerosity can be estimated as $w = \frac{\sigma}{\sqrt{2}} = \frac{1}{\sqrt{2}\beta_{num}}$.

Although DeWind et al.'s model and the ANS diffusion model share some features, they have crucial differences. Most importantly, DeWind et al.'s model is constrained to account only for accuracy, while the ANS diffusion model is constrained by both

accuracy and RT distributions. As discussed in the main text, any analysis method which does not fully consider both of the measures is incomplete.

In examining the effects of perceptual features in a visual stimulus, DeWind et al.'s model used pre-defined variables, namely size and spacing. They defined these variables to be orthogonal to numerosity but it is not easy to find a meaningful interpretation of these variables. For example, size is defined as a product of the total surface area and the item surface area, but the total surface area is equal to the item surface area multiplied by the number of dots. Thus, size is the product of the number of dots and the squared item surface area. It is not clear how to interpret this variable and what it means in numerosity judgment tasks.

Furthermore, it is not clear if these variables (size and spacing) are really orthogonal to numerosity. To define these variables, they first fixed the numerosity and found axes which are perpendicular to *iso*-numerosity lines (lines connecting the same numerosity in the space of intrinsic and extrinsic perceptual variables). Thus, these axes would be orthogonal to numerosity (see Fig. 1C-E in DeWind et al., 2015). However, the algebraic definition they provided does not match this property. Specifically, because total surface area is a product of item surface area and the number of items and sparsity is a field area divided by the number of items, size and spacing can be represented as below.

$$log_2(size) = log_2(n) + 2 * log_2(item\ surface\ area)$$

$$log_2(spacing) = -log_2(n) + 2 * log_2(field\ area)$$

If the variables are orthogonal, a change in one variable should not affect another variable. However, the expressions above show that there is a linear relationship between log numerosity, log size, and log spacing. Conceptually, there could be variables orthogonal to numerosity as they claimed. However, it is doubtful that size and spacing as defined in DeWind et al. have this property.

In contrast, the ANS diffusion model uses direct measures of non-numeric perceptual features in the regression model of the drift rate. This approach has many advantages over using orthogonal variables because the interpretation of the effects of each variable is possible. For example, the effect of dot area on the drift rate can be directly measured as its drift rate coefficient. Also, different representations of the cognitive components, particularly drift rate which contributes to the choice probability, can be built and compared in the ANS diffusion model framework. As in Ratcliff and McKoon (2018), the log and linear representations of the effect of numerosity on drift rate can be compared. As in the current article, drift rate can be assumed to be a linear combination of linear or log functions of numerosity and perceptual features and interaction between the variables can also be examined. The approach of DeWind et al. (2015) does not allow such evaluation because, for example, the interaction of numerical and perceptual features cannot be explored since the variables were designed to be independent.

The two methods also differ in evaluating relative contributions of the variables. In DeWind et al. (2015), the coefficients of their variables were directly compared and it was shown that the slope for numerosity was much larger than the other coefficients. However, each independent variable, numerosity, size, and spacing has a different scale and so the evaluation based on coefficients was not free from the scale problem. For example, the coefficient for the variable with larger values tends to be small, but it does not mean that the variable has a small effect on the dependent variable. On the other hand, in the ANS diffusion model analysis presented in this paper, we computed ranges of the variables multiplied by their coefficients. By doing so, the relative contribution of the variables in the experiments can be compared on the same scale.

The models can also be compared in another aspect of measuring the numerical acuity. DeWind et al.'s model estimates the numerical acuity as the standard deviation (i.e., internal Weber fraction) of the normal distribution that represents the log numerosity. The smaller the standard deviation is, the narrower the internal numerosity distribution is, and so more accurate numerosity discrimination can be made. However, the model with a constant standard deviation of the numerosity distribution has a conceptual problem (Rouder & Geary, 2014; Ratcliff & McKoon, 2018). Because the distribution is assumed to be normal, the model can represent a number as negative in some cases. Although the proportion of negative numbers might be small, it is nonzero and if the mean of the numerosity distribution is small with a constant standard deviation, the proportion of negative numbers can be high. In contrast, the modeling approach presented in this paper models drift rate instead of the Weber fraction or the internal Weber fraction. Because the drift rate drives a process of numerosity judgment, this can be used as a measure of numerical acuity. Also, because drift rate can be negative, this measure is free from the scaling issue. Under the ANS diffusion model framework, drift rate can be easily calculated over different conditions of numerosity and perceptual features, given the drift rate coefficients.

Appendix C. Model fits of individual subjects and conditions

The quantile probability plots presented in the body of the article show general patterns in data and predictions. But they do not

show how well the model accounts for the behavior of individual subjects across different conditions. Figs. C1 and C2 show the model prediction against data for each condition, for each subject, and for Experiment 1 and 2 (Model 12 in Table 3 for Experiment 1 and Model 3 in Table 8 for Experiment 2). The small bars on the bottom right represent plus or minus 2 standard deviations in the data. For accuracy, this was calculated using the standard deviation formula for Bernoulli probability. For quantiles, a bootstrapping method was used. Reaction times in each condition were randomly resampled with replacement and their quantiles were calculated. This was repeated and the standard deviations of quantiles across repetitions were calculated. The standard deviation was obtained separately for each individual data point and then averaged. The good agreement between data and theory (prediction) illustrates that the ANS diffusion model can explain numerous conditions with only a few parameters. The model prediction in Experiment 3 and 4 showed similar patterns.

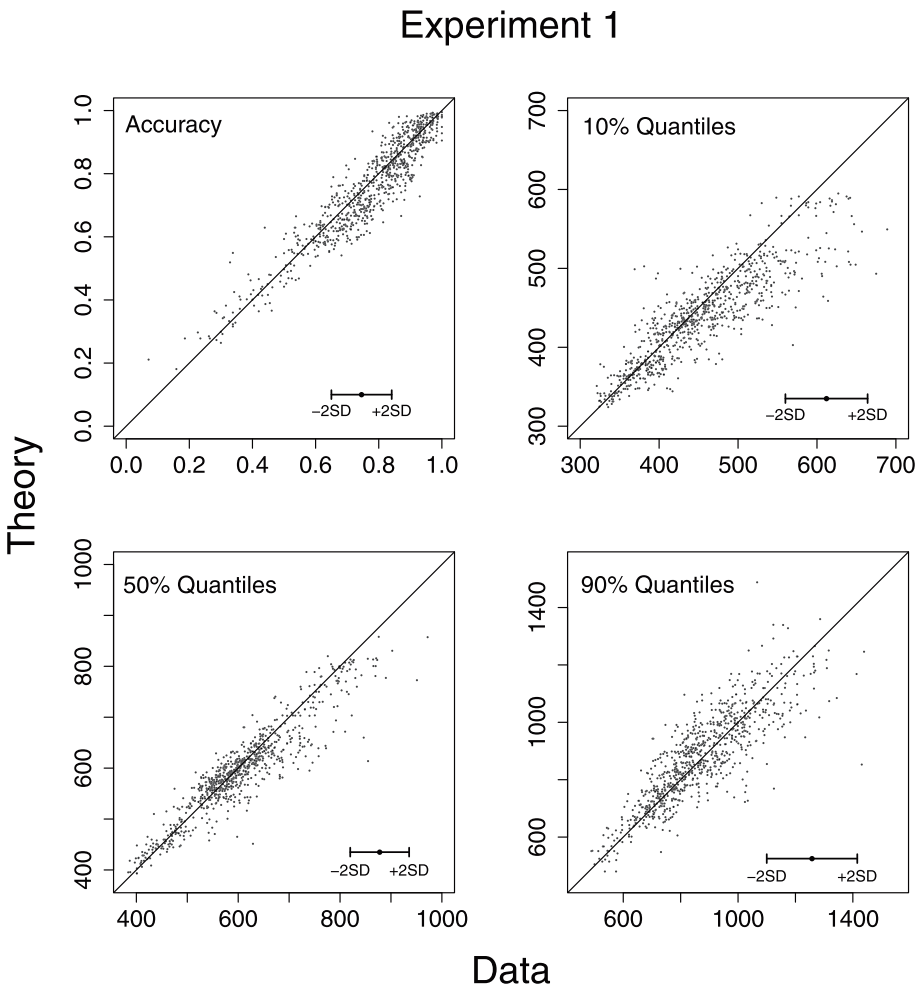

Fig. C1. Model prediction for accuracy and the 0.1, 0.5, and 0.9 quantile RTs in Experiment 1 plotted against data for each condition and for each subject. The number of subjects is 29 and the number of conditions is 28.

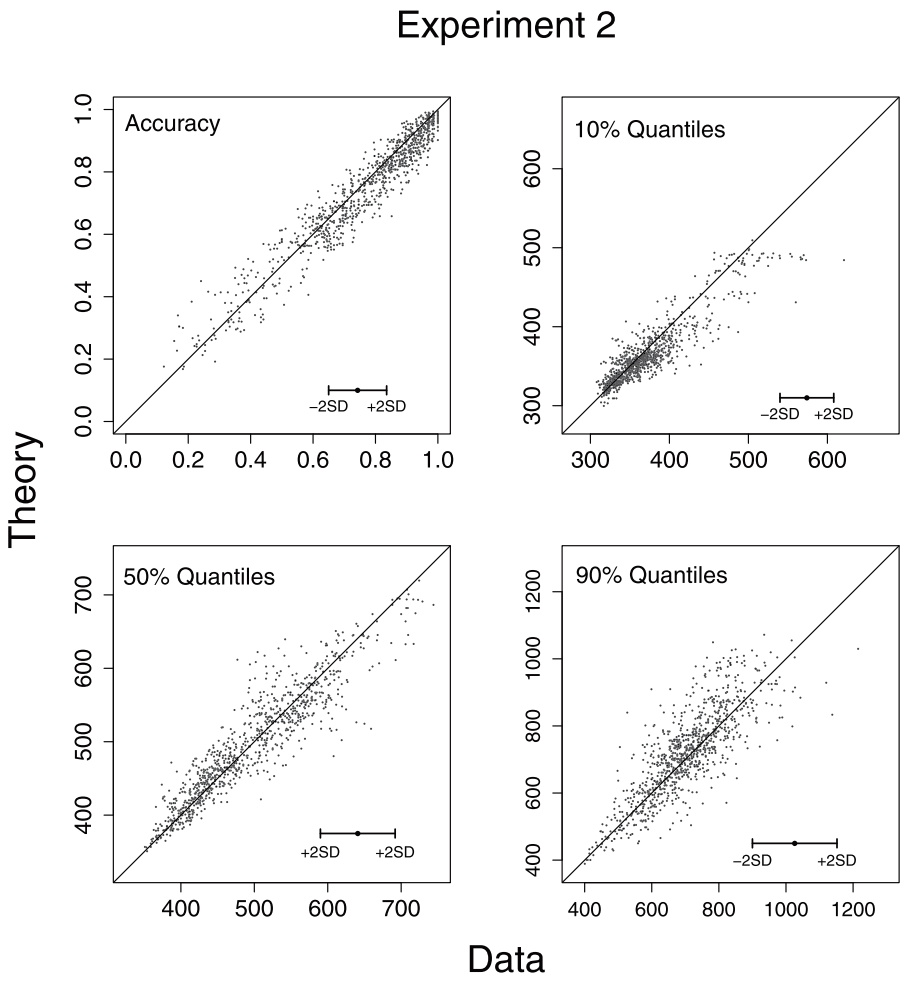

Fig. C2. Model prediction for accuracy and the 0.1, 0.5, and 0.9 quantile RTs in Experiment 2 plotted against data for each condition and for each subject. The number of subjects is 34 and the number of conditions is 28.

Appendix D. Parameter recovery for the ANS diffusion model

In this Appendix, we present parameter recovery for the ANS diffusion model. Often (but not always, see Ratcliff, 2014), the diffusion model is fit to experimental data with several conditions with several hundred observations per condition and with drift rate freely estimated for each condition. In the application presented here, as the conditions are grouped with finer and finer divisions among conditions (e.g., 28 conditions in Experiment 1), there are fewer observations per condition.

For the parameter recovery study, we generated model parameters from Model 12 (the most successful model) in Experiment 1 and used these to produce simulated data. Parameter values were randomly sampled from uniform distributions with their ranges determined by the ranges from individual differences in the parameters from the fits to the data. We examined the recovery with two sample size conditions, namely, N = 100 and N = 500 for each of the 28 conditions in Experiment 1. The model was fit to the

Table D1
Parameter recovery: correlations between the parameters and the estimates from the model fit to the simulated data. a: boundary separation, s_z: across-trial variability in starting point, s_t: across-trial variability in nondecision time, v_i's: drift rate coefficients, t_0 and t_1: intercept and slope of the nondecision time regression, and η_0 and σ_1: intercept and slope of the across-trial variability in drift rate. MAD: mean absolute deviation between the parameter values and the corresponding estimates. SD: standard deviation of the absolute deviation.

	a	t_0	η_0	s_z	s_t	v_1	v_2	v_3	σ_1	t_1
N = 100	0.851	0.987	0.747	0.612	0.994	0.898	0.761	0.888	0.784	0.935
MAD	0.011	0.008	0.065	0.023	0.010	0.005	0.127	0.0005	0.108	0.014
SD	0.008	0.006	0.051	0.016	0.009	0.003	0.103	0.0004	0.097	0.012
N = 500	0.942	0.996	0.900	0.807	0.998	0.964	0.903	0.960	0.919	0.984
MAD	0.004	0.004	0.026	0.010	0.005	0.002	0.115	0.0002	0.052	0.007
SD	0.004	0.004	0.026	0.009	0.004	0.002	0.071	0.0002	0.050	0.005

simulated data using the G^2 method. This procedure was repeated 100 times for each of the sample size conditions.

Table D1 shows correlations between the estimates and the parameter values. When N = 100, the parameters related to the main cognitive components such as boundary separation, nondecision time, and drift rate coefficients have high correlations (0.851–0.987) except for the drift rate coefficient for the moderated numerosity difference (0.761). Across-trial variability parameters have a lower

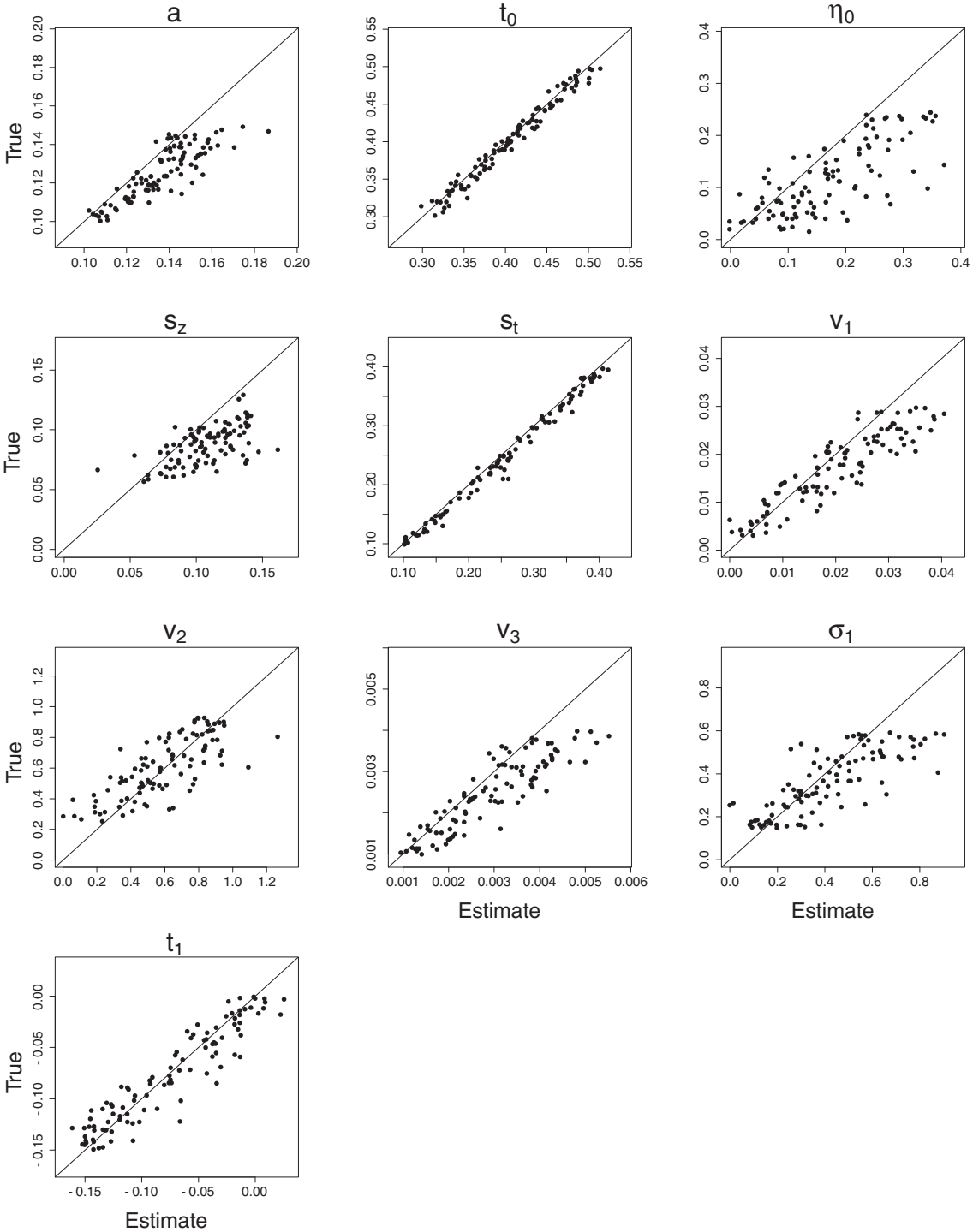

Fig. D1. Parameter recovery of the ANS diffusion model (N = 100). a: boundary separation, s_z: across-trial variability in starting point, s_t: across-trial variability in nondecision time, v_i's: drift rate coefficients, t_0 and t_1: intercept and slope of the nondecision time regression, and η_0 and σ_1: intercept and slope of the across-trial variability in drift rate.

but acceptable level of correlations. In particular, s_z is relatively difficult to estimate, which has been pointed out in previous literature (e.g., Ratcliff & Tuerlinckx, 2002). When N = 500, parameter recovery is improved with most of the correlations higher than 0.9. This large sample size corresponds to experiments with multiple sessions.

Figs. D1 and D2 show the scatter plots of the parameters when N = 100 and N = 500. The estimates are plotted on the x-axis

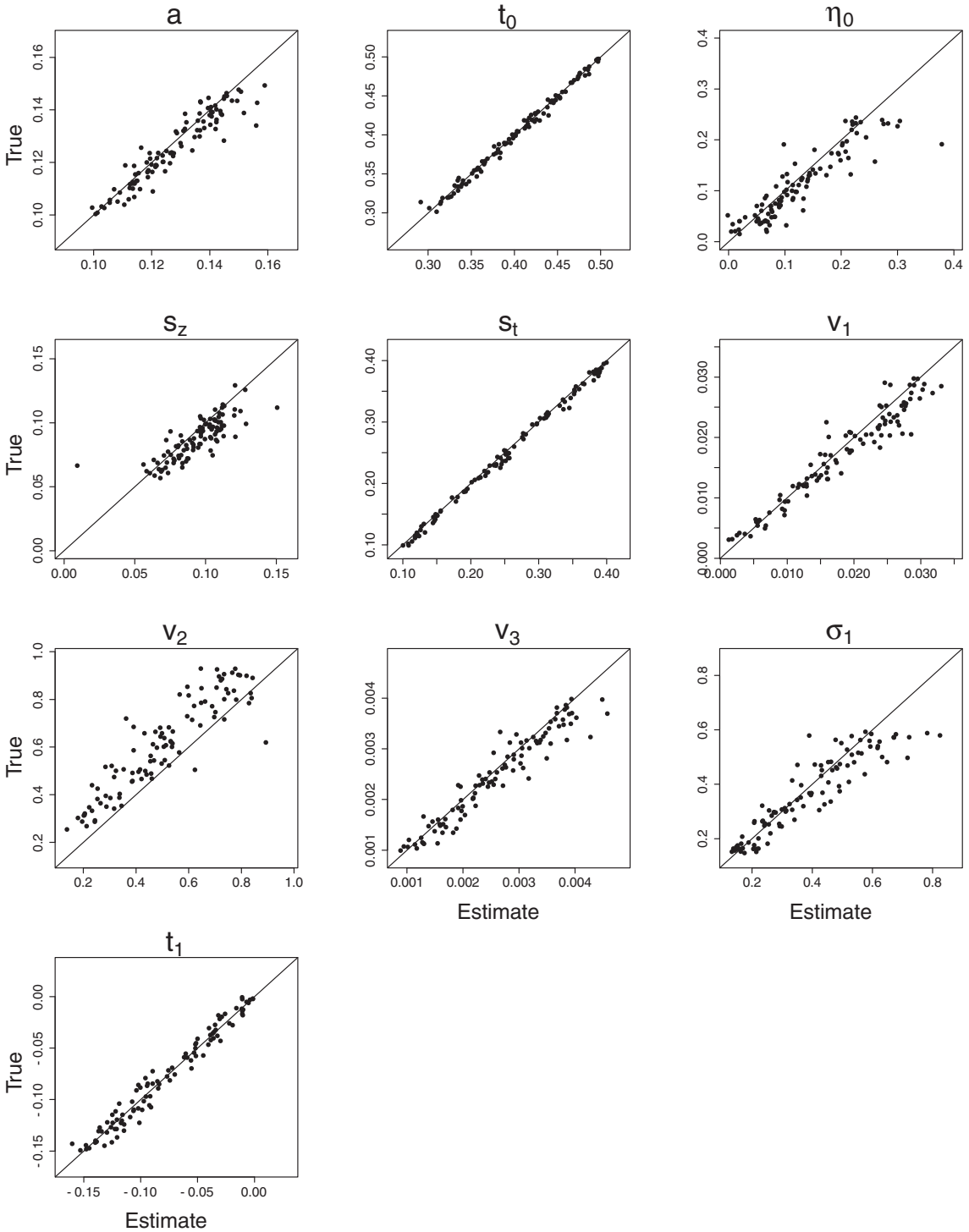

Fig. D2. Parameter recovery of the ANS diffusion model (N = 500). a: boundary separation, s_z: across-trial variability in starting point, s_t: across-trial variability in nondecision time, v_i's: drift rate coefficients, t_0 and t_1: intercept and slope of the nondecision time regression, and η_0 and σ_1: intercept and slope of the across-trial variability in drift rate.

while the parameter values used to generate the simulated data are plotted on the y-axis. When N = 100, there are biases in the estimates for some parameters such as a, η, and s_z. However, most of these biases are reduced or eliminated when N = 500. One exception is that v_2 is slightly underestimated. In general, parameter recovery results show that the model is reasonably well estimated for parameters like those that come from fits to experimental data.

Appendix E. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2020.101288.

References

Abreu-Mendoza, R. A., & Arias-Trejo, N. (2015). Numerical and area comparison abilities in Down syndrome. *Research in Developmental Disabilities, 41–42*, 58–65. https://doi.org/10.1016/j.ridd.2015.05.008.

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General, 143*(4), 1476–1488. https://doi.org/10.1037/a0035813.

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica, 148*, 163–172. https://doi.org/10.1016/j.actpsy.2014.01.016.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332–361. https://doi.org/10.1037/0033-295X.97.3.332.

Dehaene, S. (1997). *The Number Sense: How the mind creates mathematics.* Oxford University Press.

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number. *Trends in Cognitive Sciences, 7*(4), 145–147. https://doi.org/10.1016/s1364-6613(03)00055-x.

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

Dehaene, S., Izard, V., & Piazza, M. (2005). Control over non-numerical parameters in numerosity experiments. Unpublished Manuscript, <www.Unicog.Org>.

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2012). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology, 82*(1), 64–81. https://doi.org/10.1348/2044-8279.002002.

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.

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in Human Neuroscience, 6*. https://doi.org/10.3389/fnhum.2012.00068.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*, 143–149.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science, 13*, 150–156. https://doi.org/10.1111/1467- 9280.00427.

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.

Fennell, A., & Ratcliff, R. (2019). Does Response Modality Influence Conflict? Modelling Vocal and Manual Response Stroop Interference. Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication. http://dx.doi.org/10.1037/xlm0000689.

Ferreira, F.d. O., Wood, G., Pinheiro-Chagas, P., Lonnemann, J., Krinzinger, H., Willmes, K., & Haase, V. G. (2012). Explaining school mathematics performance from symbolic and nonsymbolic magnitude processing: Similarities and differences between typical and low-achieving children. *Psychology & Neuroscience, 5*(1), 37–46. https://doi.org/10.3922/j.psns.2012.1.06.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *Journal of Neuroscience, 35*(2), 485–494. https://doi.org/10.1523/JNEUROSCI.2036-14.2015.

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

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*(2), 59–65. https://doi.org/10.1016/s1364-6613(99)01424-2.

Gebuis, T., & Reynvoet, B. (2012a). 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.

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

Gebuis, T., & Reynvoet, B. (2013). The Neural Mechanism Underlying Ordinal Numerosity Processing. *Journal of Cognitive Neuroscience, 26*(5), 1013–1020. https://doi.org/10.1162/jocn_a_00541.

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., Cohen Kadosh, R., & Gevers, W. (2017). Why try saving the ANS? An alternative proposal. *Behavioral and Brain Sciences, 40*, E171. https://doi.org/10.1017/S0140525X16002107.

Gevers, W., Cohen Kadosh, R., & Gebuis, T. (2016). Sensory integration theory: An alternative to the approximate number system. In A. Henik (Ed.). *Continuous issues in numerical cognition* (pp. 405–418). Academic Press, San Diego.

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

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition, 115*(3), 394–406. https://doi.org/10.1016/j.cognition.2010.02.002.

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–U662. https://doi.org/10.1038/nature07246.

Hübner, R., Steinhauser, M., & Lehle, C. (2010). A dual-stage two-phase model of selective attention. *Psychological Review, 117*(3), 759–784. https://doi.org/10.1037/a0019471.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103*(1), 17–29. https://doi.org/10.1016/j.jecp.2008.04.001.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92–107. https://doi.org/10.1016/j.cognition.2013.12.007.

Im, H. Y., Zhong, S.-H., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research, 126*,

291–307. https://doi.org/10.1016/j.visres.2015.08.013.
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*(6), 1222–1229. https://doi.org/10.3758/s13423-011-0154-1.
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*, 1–16. https://doi.org/10.1017/s0140525x16000960.
Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science, 14*(6), 1292–1300. https://doi.org/10.1111/j.1467-7687.2011.01080.x.
Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability? *Learning and Individual Differences, 25*, 126–133. https://doi.org/10.1016/j.lindif.2013.02.001.
Logan, G. D. (1980). Attention and Automaticity in Stroop and Priming Tasks: Theory and Data. *Cognitive Psychology, 12*, 523–553.
Lonnemann, J., Linkersdoerfer, J., Hasselhorn, M., & Lindberg, S. (2011). Symbolic and non-symbolic distance effects in children and their connection with arithmetic skills. *Journal of Neurolinguistics, 24*(5), 583–591. https://doi.org/10.1016/j.jneuroling.2011.02.004.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203. https://doi.org/10.1037/0033-2909.109.2.163.
MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 126–135. http:// dx.doi.org/10.1037/0278-7393.14.1.126.
Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Preschoolers' Precision of the Approximate Number System Predicts Later School Mathematics Performance. *Plos One, 6*(9), https://doi.org/10.1371/journal.pone.0023749.
Mewhort, D. J., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop Task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 872–882. https://doi.org/10.1037/0096-1523.18.3.872.
Nieder, A., & Miller, E. K. (2003). Coding of Cognitive Magnitude: Compressed Scaling of Numerical Information in the Primate Prefrontal Cortex. *Neuron, 37*(1), 149–157. https://doi.org/10.1016/S0896-6273(02)01144-3.
Nelder, J. A., & Mead, R. (1965). A SIMPLEX-METHOD FOR FUNCTION MINIMIZATION. *Computer Journal, 7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308.
Park, J., Bermudez, V., Roberts, R. C., & Brannon, E. M. (2016). Non-symbolic approximate arithmetic training improves math performance in preschoolers. *Journal of Experimental Child Psychology, 152*, 278–293. https://doi.org/10.1016/j.jecp.2016.07.011.
Park, J., & Brannon, E. M. (2013). Training the Approximate Number System Improves Math Proficiency. *Psychological Science, 24*(10), 2013–2019. https://doi.org/10.1177/0956797613482944.
Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition, 133*(1), 188–200. https://doi.org/10.1016/j.cognition.2014.06.011.
Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychon Bull Rev, 24*, 1234–1251. https://doi.org/10.3758/s13423-016-1199-y.
Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences, 14*(12), 542–551. https://doi.org/10.1016/j.tics.2010.09.008.
Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S., & Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition, 116*(1), 33–41. https://doi.org/10.1016/j.cognition.2010.03.012.
Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*(3), 547–555.
Praet, M., Titeca, D., Ceulemans, A., & Desoete, A. (2013). Language in the prediction of arithmetics in kindergarten and grade 1. *Learning and Individual Differences, 27*, 90–96. https://doi.org/10.1016/j.lindif.2013.07.003.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59.
Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88*, 552–572.
Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 870–888.
Ratcliff, R., & Childers, R. (2015). Individual Differences and Fitting Methods for the Two-Choice Diffusion Model of Decision Making. *Decision, 2*, 237–279.
Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Comput. 24*, 1186–1229.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420.
Ratcliff, R., & McKoon, G. (2018). Modeling Numerosity Representation with an Integrated Diffusion Model. *Psychological Review, 125*(2), 183–217. https://doi.org/10.1037/rev0000085.
Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science, 9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067.
Ratcliff, R., Sederberg, P. B., Smith, T. A., & Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia, 93*, 128–141. https://doi.org/10.1016/j.neuropsychologia.2016.09.026.
Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction times. *Psychological Review, 111*(2), 333–367.
Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*, 70–94.
Ratcliff, R., Smith, P. L., & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychological Science, 24*, 458–470.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438–481. https://doi.org/10.3758/bf03196302.
Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261–300. https://doi.org/10.1037/0033-295X.106.2.261.
Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cognitive Psychology, 103*, 1–22.
Rouder, J. N., & Geary, D. C. (2014). Children's cognitive representation of the mathematical number line. *Developmental Science, 17*(4), 525–536.
Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology, 30*(2), 344–357. https://doi.org/10.1111/j.2044-835X.2011.02048.x.
Sasanguie, D., Gobel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number-space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology, 114*, 418–431.
Servant, M., Montagnini, A., & Burle, B. (2014). Conflict tasks and the diffusion framework: Insight in model constraints based on psychological laws. *Cognitive Psychology, 72*, 162–195. https://doi.org/10.1016/j.cogpsych.2014.03.002.
Servant, M., White, C., Montagnini, A., & Burle, B. (2015). Using covert response activation to test latent assumptions of formal decision-making models in humans. *Journal of Neuroscience, 35*(28), 10371–10385. https://doi.org/10.1523/jneurosci.0078-15.2015.
Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 506–526. https://doi.org/10.1037/0096-1523.26.2.506.
Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the Stroop task: Evidence from ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1398–1412. https://doi.org/10.1037/a0016467.
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.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662. https://doi.org/10.1037/h0054651.
Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods Instruments & Computers, 36*(4), 702–716. https://doi.org/10.3758/bf03206552.
Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review, 122*(2), 312–336. https://doi.org/10.1037/a0038894.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.
Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience, 16*(9), 1493–1504. https://doi.org/10.1162/0898929042568497.
White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology, 63*(4), 210–238. https://doi.org/10.1016/j.cogpsych.2011.08.001.