Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

2020, Vol. 46, No. 11, 2128–2152
http://dx.doi.org/10.1037/xlm0000937

Examining Aging and Numerosity Using an Integrated Diffusion Model

Roger Ratcliff and Gail McKoon
The Ohio State University

Two experiments are presented that use tasks common in research in numerical cognition with young adults and older adults as subjects. In these tasks, one or two arrays of dots are displayed, and subjects decide whether there are more or fewer dots of one kind than another. Results show that older adults, relative to young adults, tend to rely more on the perceptual feature, area, in making numerosity judgments when area is correlated with numerosity. Also, convex hull unexpectedly shows different effects depending on the task (being either correlated with numerosity or anticorrelated). Accuracy and response time (RT) data are interpreted with the integration of the diffusion decision model with models for the representation of numerosity. One model assumes that the representation of the difference depends on the difference between the numerosities and that standard deviations (*SD*s) increase linearly with numerosity, and the other model assumes a log representation with constant *SD*s. The representational models have coefficients that are applied to differences between two numerosities to produce drift rates and *SD*s in drift rates in the decision process. The two tasks produce qualitatively different patterns of RTs: One model fits results from one task, but the results are mixed for the other task. The effects of age on model parameters show a modest decrease in evidence driving the decision process, an increase in the duration of processes outside the decision process (nondecision time), and an increase in the amount of evidence needed to make a decision (boundary separation).

Keywords: integrated diffusion model, approximate number system, response time and accuracy, aging and numeracy, perceptual variables and aging

Many researchers have found that numeracy abilities decline with age, especially in tasks that assess high-level reasoning-type decision-making (e.g., financial and health literacy, risk assessment, drug choice, Boyle et al., 2013; Delazer, Kemmler, & Benke, 2013; Finucane & Gullion, 2010; Li et al., 2015; Szrek & Bundorf, 2013; arithmetic, Charron, Fischer, & Meljac, 2008; Wood & Hanock, 2012; a charity-giving survey, Bruine de Bruin, McNair, Taylor, Summers, & Strough, 2015; subjective numeracy, Fraenkel, Cunningham, & Peters, 2015). Yet others have found that these skills are maintained, especially lower-level skills (e.g., estimation, Gandini, Lemaire, & Dufau, 2008; Gandini, Lemaire, & Michel, 2009; Lemaire & Lecacheur, 2007; number discrimination with one-digit numbers, Trick, Enns, & Brodeur, 1996; subitizing, Watson, Maylor, & Bruce, 2005; Watson, Maylor, & Manson, 2002). Across these studies, a variety of skills in a variety of tasks were examined, with high-level tasks showing more consistent declines. If numeracy abilities are not a single skill, we might expect that different tasks and different aspects of numerosity might be preserved or might decline with age. Of clinical interest, numeracy impairments are found across all ages for patients with mild cognitive impairment (Delazer et al., 2013; Griffith et al., 2003; Kaphingst, Goodman, MacMillan, Carpenter, & Griffey, 2014; Pertl et al., 2014; Triebel et al., 2009), which is often a precursor of Alzheimer's disease, and also patients with early Alzheimer's disease (Maylor, Sheehan, Watson, & Henderson, 2008; Maylor, Watson, & Muller, 2005).

Some studies have examined age differences in adults using low-level numerosity tasks of the kind we used in this study (e.g., Cappelletti, Didino, Stoianov, & Zorzi, 2014; Halberda, Ly, Wilmer, Naiman, & Germine, 2012). Halberda et al. used a numerosity discrimination task in which an array of blue and yellow dots was displayed, and subjects were asked to decide whether there were more blue or more yellow dots. The dots were selected randomly from several sizes, and they were displayed for 200 ms. The data were collected from the Internet for over 10,000 individuals. Results showed that accuracy (the Weber fraction) decreased with age from about age 30 on up, and response time (RT) increased with age from about age 18 on up. Cappelletti et al. performed a similar study with the same blue-dot and yellow-dot displays. In one condition, dots were selected randomly from several sizes such that area was proportional to numerosity, and in another condition, the total area of blue dots was equated to the total area of yellow dots. This was done by making the area of each of the dots for the more-numerous color smaller on average than those for the less-numerous color. Their results showed no aging effect for accuracy on the proportional-area condition but did for the equal-area condition. Also, RTs differed between the two conditions for older adults but only to a small degree for young adults.

This article was published Online First July 30, 2020.

Roger Ratcliff and Gail McKoon, Department of Psychology, The Ohio State University.

Preparation of this article was supported by the National Institute on Aging Grants R01-AG041176 and R01-AG057841.

Correspondence concerning this article should be addressed to Roger Ratcliff, Department of Psychology, The Ohio State University, 291 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210. E-mail: ratcliff.22@osu.edu

Simple numerosity discrimination tasks have also been important in the developmental literature because performance has been predictive of math skills for children (e.g., Halberda et al., 2012; Halberda, Mazzocco, & Feigenson, 2008). Furthermore, training on these tasks has been found to improve math performance for children (Park & Brannon, 2013, 2014). Results such as these suggest it is worth using these tasks to examine whether there are declines in numerosity processing in aging and whether they can be usefully explored with quantitative models. This is accomplished in this article by showing that our models apply to older adults, and this allows us to examine aging effects by comparing model parameters for different age groups. In our experiments, we used two groups of older adults, 60–69- and 70–90-year-olds, and when we use the term "older adults," we mean this to apply to both groups.

In the experiments for this article, we examined the performance of college-age adults, 60–69-year-old adults, and 70–90-year-old adults. In Experiment 1, subjects were to decide whether there were more blue dots than yellow dots in a single array (B/Y task, Figure 1A), and in Experiment 2, subjects were to decide whether there were more yellow dots in the left or right of two horizontally spatially separated arrays (L/R task, Figure 1B). The standard findings (replicated in our experiments) are that it is easier to discriminate 10 from 20 objects than 15 from 20 (accuracy decreases as the difference in two numerosities decreases; the "distance" effect) and that it is easier to discriminate 10 objects from 20 than 60 objects from 70 (accuracy decreases as numerosities increase; the "size" effect). In research using tasks like ours, "easier" has almost always meant more accurate. For college-age adults (Ratcliff & McKoon, 2018), the patterns of accuracy and RTs were qualitatively different for the two tasks, and the experiments here examined whether older adults show the same patterns.

Two popular, and competing, models have been proposed to explain the size and distance effects (Dehaene & Changeux, 1993; Gallistel & Gelman, 1992). Both have been developed in the context of a postulated "approximate number system" (ANS) that provides representations of numeracy to cognitive processes, with the representation of each number normally distributed on a numerosity scale. The differences between the models lie in their assumptions about scale and variability. In one (the "linear" model), numerosity is represented on a linear scale, and the variability around numerosities (i.e., their standard deviations [*SD*s]) increases linearly as numerosity increases. In the other (the "log" model), numerosity is represented on a logarithmic scale with equal variability around all numerosities (Figure 1C). Both models conform to Weber's law, that is, the difference in two numerosities divided by the total number is constant (Dehaene & Changeux, 1993; Gallistel & Gelman, 1992; see Zorzi, Stoianov, & Umilta, 2005, for a review).

It has been claimed that the two models cannot be discriminated on the basis of empirical data (e.g., Dehaene, 2003), but that argument has been based solely on accuracy, not RTs. In fact, the models can be discriminated when RTs are taken into account. Ratcliff and McKoon (2018) integrated the ANS representation models with Ratcliff's diffusion decision model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Figure 1D and Appendix A). In the experiments presented here, stimuli varied on two dimensions, numerosity and area. In the B/Y task, we varied the overall areas of the blue and yellow dots, the differences in numerosity between the blue and yellow dots, and the overall numerosity of both kinds of dots. For the B/Y task, Ratcliff and McKoon (2018) found that as overall numerosity increased, for a fixed difference of 5 between the numerosities of the blue and yellow dots, not surprisingly, accuracy decreased—but surprisingly, RTs also decreased. It was not the case that RTs increased with difficulty (i.e., overall numerosity). As we show soon, the linear model can account for this pattern of data, but the log model cannot. For the L/R task, stimuli varied in numerosity and area in the same way as for the B/Y task. Ratcliff and McKoon found that as overall numerosity increased, with a fixed difference between the numerosities of the two arrays, accuracy decreased and RTs increased, the usual relation between accuracy and RTs. The log model could account for this pattern of data, and the linear model could not.

In both the B/Y and L/R experiments, there were six possible dot sizes, and each stimulus had a mixture of sizes. For both experiments, there were two conditions for the area variable. In one, the size of each dot was selected randomly from the six possible sizes, and this means that the total area of the dots of one color (B/Y task) or one side (L/R task) was on average proportional to the number of dots. For the second condition, dot sizes differed in the stimulus display, but the total area of the dots was equal for the dots of the two colors (B/Y) or both sides (L/R).

The experiments and model-based analyses were designed to address five issues of importance for older adults' numerical cognition: whether the older adults produced the same patterns of results as the young adults, whether the older adults showed a deficit in the numerosity information available to them to make decisions relative to young adults, whether the older adults relied on area (a perceptual variable) more than the young adults in making numerosity decisions, whether young and older adults were affected by the perceptual variable convex hull (this is discussed in detail in a later section), and whether there were consistent individual differences across the tasks. In the discussion, we suggest the following way of understanding the conclusions: For the B/Y task, the linear integrated model applies, and we argue that this is because differences in the number of dots are all that can be used to make decisions. For the L/R task, in Ratcliff and McKoon (2018), the log integrated model applied, and it was argued that this is because approximate magnitudes for the two arrays can be separately computed. However, the results presented here are ambiguous compared with the results from Ratcliff and McKoon, and there is no clear-cut winner, but there is a tendency for very old adults to use differences rather than separate representations. In the next sections, we describe the ANS models and the diffusion model and how they are integrated.

The ANS-Diffusion Models

In the ANS-diffusion models, drift rate (evidence driving the accumulation process) and the *SD* in drift rate across trials are provided by the ANS model (Figure 1C). Boundary settings, nondecision times, and the ranges in starting point and nondecision time come from the diffusion model (see Appendix A).

Figure 1C shows how drift rates for the two models are computed. For the linear model, drift rate (v) is a coefficient (v_I) multiplied by the numerical difference between the blue and yellow dots for the B/Y task and the numerical difference between the

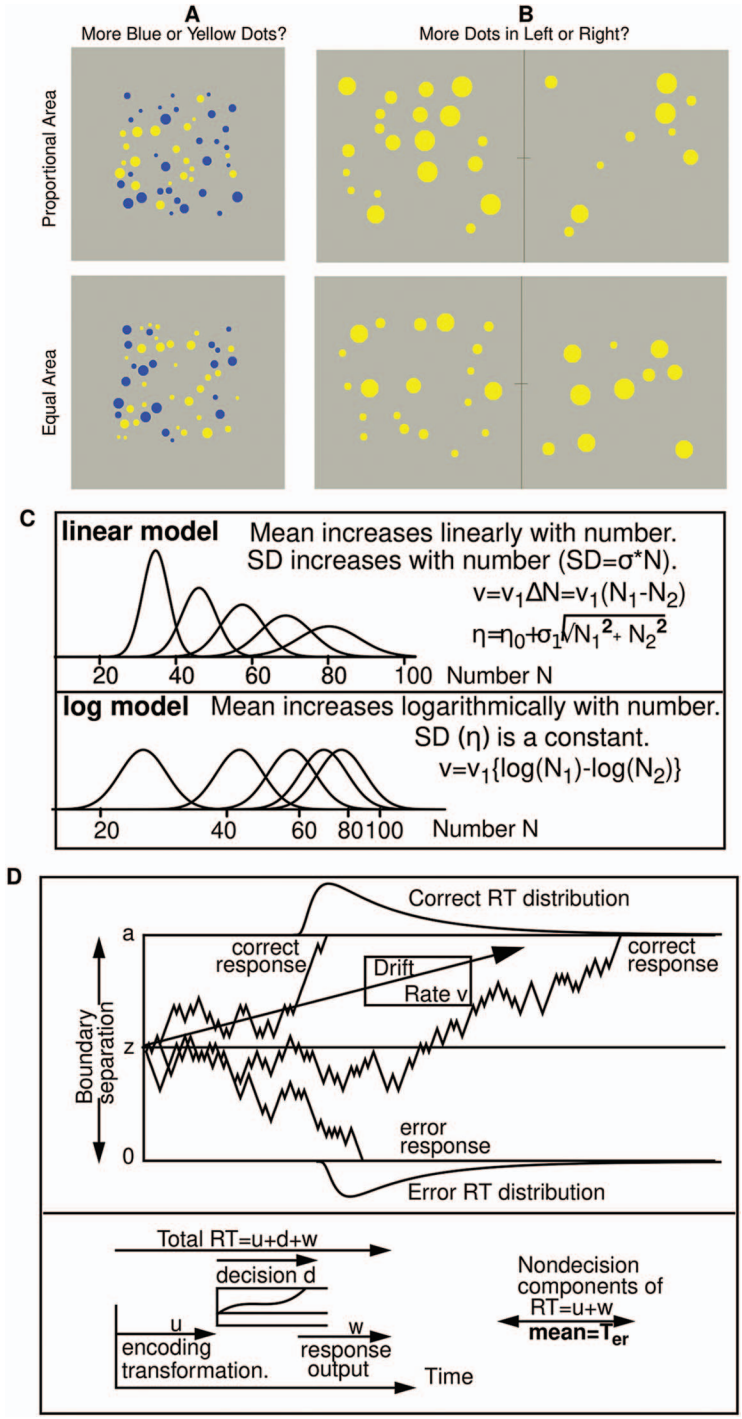

Figure 1. Panel A: Examples of stimuli for Experiment 1, the B/Y task. Panel B: Examples of stimuli for Experiment 2, the L/R task. Panel C: Models of numerosity representation and the equations that translate numerosity to drift rate and across-trial variability in drift rate. Panel D: Illustration of the diffusion decision model. The top panel of 1D shows the decision process with three simulated paths and with model parameters. The bottom one shows the additional components of the decision model that produce the total RT. *SD* = standard deviation; RT = response time. v is drift rate, v_1 is the drift rate coefficient, N_1 and N_2 are the numbers of dots in the two arrays, η is across trial variability in drift rate, σ_1 is the across trial variability coefficient, η_0 is the constant across trial variability in drift rate, and T_{er} is nondecision time. See the online article for the color version of this figure.

number of dots on the left and those on the right for the L/R task. For the log model, drift rate is the difference in the logs of two numerosities multiplied by a coefficient (v_1). For both models, the coefficient of drift rate is most related to accuracy and so it is most likely to be most related to the numerosity abilities of individuals; a larger coefficient gives higher accuracy. Figure 1C also shows how across-trial *SD* in drift rate (η) is computed. For the linear model, it is a constant (η_0) plus a coefficient (σ_1) multiplied by the square root of the sum of squares of the two numerosities (the square root of the sum of squares is how *SD*s are combined; variances are added).

For the log model, we might assume that η remains constant as numerosity increases, as in the log model for accuracy (e.g., Dehaene, 2003). However, there is no guarantee that an ANS model combined with the diffusion model would fit data the best with a constant value of η. Therefore, we gave our log ANS diffusion model the same flexibility in across-trial variability as the linear model, with the same expression for η as for the linear model. Then, when the model is fit to data, the result could be a constant value of η as numerosity increases (i.e., with σ_1 near zero) or a value of η that increases with numerosity.

This has the advantage of giving the linear and log models the same number of parameters, and this makes model selection less ambiguous because different measures such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) give the same results. The expressions for AIC and BIC are a likelihood (we use a multinomial likelihood measure for model fitting that is based on fits to accuracy values and RT quantiles; see later) minus a term based on the number of parameters. If the number of parameters is the same, then model selection through AIC, BIC, and likelihood is the same. Thus, the only difference between our linear and log models was that the drift rate assumption was different: linear versus log. We examine this issue in more detail in Appendix C including an analysis in which we fit the log model with across-trial *SD* in drift rate set to a constant that reduces the number of parameters by one and allows us to use AIC and BIC in model comparison.

With the standard diffusion model with no representation model to provide drift rates, there would be 20 different drift rates and 20 different across-trial *SD*s in drift rates for the conditions of both experiments (10 equal-area conditions and 10 proportional-area conditions). However, in each integrated model, drift rates (and their *SD*s) are set by the representation model. For example, there would be one coefficient for the 10 equal-area conditions and one for the 10 proportional-area conditions, plus two parameters for across-trial variability in drift rate (a coefficient to multiply the square root of the sum of the squared numerosity values plus a constant value; Figure 1C). If the equal-area conditions are more difficult than the proportional-area ones, then the drift-rate coefficient would be smaller for the equal-area conditions.

Overall, for the integrated models for both experiments, there are six free parameters plus one drift-rate coefficient for each of the area conditions (equal and proportional). From the diffusion model, there are the distance between the boundaries, nondecision time, and the ranges in the starting point and nondecision time. For the experiments presented here, the starting point can be set to halfway between the two boundaries because "blue" responses to blue stimuli were symmetric in accuracy and RTs to "yellow" responses to yellow stimuli in the B/Y task and "left" responses to left stimuli were symmetric with "right" responses to right stimuli in the L/R task. We used this symmetry to combine data from the pairs of conditions to provide half the number of conditions and double the number of observations per condition.

Why Does the Linear Model Produce Shorter RTs as Accuracy Decreases?

To illustrate how the linear ANS model produces shorter RTs as accuracy decreases when there is a constant numerosity difference, we used the simple case for which the boundaries of the diffusion process are equidistant from the starting point. We used two values of across-trial *SD* in drift rate, one large and one small, to show how the RTs shorten and accuracy decreases as across-trial *SD* in drift rate increases. In both examples in Figure 2, the distributions of across-trial variability were normal and centered at 0.1. The red solid line represents the larger across-trial *SD* in numerosity, and the blue dashed line represents the smaller across-trial *SD* in numerosity. Two values of drift rate were selected from each distribution at about plus or minus one *SD*, −0.05 and 0.25 for the larger numerosity and 0.05 and 0.15 for the smaller.

Figures 2A and 2B show the RT distributions for correct and error responses, with the two values of *v* for the smaller numerosity (Figure 2A) and the two values for the larger numerosity (Figure 2B). For the smaller numerosity, the 0.15 and 0.05 drift rates produce accuracy values of 0.86 and 0.65, respectively, which average to 0.76, and they produce RTs of 685 ms and 748 ms for correct responses, which, when weighted by their probabilities (0.86 and 0.65), average to 717 ms. For the larger numerosity, the 0.25 and −0.05 drift rates produce accuracy values of 0.95 and 0.35, which average to 0.65. They produce RTs of 616 ms and 748 ms for correct responses, which, when weighted by their probabilities (0.95 and 0.35), average to 652 ms. Thus, accuracy is lower for the larger numerosity, 0.65, than the smaller, 0.76, and this produces the counterintuitive result: RTs are shorter for the larger numerosity, 652 ms versus 717 ms. The computations for RTs for errors are shown at the bottom boundary in the figures.

To explain this generally, when the distribution of drift rates has a large *SD*, then drift rates in the left tail are negative (Figure 2B). Responses are slower than responses in the right tail at the same distance from the mean, but they have lower probabilities of being correct (because their drift rate is toward the error boundary). This means that fast correct responses in the right tail are weighted more heavily (there are more of them) than slower responses in the left tail, which leads to overall faster responses. As numerosity increases, the *SD* increases, which leads to lower accuracy and faster responses.

Experiments

In the two experiments, we collected data from three groups of subjects, which allowed us to compare the log and linear ANS-diffusion models for the B/Y task (Experiment 1) and the L/R task (Experiment 2) as a function of age. The same subjects were tested in both experiments, which allowed individual differences in model parameters (correlations) to be compared across the two tasks.

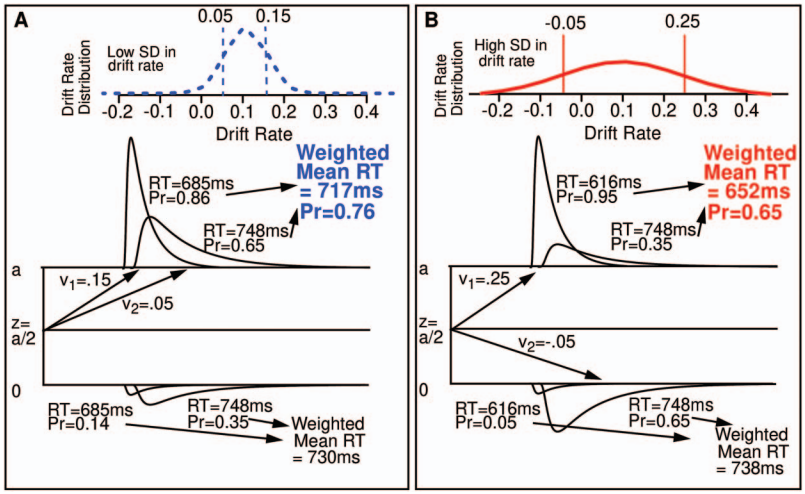

Figure 2. An illustration of how the predictions of the linear model arise. The distributions of drift rate (across trials) for high numerosity (wide solid distribution) and low numerosity (narrow dashed distribution) are shown at the tops of the panels. To illustrate averaging over these distributions, two drift rates are chosen (v_1 and v_2), and accuracy values and mean response times (RTs) are shown. Accuracy for the mixture is the average of the two accuracy values, and mean RT is a weighted sum of the two mean RTs. Panel A shows the averages for the low-standard-deviation (*SD*) condition, and Panel B shows the averages for the high-*SD* condition with the averages for correct responses. For completeness, error responses are also shown; note that for boundaries equidistant from the starting point, for a single drift rate, correct and error RTs are the same. Pr is probability, a is boundary separation and z is the starting point of the process. See the online article for the color version of this figure.

Subjects

Sixty older adults, ages 60 to 90, were recruited from senior citizen centers in the Columbus area and paid $18 per session. Thirty had ages between 60 and 69, and 30 had ages between 70 and 90. We collected data from the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975) and the Vocabulary and Matrix Reasoning subtests of the Wechsler Adult Intelligence Scale–Third Edition (Wechsler, 1997). The subjects met the following criteria: a score of 26 or above on the Mini-Mental State Examination, no evidence of disturbances in consciousness or medical or neurological diseases that could impair cognition, no head injuries with loss of consciousness, and no current psychiatric disorder. Their static visual acuity was screened to ensure a minimum corrected visual acuity of 20/30 using a Snellen "E" chart. The 30 young adults were recruited from The Ohio State University student body and were paid $12 per session for participation. All participants provided informed consent under a protocol approved by The Ohio State University's institutional review board.

Each subject took part in two sessions, with approximately three fourths of each session completing an experiment and one fourth collecting demographic information and IQ measures. The order of the experiments was randomized across subjects. The subjects were being tested in several experiments, and for the young adults, 60–69-year-olds, and 70–90-year-olds, the B/Y task was the first tested on eleven, eight, and eight subjects, respectively. The L/R task was the first tested on four, six, and six subjects (for the three subject groups), and some other task preceded these two tasks for 15, 16, and 16 subjects, respectively. Demographic and IQ scores are shown in Table 1.

Table 1
Subject Characteristics

Measure	Younger adults		60–69-year-olds		70–90-year-olds	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age	20.5	2.0	63.5	2.7	75.3	15.7
Years education	13.7	1.6	15.4	2.2	15.7	2.5
MMSE	29.3	0.9	29.1	0.9	28.9	1.1
WAIS-III Vocabulary (scaled score)	14.2	2.4	14.0	2.2	13.8	2.2
WAIS-III Matrix Reasoning (scaled score)	13.7	2.5	12.8	2.3	13.1	3.1

Note. MMSE = Mini-Mental State Examination; WAIS-III = Wechsler Adult Intelligence Scale—Third Edition.

Apparatus and Stimuli

The stimuli were presented on ASUS Chromebook computers. These have an 11.6-in. diagonal screen with a width of 25.7 cm and height of 14.5 cm. The resolution was 1366 × 768 pixels. The dots had radii of 6, 8, 10, 12, 14, and 16 pixels, and one pixel was 0.0188 cm per side, which corresponds to 0.0203° at 53-cm viewing distance. Within an array, the minimum distance between dot edges was five pixels, and the maximum was 360 pixels.

For the B/Y stimuli (Experiment 1; Figure 1B), the array of dots was displayed on a gray pedestal of 640 × 640 pixels. For the L/R stimuli (Experiment 2; Figure 1C), the two arrays were displayed side by side with a thin line between them. The minimum distance between dots in the two patches was 80 pixels, and the stimuli were presented on a pedestal of 1262 × 640 pixels.

Method

There were 20 blocks of 100 trials for both experiments, giving a possible total of 2,000 observations per subject. Subjects initiated each block of trials by pressing the space bar on the keyboard. The first block of trials and the first response in each block were discarded, which gave a total of 1,881 possible trials per session. Subjects were tested for about 40 min on each of these tasks, and they usually did not finish all the possible trials available (because demographics and IQ were collected in the remaining time of the hour testing time). This resulted in mean numbers of responses for the B/Y and L/R tasks, respectively: for young adults, 1,480 and 1,575; for 60–69-year-old adults, 1,412 and 1,597; and for 70–90-year-olds, 1,352 and 1,427. Ratcliff and Childers (2015; Figure 2) conducted an analysis of diffusion model parameters as a function of the number of trials in two experiments (numerosity discrimination and lexical decision) and found little difference in model parameters even with differences in the numbers of observations as large as 2:1.

Stimuli in Experiments 1 (B/Y) and 2 (L/R) were presented for 300 ms and 250 ms, respectively, and then the screen returned to the background color. The short display time was intended to reduce the possibility of subjects using slow, strategic search processes to make their decisions. Subjects were instructed to respond as quickly and accurately as possible, and responses were collected by key presses on the Chromebook keyboard using the "/" and "z" keys, one for each choice. For both tasks for practice and training, there were four trials in which a stimulus was presented until the response key was pressed. For these trials, the research assistant administering the task instructed the subject about the nature of the stimuli. For each of the four trials, a message indicated what kind of stimulus it was (e.g., "an example of a large number of yellow dots"). As noted above, the first block of trials was considered practice and eliminated.

In both experiments, if an RT was longer than 1,500 ms, the message "too slow" was presented for 500 ms. If an RT was shorter than 280 ms, a "too fast" message was presented for 1,500 ms. Accuracy feedback in the form of the word "correct" or "error" was presented on each trial for 250 ms, followed by a blank screen for 250 ms.

The stimuli in Experiment 1 were blue and yellow dots intermingled in a single array (B/Y task; Figure 1B), and subjects decided whether there were more blue or more yellow dots. The total numbers of dots and the differences between them varied across conditions. For a difference of five, the combinations were 15/10, 20/15, 25/20, 30/25, and 40/35. For a difference of 10, they were 20/10, 30/20, and 40/30, and for a difference of 20, they were 30/10 and 40/20. These numerosities were chosen to produce a range of accuracy values from low to high.

The areas of both the blue and yellow dots were either randomly selected from the six radii, in which case area was proportional to the number of dots, or the total area of the two colors was controlled to be equal to the total area that would be obtained from 25 dots randomly selected. This means that the area of each of the dots in the smaller number was larger on average than the area of each of the dots in the larger number.

The stimuli in Experiment 2 were side-by-side arrays of dots, all of them yellow, and the task was to decide whether there were more dots on the left or the right (L/R task; Figure 1C). There were the same combinations of numbers of dots and area manipulations as for Experiment 1. In both experiments, the 20 conditions represented by the 10 numerosities and two area conditions were presented five times in a block of 100 trials in random order.

Results

In this section, we present the results of the experiments in three ways. First, the data are shown as plots of RTs for correct responses against the corresponding proportions of correct responses; these display the main features of the RTs and accuracy data. Table 2 shows mean accuracy and RT across subjects and numerosity conditions for both experiments. Accuracy shows a decrease

Table 2
Accuracy, Drift Rate, and Mean RT as a Function of the Area Variable

Task and age	Proport.-area prob.	Equal-area prob.	v_p	v_e	Proport.-area mean RT	Equal-area mean RT
B/Y young	0.813	0.699	0.0319	0.0169	604	635
B/Y 60–69	0.792	0.633	0.0287	0.0099	728	796
B/Y 70–90	0.725	0.579	0.0233	0.0070	710	752
L/R young	0.903	0.859	1.1220	0.9072	454	465
L/R 60–69	0.892	0.817	1.1805	0.8033	537	566
L/R 70–90	0.868	0.776	0.9446	0.5965	564	595

Note. Proport. = proportional; prob. = probability; RT = response time; B/Y = Experiment 1; L/R = Experiment 2. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition. The drift-rate coefficients are for the linear model for the B/Y task and the log model for the L/R task.

with age, a decrease from the proportional-area condition to the equal-area condition, and an increase in the size of the area difference with age. There are parallel effects for mean RT with an increase with age, an increase from the proportional-area condition to the equal-area condition, and with the difference in mean RT for the two area conditions increasing with age. The second way results are presented is in the fits of the models to the data, which are displayed in quantile-probability plots (for both correct and error responses) with predictions and data averaged across subjects in exactly the same way. The third way results are presented is in plots of predictions against data for each subject and each condition for accuracy and the 0.1, 0.5, and 0.9 quantile RTs. These show whether there is any serious deviation between theory and data for individual subjects or individual conditions. The mean values of the model parameters and the mean G^2 statistic over subjects are reported in Table 3. Statistical analyses are given for the data (accuracy and mean RTs) and the model parameters. These focus on aging effects and the effects of the area variable. Finally, we discuss what the models tell us about decision-making in numerosity tasks.

For all the analyses, we removed trials with RTs shorter than 300 ms and longer than 4,000 ms for the B/Y and L/R tasks for the older adults, 300 ms and 2,000 ms for the B/Y task for young adults, and 250 ms and 2,000 ms for the L/R task for young adults. The lower cutoff was set at 250 ms instead of 300 ms because accuracy was above chance for many of the young adults at 300 ms for the L/R task. For the B/Y and L/R tasks, for young adults, the proportions of responses eliminated were 0.019 and 0.016, respectively; for 60–69-year-old adults, the proportions were 0.005 and 0.012, respectively; and for 70–90-year-old adults, the proportions were 0.017 and 0.016, respectively.

Accuracy and RT Results

Figure 3 shows latency-probability plots of the data for correct responses (Audley & Pike, 1965; Ratcliff, Van Zandt, & McKoon, 1999; Vickers, Caudrey, & Willson, 1971). Mean RTs for correct responses are plotted against accuracy, that is, response proportions, for equal-area conditions (the Xs) and proportionate-area conditions (the Os). As difficulty decreases, the proportion of correct responses increases. The lines connect data points that have the same difference in numerosities: A line of five points represents the 10/15, 15/20, 20/25, 25/30, and 35/40 conditions; a line of three points represents the 10/20, 20/30, and 30/40 conditions; and a line of two points represents the 10/30 and 20/40 conditions.

For the B/Y task, the three groups of subjects show the same pattern of results (Figure 3A−3C). As expected, accuracy decreased as difficulty increased. Specifically, accuracy decreased both as the total of the numerosities of the dots of the two colors increased (e.g., from 15/10 to 40/35) and as the difference between the numerosities of the dots of the two colors decreased (e.g., from the conditions with a difference of 20, to 10, to five), the standard results with these manipulations. Also as expected, equal-area discriminations were more difficult than proportional-area discriminations.

What was not expected (but replicates the results in Ratcliff & McKoon, 2018) is that RTs decreased as difficulty increased for a constant numerosity difference and increasing total numerosity. The usual and intuitive effect is that RTs increase as difficulty increases. Figures 3A, 3B, and 3C illustrate the counterintuitive result. For numerosity differences of five and 10 (the lines with five and three points), as accuracy decreased, RTs also decreased for 10 of the 12 plots in the figures (for proportional areas for the young adults, the plot was slightly increasing). For differences of 20, results were mixed, with four increasing and two decreasing functions. The decrease in RTs with decreasing accuracy is obtained for both area conditions and for young and older adults, although the effect is smaller for the young adults.

The highly unexpected pattern is not obtained in almost all studies with single stimuli changing on one dimension (although it has been obtained in perceptual and value-based tasks in which there are two stimuli and magnitudes and differences between the two stimuli are manipulated; Hunt et al., 2012; Niwa & Ditterich, 2008; Ratcliff & McKoon, 2018; Ratcliff, Voskuilen, & Teodorescu, 2018). Here, the importance of RT data is that they give an

Table 3
Mean Values of Integrated Diffusion Model Parameters for the Three Age Groups, Two Tasks, and Two Models

Task and age	Model	a	T_{er}	$100\sigma_1$	s_z	s_t	v_p	v_e	η_0	G^2	G^2 $\sigma_1 = 0$
B/Y young	Linear	0.100	0.466	0.540	0.046	0.260	0.0319	0.0169	0.023	261.4	
B/Y 60–69	Linear	0.121	0.537	0.623	0.041	0.294	0.0287	0.0099	0.036	258.3	
B/Y 70–90	Linear	0.117	0.518	0.715	0.052	0.277	0.0233	0.0070	0.022	263.6	
B/Y young	Log	0.096	0.462	0.175	0.037	0.257	0.5751	0.3054	0.091	276.2	277.6
B/Y 60–69	Log	0.117	0.529	0.326	0.030	0.291	0.5023	0.1752	0.091	275.3	281.0
B/Y 70–90	Log	0.116	0.512	0.564	0.045	0.275	0.3859	0.1088	0.072	277.2	281.4
L/R young	Linear	0.091	0.367	0.781	0.054	0.173	0.0692	0.0548	0.022	258.1	
L/R 60–69	Linear	0.115	0.419	0.880	0.056	0.184	0.0722	0.0489	0.066	251.8	
L/R 70–90	Linear	0.112	0.422	0.784	0.049	0.169	0.0570	0.0354	0.045	243.2	
L/R young	Log	0.083	0.363	0.069	0.038	0.172	1.1220	0.9072	0.121	256.3	256.4
L/R 60–69	Log	0.107	0.412	0.102	0.043	0.179	1.2265	0.8311	0.212	240.4	245.6
L/R 70–90	Log	0.107	0.416	0.174	0.036	0.164	0.9446	0.5965	0.158	257.0	261.2

Note. The parameters were boundary separation a, starting point $z = a/2$, and mean nondecision component of response time T_{er}. The constant coefficient of standard deviation in drift across trials is η_0, and the coefficient that multiplies the square root of the sum of the squared numerosities is σ_1. Range of the distribution of starting point is s_z, and range of the distribution of nondecision times is s_t. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition. G^2 is the multinomial likelihood statistic. The B/Y task is Experiment 1, and the L/R task is Experiment 2.

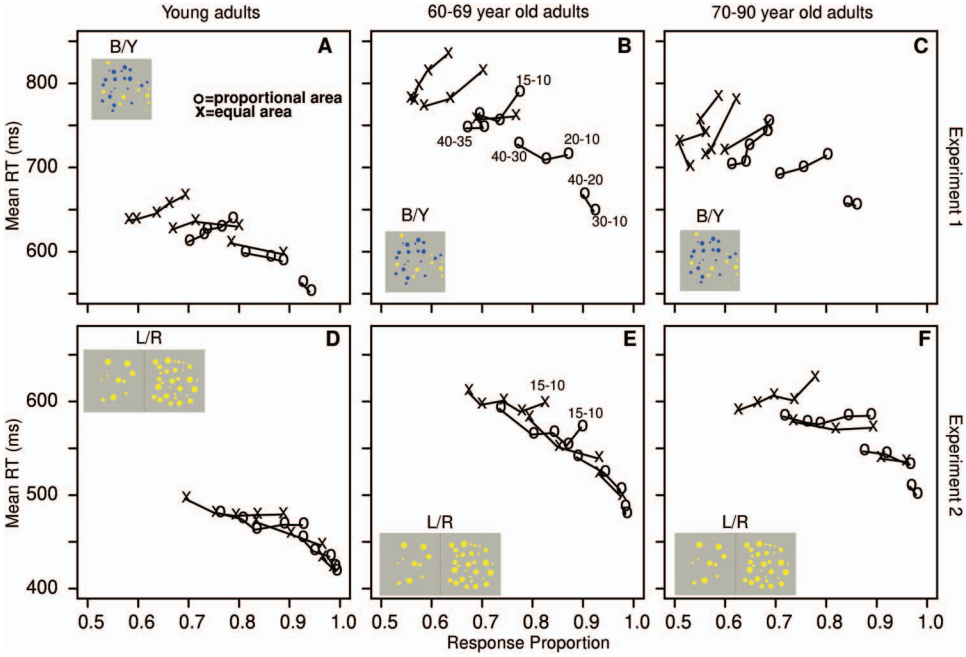

Figure 3. Plots of mean response time (RT) against accuracy for Experiments 1 (B/Y) and 2 (L/R). The Xs are for equal-area conditions, and the Os are for proportional-area conditions. For panels A−D (Experiments 1 and 2), the conditions with differences of five (10/15, 15/20, 20/25, 25/30, 35/40) are represented by the groups of five points joined by lines with the conditions arranged from right to left, with smaller numbers to the right (10/15). The groups of three points are differences of 10 (10/20, 20/30, 30/40), with smaller numbers to the right, and the pairs of dots joined by a line are to the right 10/30 and to the left 20/40. The end points of these pairs are labeled in Panel B for the proportional-area conditions. See the online article for the color version of this figure.

understanding of these numerosity tasks that is quite different from that based on accuracy alone.

Unlike the results for the B/Y task, the L/R task shows a pattern of results that is more like the pattern that might be expected from single-stimulus experiments. Figures 3D, 3E, and 3F show plots for the L/R task that are equivalent to those from the B/Y task. In Ratcliff and McKoon (2018), the functions corresponding to those in Figures 3D−3F had RTs increasing as accuracy decreased for a constant numerosity difference, with increasing overall numerosity. The curves all fell on a single latency-probability function to a good approximation. However, the results from the L/R task here are somewhat mixed. For young adults with differences in numerosity of five, the functions are flat for the higher accuracy values, with the right-hand points lying a little off what looks like a single function. For 60–69-year-old adults, this deviation becomes more pronounced, with two points lying off the single function (these are for the 15/10 conditions). For the 70–90-year-olds, the functions look more like those for young adults for the B/Y task. What seems to be happening for the L/R task is a transition in how the task is performed as a function of age, which will be discussed later.

In the next two paragraphs, we present analyses of variance on accuracy and mean RTs, collapsing over the different numerosity conditions for each experiment, to examine the effects of the age and area variables. For the B/Y task, analysis of variance on accuracy with two factors, the two area conditions and age, showed a significant effect of age, $F(2, 87) = 17.5$, $p = 4.2 \times 10^{-7}$, $\eta_p^2 = .153$, and area, $F(1, 87) = 815.3$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .414$, and a significant interaction between age and area, $F(2, 87) = 7.6$, $p = 1.8 \times 10^{-4}$, $\eta_p^2 = .008$. The L/R task showed similar effects on accuracy with significant effects of age, $F(2, 87) = 9.5$, $p = 9.0 \times 10^{-4}$, $\eta_p^2 = .117$, and area, $F(1, 87) = 282.0$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .252$, and a significant interaction between age and area, $F(2, 87) = 10.8$, $p = 6.3 \times 10^{-4}$, $\eta_p^2 = .019$.

Similar results were obtained for mean RTs. For the B/Y task, an analysis of variance with the two area conditions and age showed a significant effect of age, $F(2, 87) = 8.0$, $p = 6.5 \times 10^{-4}$, $\eta_p^2 = .149$, and area, $F(1, 87) = 160.2$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .022$, and a significant interaction between age and area, $F(2, 87) = 8.7$, $p = 3.6 \times 10^{-4}$, $\eta_p^2 = .002$. The L/R task showed similar effects on mean RT with significant effects of age, $F(2, 87) = 15.0$, $p = 2.6 \times 10^{-6}$, $\eta_p^2 = .249$, and area, $F(1, 87) = 105.6$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .013$, and a significant interaction between age and area, $F(2, 87) = 7.4$, $p = .0011$, $\eta_p^2 = .002$.

Quantile-Probability Plots

Quantile-probability plots are a way of displaying the joint behavior of RT distributions and accuracy. We also use them in Figures 4 and 5 to show fits of the models to RT distributions and accuracy values. Here, we show plots for data averaged

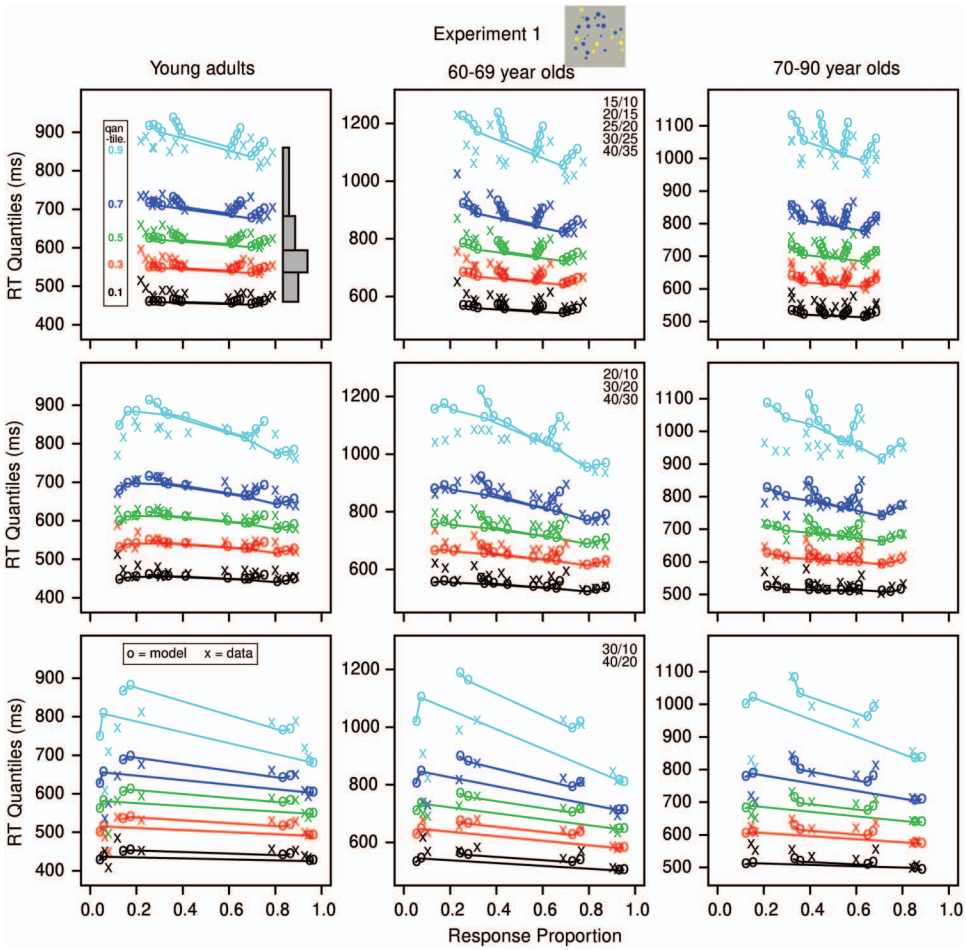

Figure 4. Quantile-probability functions for the linear model for Experiment 1 (B/Y) for the young adults, 60–69-year-olds, and 70–90-year-olds. These plot RT quantiles against response proportions (correct responses to the right of 0.5 and errors to the left). The green/central lines are the median RTs. The number of dots in the conditions in the plots is shown in the top right corner, with the top one in each condition corresponding to the right-hand point in the plot. The more extreme functions (more visible for Experiment 1) are for proportional-area conditions, and the less extreme for equal-area conditions. The quantiles are labeled on the left-hand side of the top left plot and equal-area rectangles drawn between the quantiles are shown on the right side of the plot. RT = response time. See the online article for the color version of this figure.

over subjects and predicted values averaged over subjects in the same way. The 0.1, 0.3, 0.5, 0.7, and 0.9 RT quantiles are computed from the data for each condition and plotted vertically on the *y*-axis above the value of the choice proportion for that condition plotted on the *x*-axis. There is 0.2 probability mass between each pair of these quantiles, and drawing equal-area rectangles between them produces an approximation to RT distributions (see the top left panel of Figure 4, and for examples and a complete description, see Figure 3; Ratcliff & McKoon, 2018). Correct responses are on the right of 0.5 accuracy, and errors are on the left (because we have grouped correct "blue" responses with correct "yellow" responses and correct "right" responses with correct "left" responses and the same for errors, the error probabilities are symmetric with the correct probabilities, i.e., error probability is 1 minus the correct probability). Note that these figures include error RTs, while Figure 3 only contains correct RTs.

Quantile-probability plots make it easy to see changes in RT distribution locations and spread as a function of response probabilities. If changes in RT are due to the distributions spreading and not shifting, the 0.1 quantile (leading edge) changes a little, but the 0.9 quantile changes a lot. If the distributions shift, the 0.1 quantile as well as the others change. Also, the relative speeds of correct and error responses can be observed by comparing quantiles across the two halves of the plots (through the lines that join them; see Ratcliff & McKoon, 2008). In these ways, quantile-probability plots allow all the important aspects of both the accuracy and RT data to be read from a single plot.

Figures 4 and 5 show the quantile-probability functions for the B/Y and L/R tasks, respectively, and the fits of the models to them. The Xs are the data, and the Os and the lines joining them (within each group of correct responses or error responses) are the predictions of the models. The proportional-

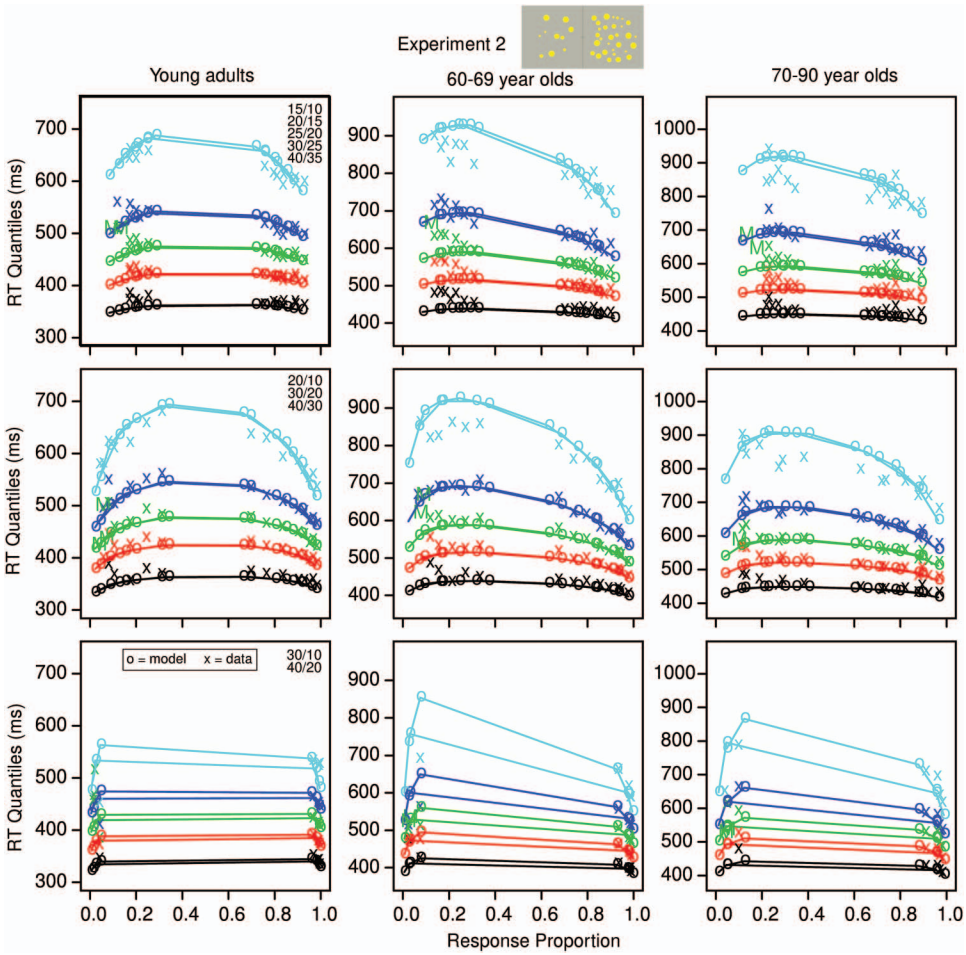

Figure 5. Quantile-probability functions for the log model for Experiment 2 (L/R) for the young adults, 60–69-year-olds, and 70–90-year-olds. RT = response time. The symbols and other details are the same as for Figure 4. See the online article for the color version of this figure.

area conditions are farther to the right for correct responses and further to the left for errors because they have higher accuracy (and so lower error rates) than the equal-area conditions, which are nearer the center. The horizontal lines that connect correct and error responses across 0.5 are not meaningful; they are there only to show which correct responses correspond to which error responses. The median RTs, the middle rows of quantiles vertically, approximately match the means shown in Figure 3 (note that in both Figures 4 and 5, the vertical scales differ among the panels).

The quantile-probability functions for the B/Y task show that the decrease in RT with decreasing accuracy for the constant numerosity differences of five and 10 occurs for all of the five quantiles. This is true for both the equal-area conditions and the proportional-area conditions. In contrast, the functions for the L/R task show the typical inverted-U-shaped functions, with RTs increasing as accuracy decreases over all the quantiles but with the exception for the 70–90-year-olds and the exception described earlier for Figure 3: For the conditions with a difference in numerosity of five, the functions are somewhat U-shaped in all the quantiles.

Fits of the Integrated Models to the Results of the Experiments

Details of the fitting method are given in Appendix B. The first thing to note is that the models are highly constrained by the data. There are eight model parameters that, through the model, must account for 220 degrees of freedom in the summary of the data. This is a massive reduction in degrees of freedom, a degree of reduction that is rarely seen in modeling in psychology. It is also important to note that changing a single parameter in the model changes all aspects of the data, so it is not possible to alter the model, for example, to fit a few deviant data points.

Four of the eight parameters were from the diffusion model: the distance between the boundaries, across-trial range in the starting point, nondecision time, and across-trial range in nondecision time. The other parameters were derived from the ANS models: a drift-rate coefficient (v_e) for the equal-area conditions, a drift-rate coefficient for the proportional-area conditions (v_p), the *SD* coefficient (σ_1), and the constant component of the across-trial *SD* in drift rate (η_0). The 220 degrees of freedom in the data were derived from the number of conditions multiplied by the 11 degrees of

freedom for the proportions of responses between and outside the .1, .3, .5, .7, and .9 bins for correct and error responses minus 1 because the proportions add to 1 and minus the number of parameters. Then, the number of degrees of freedom from applying the models to data was 220 – 8 = 212.

Table 3 shows the parameter values of the models that best fit the data, and Table 4 shows the *SD*s across subjects. The critical value of the chi square for 212 degrees of freedom is 246.0. The mean G^2 values for the two models and three subject groups were only moderately larger than the critical value, which is typical of fits of the diffusion model to data (because of the conservativeness of the chi-square statistic to even small deviations; see Ratcliff, Thapar, Gomez, & McKoon, 2004) and so indicates a good fit of the model to data.

Using the mean G^2 values, for the B/Y task, the linear model fit the data modestly better than the log model (mean differences in G^2 between 15 and 18 for the three subject groups). For the L/R task, the results were not conclusive, with the G^2 values about the same for the two models for the young adults and 60–69-year-olds, but for the 70–90-year-old adults, the linear model fit better than the log model. Detailed discussion of this including analyses based on other fit statistics, fits with the *SD* in drift-rate coefficient set to zero, the number of subjects fit better by each model, and an analysis of qualitative trends in the data is presented in Appendix C.

In Figure 4, we show the fits of the linear model for the B/Y task, and in Figure 5, fits of the log model for the L/R task. The results for the linear model show a good qualitative and quantitative match between theory and data. The model produces the decreases in RT quantiles as accuracy decreases (the counterintuitive finding), the larger and sharper decreases for the equal-area conditions than the proportional-area conditions, and the larger decreases for the higher than the lower quantiles. It also produces the flattening of the functions as the difference in numerosities between the blue and yellow dots increases (from five to 10 to 20). The systematic differences between the model and data are an overprediction for the 0.9 error RT quantiles for older adults in the B/Y task (note that the 0.7 quantiles are not systematically misfit) and the misfit in the RTs for the log model for the L/R task. Apart from these, there were no systematic differences.

The G^2 goodness-of-fit measures are not greatly different for the log and linear models, even though the models produce quite different qualitative predictions. This can be understood by examining the top left panels of Figures 4 and 5. The lines in Figure 4 capture the decreasing quantiles and accuracy, but the lines in Figure 5 do not. If the predictions were switched, then the lines would go through the majority of the data missing the decreasing or increasing functions, but these misses would not be too large. This illustrates why the G^2 values are not greatly different for the two models.

The fit of the models to the data is impressive for several reasons. First, for the two experiments, there is only one drift-rate coefficient for the 10 equal-area conditions and only one for the 10 proportional-area conditions; drift rate is determined by the coefficient and the two numerosities being compared. Second, the values of the four parameters from the diffusion model and the constant component of the across-trial *SD* in drift rate are fixed across all 20 conditions. Third, the eight parameters for the models here contrast sharply with the numbers of parameters that would usually be used to fit the diffusion model to data (20 drift-rate parameters and possibly 20 parameters for across-trial *SD* in drift rates). Integrating the linear and log models with the diffusion model reduces this to two drift-rate coefficients and two *SD* coefficients, the constant *SD* coefficient (η_0) and the one that specifies how the *SD* changes with numerosity (σ_1).

Figures 6 and 7 show plots of model predictions against data for accuracy and the 0.1, 0.5, and 0.9 quantile RTs for each of the experiments for the three subject groups for correct responses. These plots show the data and predictions for the linear model for the B/Y task and the log model for the L/R task. The plots provide a way to seeing if there are any systematic misfits for specific subjects or conditions. If so, there would be a cluster of data points away from the diagonal lines that represent equality. The general

Table 4
Standard Deviations of Integrated Diffusion Model Parameters for the Three Age Groups, Two Tasks, and Two Models

Task and age	Model	a	T_{er}	$100\sigma_1$	s_z	s_t	v_p	v_e	η_0	G^2
B/Y young	Linear	0.015	0.074	0.237	0.029	0.088	0.0102	0.0062	0.043	24.1
B/Y 60–69	Linear	0.029	0.107	0.560	0.034	0.101	0.0093	0.0048	0.039	21.1
B/Y 70–90	Linear	0.024	0.117	0.505	0.038	0.103	0.0162	0.0055	0.044	39.6
B/Y young	Log	0.018	0.073	0.212	0.030	0.089	0.2355	0.1192	0.090	27.8
B/Y 60–69	Log	0.029	0.108	0.559	0.024	0.101	0.1555	0.0919	0.064	23.9
B/Y 70–90	Log	0.025	0.116	0.493	0.038	0.102	0.1570	0.0796	0.102	38.3
L/R young	Linear	0.018	0.057	0.467	0.023	0.078	0.0326	0.0272	0.039	28.0
L/R 60–69	Linear	0.030	0.083	0.564	0.026	0.093	0.0306	0.0239	0.095	23.1
L/R 70–90	Linear	0.022	0.082	0.531	0.036	0.069	0.0319	0.0213	0.070	22.9
L/R young	Log	0.013	0.057	0.122	0.023	0.080	0.4670	0.3873	0.112	26.0
L/R 60–69	Log	0.031	0.081	0.303	0.028	0.090	0.4837	0.3568	0.116	20.3
L/R 70–90	Log	0.023	0.083	0.230	0.033	0.071	0.4932	0.3403	0.126	24.1

Note. The parameters were boundary separation a, starting point $z = a/2$, and mean nondecision component of response time T_{er}. The constant coefficient of standard deviation in drift across trials is η_0, and the coefficient that multiplies the square root of the sum of the squared numerosities is σ_1. Range of the distribution of starting point is s_z, and range of the distribution of nondecision times is s_t. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition. G^2 is the multinomial likelihood statistic. The B/Y task is Experiment 1, and the L/R task is Experiment 2.

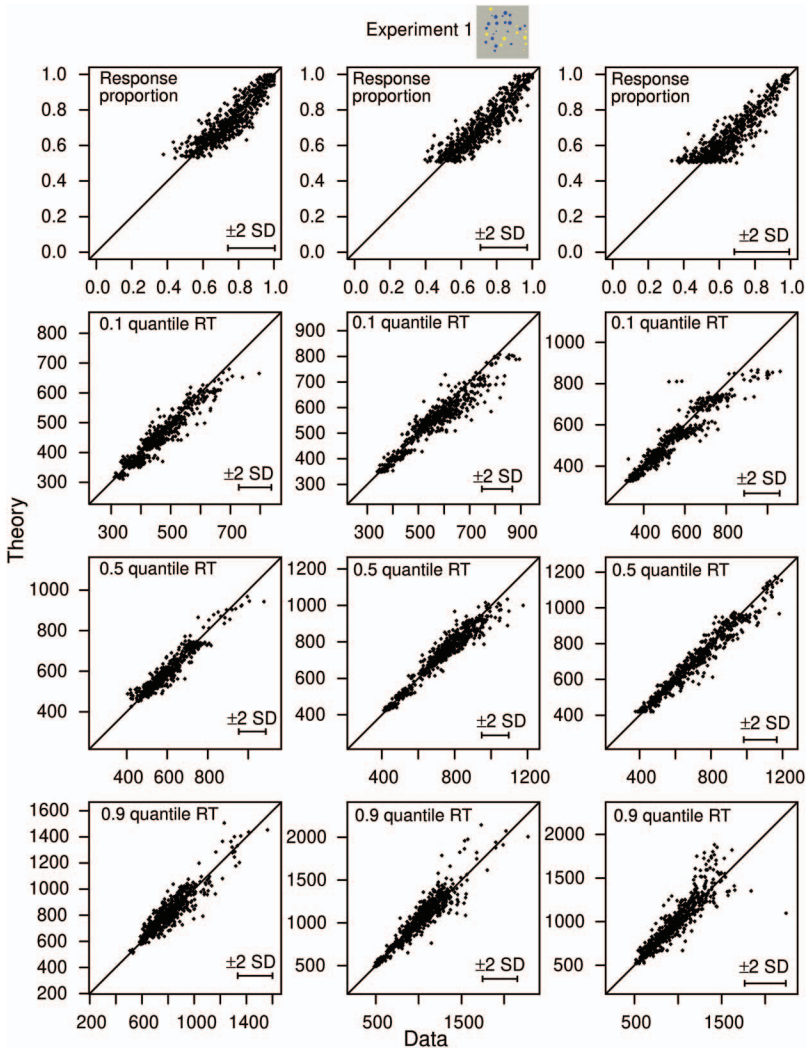

Figure 6. Plots of accuracy, the 0.1, 0.5 (median), and 0.9 quantile correct response times (RTs) for every subject and every condition for Experiment 1 (B/Y task) for the young adults, 60–69-year-olds, and 70–90-year-olds. The two-standard-deviation (*SD*) error bars for RTs are computed from a bootstrap method, and for accuracy, they are based on binomial probabilities. See the online article for the color version of this figure.

shapes of the spreads in data points around the diagonal lines are symmetric, which shows no systematic deviations between theory and data. In the bottom right corner are error bars of plus or minus two *SD* on the data. For accuracy, we used a simple binomial *SD* (sqrt[$p\{1 - p\}/N$]) with $p = .7$. For RT quantiles, the *SD*s were computed using a bootstrap method: Random samples of the RTs were selected with replacement from the RTs from each condition (the number selected was equal to the number of RTs). Then the 0.1, 0.5, and 0.9 quantiles were computed, and this process was repeated 100 times. The two *SD*s were the means over conditions from the *SD*s computed from the 100 quantiles. Generally, the spread of the data was about the same and in the plus or minus two *SD* range.

The next paragraphs give the results from analyses of variance for the model parameters for the two experiments (the means are shown in Table 3 and the *SD*s across subjects in Table 4). The age effects in boundary separation were significant for the B/Y task, $F(2, 87) = 6.9$, $p = .0017$, $\eta_p^2 = .136$, and for the L/R task, $F(2, 87) = 9.9$, $p = 1.3 \times 10^{-4}$, $\eta_p^2 = .186$. Boundary separation increased from young to older adults but not from 60–69- to 70–90-year-old adults. Age effects in nondecision time were significant for the B/Y task, $F(2, 87) = 3.9$, $p = .024$, $\eta_p^2 = .082$, and for the L/R task, $F(2, 87) = 4.6$, $p = .013$, $\eta_p^2 = .095$. As for boundary separation, nondecision time was shorter for young adults than older adults, but the differences between 60−69- and 70–90-year-old adults was small.

For the drift-rate coefficients, for the B/Y task, age was significant, $F(2, 87) = 9.0$, $p = 2.9 \times 10^{-4}$, $\eta_p^2 = .083$, and area was significant, $F(1, 87) = 333.5$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .405$, but the interaction between age and area was not significant, $F(2, 87) = 1.6$, $p > .05$. For the L/R task, age was significant, $F(2, 87) = 3.2$, $p = .044$, $\eta_p^2 = .057$, area was significant, $F(1, 87) = 177.4$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .115$, and the interaction between age and area was significant, $F(2, 87) = 4.5$, $p = .013$, $\eta_p^2 = .006$. These results

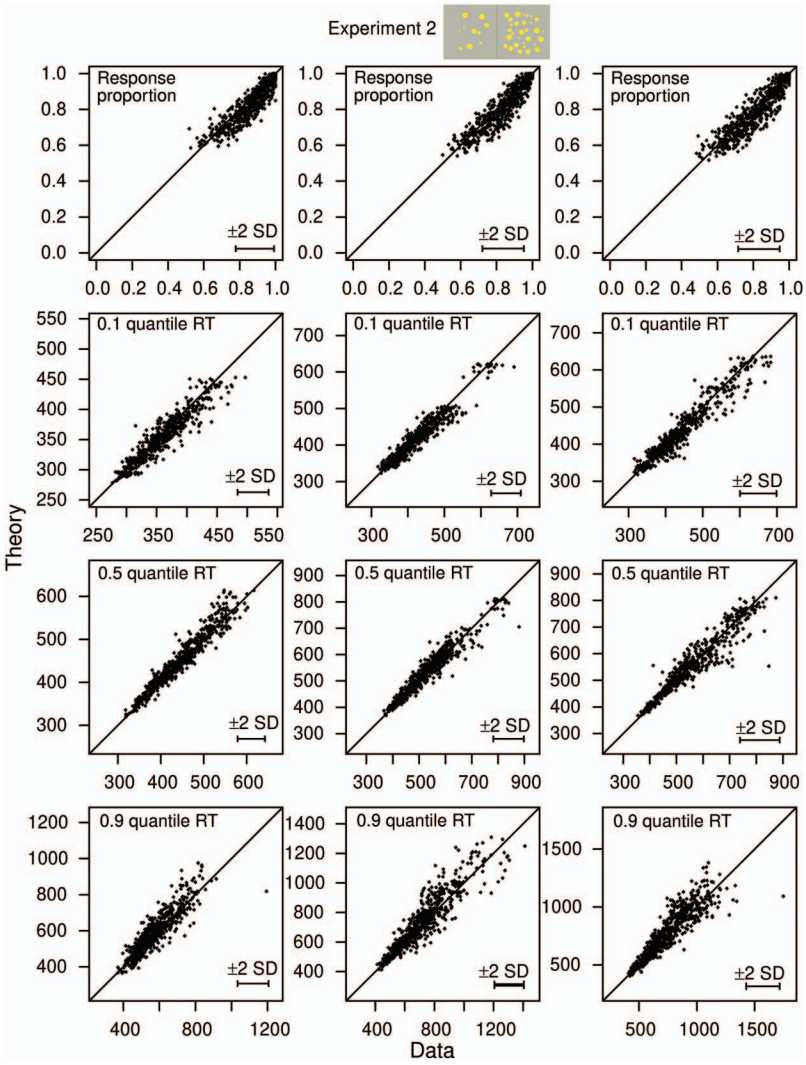

Figure 7. Plots of accuracy, the 0.1, 0.5 (median), and 0.9 quantile response times (RTs) for every subject and every condition for Experiment 2 (L/R task) for the young adults, 60–69-year-olds, and 70–90-year-olds. The two-standard-deviation (*SD*) error bars for RTs are computed from a bootstrap method, and for accuracy, they are based on binomial probabilities. See the online article for the color version of this figure.

show that both the proportional-area and equal-area drift-rate coefficients decrease with age. Furthermore, the difference between the proportional-area and equal-area drift-rate coefficients was larger for the B/Y task than the L/R task, replicating the results for young adults from Ratcliff and McKoon (2018). For the B/Y task, the ratio was 2:1 for young adults and 3:1 for both groups of older adults. For the L/R task, the ratio was 1.2:1 for young adults and 1.5:1 for both groups of older adults. We found age effects on drift-rate coefficients for both tasks but an interaction of age with area only for the L/R task and not the B/Y task. This interaction showed a larger difference between the equal-area and proportional-area coefficients for older adults than for young adults.

None of the other parameter differences were significant (Fs <1.6) except for the constant drift-rate coefficient for the L/R task, $F(2, 87) = 3.5$, $p = .034$, $\eta_p^2 = .075$. Even though this coefficient was significant for the L/R task, it did not change in a regular way (it had a larger value for 60–69-year-olds than the other two age groups).

Although we do not give statistics comparing data and parameter values between the two tasks, we can see large differences. Table 2 shows that accuracy was higher for the L/R task than the B/Y task by between 0.1 and 0.2. Mean RTs were about 150 ms shorter for the L/R than the B/Y task. This is reflected in model parameters with boundary separation lower for the L/R task than the B/Y task, and nondecision time was about 100 ms lower for the L/R task than the B/Y task. Some, but not all, of this difference in mean RTs may be due to the smaller stimulus presentation time for the L/R task (250 ms) than the B/Y task (300 ms). For both the linear model and the log model fits to the data from the two tasks, the drift-rate coefficients were 2–3 times larger for the L/R task than the B/Y task. Overall, the B/Y task was more difficult than the L/R task and the model parameters reflect this difference.

Correlational/Individual Differences Analyses

Ratcliff and McKoon (2018), Ratcliff, Thompson, and McKoon (2015), and Thompson, Ratcliff, and McKoon (2016) examined individual differences in diffusion model parameters for several numerosity tasks. In those studies, each parameter, boundary separation, nondecision time, and drift rates, was significantly correlated across the different tasks. Here, correlation coefficients between the B/Y and L/R tasks are shown in Table 5. For the B/Y and L/R tasks with 30 subjects per group, there were 28 (N-2) degrees of freedom, and correlation coefficients were significant at 0.36 and 0.31 for two-tailed and one-tailed tests, respectively, at the 0.05 level.

Boundary separation and nondecision time correlated across tasks for all three groups of subjects, with lower correlations for the young adults. Within each task, the correlations between subjects' drift-rate coefficients for the proportional-area and equal-area conditions were high, with means over the three subject groups of 0.79 and 0.84 for the B/Y and L/R tasks, respectively. In other words, if someone has a high drift rate for the equal-area condition, they also had a high drift rate for the proportional-area condition. The drift-rate coefficients were significantly correlated across the two tasks for the young adults and 60–69-year-olds, but for 70–90-year-olds, they were not significant. Generally, the results are similar to those in Ratcliff and McKoon (2018), Ratcliff, Thompson, and McKoon (2015), and Thompson et al. (2016).

Convex-Hull Analyses

Norris, Clayton, Gilmore, Inglis, and Castronovo (2019) pointed out that when stimuli are constructed in the usual way (outlined in Halberda et al., 2008), including the two experiments reported here, placing dots in random positions will result in differences in the sizes of convex hulls among stimuli. The way to think about convex hull is to view it as the circumference around or area enclosed by a rubber band stretched around the outer dots of a stimulus. Norris et al. tested the effects of convex-hull area with a task like our L/R task. They divided the stimuli into those for which area and numerosity were congruent, that is, the larger numerosity had the larger area, and those for which they were incongruent, that is, the smaller numerosity had the larger area. For the young adults, accuracy was 0.83 for congruent stimuli and 0.73 for incongruent stimuli, and for the older adults, accuracy values were 0.83 and 0.71, showing a significant effect of convex-hull area, which was about the same size for the young and older adults.

Table 5
Correlations Between Model Parameters for the Linear Model for the B/Y Task and the Log Model for the L/R Task

Parameter	Young adults	60–69-year-olds	70–90-year-olds
a	0.39	0.64	0.53
T_{er}	0.37	0.62	0.73
v_p	0.36	0.64	0.21
v_e	0.41	0.52	0.21

Note. The parameters were boundary separation a, starting point $z = a/2$, and mean nondecision component of response time T_{er}. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition.

We examined the effect of convex-hull area in our L/R and B/Y experiments. We used the matlab routine "convhull" to compute convex-hull areas for the left and right displays for the L/R task and the yellow and blue dots for the B/Y task separately. For both tasks, a larger numerosity was almost always associated with a larger convex-hull area for the conditions in which the differences in numerosity were 10 and 20 (89% of trials with many subjects producing less than 10 and often no responses for the larger hull and smaller numerosity conditions). Therefore, we restricted our analyses to the subset of the data for conditions with differences of five. In our experiments, 26% and 38% of the trials were incongruent for the L/R task and B/Y task, respectively.

For ease of understanding the results, we computed the radius of a circle with the same area as that of a convex hull, which allowed us to compare the areas for the congruent and incongruent trials. This also provided measures of the sizes of the squares within which the dots were presented. Note that a line of 100 pixels was 1.88 cm long. For the B/Y task the mean radius was 173 pixels for congruent trials and 162 pixels for incongruent and for the L/R task, 112 and 107 pixels for congruent and incongruent trials, respectively. For the B/Y task, the length of the side of the square in which the dots were placed was 392 pixels, and the length of the side of each of the squares for the L/R task was 262 pixels.

Table 6 shows the results for the two experiments. First, for the L/R task, accuracy for congruent stimuli was higher than for incongruent ones for all three groups of subjects, about an 8% difference averaged across subjects. RTs varied little between congruent and incongruent stimuli; differences were only about 10 ms. These results replicated those of Norris et al. described above.

The results for the B/Y task were unexpected. Accuracy over subjects for each of the age groups was higher for the incongruent condition, by about 4%, and RT was shorter, by about 20 ms. This is the opposite direction from the L/R task. The question is this: What does it mean to be incongruent in the B/Y task? One possibility is that subjects are responding to density, not the area of the convex hull. With incongruent stimuli, if the convex-hull area is smaller, then the dots are packed together in the smaller area. This suggests that the perceptual cue being used is not convex-hull area but instead density. (To reiterate, there was little effect of age on the results of either experiment).

We did not fit the data from the convex-hull analyses with the models because the numbers of observations in some of the conditions were too small to provide quantiles, especially in the conditions with larger numerosity differences (some with zero observations). However, we did generate predictions using the parameters in Table 3 and assuming that the convex-hull effect was in drift rate. For the B/Y task (fit by the linear model), with numerosity differences of five, a change of about 0.006 in the drift-rate coefficient gave about the 4% difference in accuracy shown in Table 6, but there was only about a 10 ms change in mean RTs. For the L/R task (fit by the log model) with numerosity differences of five, a change of about 0.25 in the drift-rate coefficient gave about the 8% difference in accuracy shown in Table 6 and less than a 10 ms change in mean RT. These results show that drift-rate differences can account for most of the effects of convex hull on accuracy and RT in the same way that drift-rate differences account for the effects of area.

Table 6
The Effect of Convex Hull on Accuracy and Mean RT for Numerosity Differences of Five

Task	Convex hull	Proport.-area young adults		Equal-area young adults		Proport.-area 60–69-year-olds		Equal-area 60–69-year-olds		Proport.-area 70–90-year-olds		Equal-area 70–90-year-olds	
		Accuracy	Mean RT	Accuracy	Mean RT	Accuracy	Mean RT	Accuracy	Mean RT	Accuracy	Mean RT	Accuracy	Mean RT
B/Y	Congruent	0.728	629	0.611	657	0.705	763	0.557	800	0.636	728	0.527	748
	Incongruent	0.766	617	0.665	637	0.730	741	0.624	783	0.679	705	0.573	717
L/R	Congruent	0.864	473	0.817	485	0.845	563	0.761	592	0.819	577	0.719	600
	Incongruent	0.782	480	0.738	487	0.769	576	0.688	605	0.724	586	0.645	604

Note. RT = response time. The B/Y task is Experiment 1, and the L/R task is Experiment 2. For the B/Y task, analysis of variance on accuracy with two factors, the two area conditions and age, showed a significant effect of age, $F(2, 87) = 18.8$, $p = 1.6 \times 10^{-7}$, $\eta_p^2 = .130$; area, $F(1, 87) = 442.7$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .342$; and convex hull, $F(1, 87) = 90.4$, $p = 4.0 \times 10^{-15}$, $\eta_p^2 = .054$, and a significant interaction between convex hull and area, $F(2, 87) = 4.6$, $p = .034$, $\eta_p^2 = .002$. The other interactions were not significant, with Fs < 1.9. The L/R task showed similar effects on accuracy with significant effects of age, $F(2, 87) = 11.5$, $p = 3.7 \times 10^{-5}$, $\eta_p^2 = .098$; area, $F(1, 87) = 188.0$ $p < 2 \times 10^{-16}$, $\eta_p^2 = .145$; and convex hull, $F(1, 87) = 153.4$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .175$, and a significant interaction between age and area, $F(2, 87) = 7.3$, $p = .0011$, $\eta_p^2 = .011$. The other interactions were not significant, with Fs < 1.8. Similar results were obtained for mean RTs. For the B/Y task, analysis of variance with the two area conditions and age showed a significant effect of age, $F(2, 87) = 7.0$, $p = .0016$, $\eta_p^2 = .134$; area, $F(1, 87) = 73.4$, $p = 3.5 \times 10^{-13}$, $\eta_p^2 = .008$; and convex hull, $F(1, 87) = 47.8$, $p = 7.5 \times 10^{-10}$, $\eta_p^2 = .004$, and a significant interaction between age and area, $F(2, 87) = 4.1$, $p = .020$, $\eta_p^2 = .001$. The other interactions were not significant, with Fs < 2.7. The L/R task showed similar effects on mean RT with significant effects of age, $F(2, 87) = 14.4$, $p = 3.9 \times 10^{-6}$, $\eta_p^2 = .239$; area, $F(1, 87) = 34.9$, $p < 2 \times 10^{-16}$, $\eta_p^2 = .013$; and convex hull, $F(1, 87) = 22.4$, $p = 8.4 \times 10^{-6}$, $\eta_p^2 = .002$, and a significant interaction between age and area, $F(2, 87) = 7.4$, $p = .0011$, $\eta_p^2 = .002$. None of the other interactions were significant, with Fs < 2.6.

Discussion

The two experiments in this article examined the effects of age and perceptual variables (area and convex hull) on two numerosity discrimination tasks with integrated ANS-diffusion models. The linear model fit data from the B/Y task better than the log model and the log model fit data from the L/R task about the same as the linear model, except for the 70–90-year-old adults, for whom the linear model fit a little better. For both models, older adults had wider boundaries and longer nondecision times than young adults, results that have been obtained in many other experiments (e.g., Ratcliff, Thapar, Gomez et al., 2004; Ratcliff, Thapar, & McKoon, 2001, 2003, 2004, 2010, 2011; Reike & Schwarz, 2019; Spaniol, Madden, & Voss, 2006). Wider boundaries mean that older adults required more information before making decisions than young adults. Boundary settings are assumed to be under the control of subjects, which means that it may be possible to change the settings. In earlier research with other tasks, we have manipulated speed-accuracy instructions and found that older adults can increase their speed, but sometimes it took one or even two sessions of training to accomplish this (Ratcliff et al., 2001, 2003; Ratcliff, Thapar, & McKoon, 2004; Thapar, Ratcliff, & McKoon, 2003). The longer nondecision times mean that the processes of encoding a stimulus, extracting decision-relevant information from the stimulus, and making a response took longer for the older adults than for young adults.

To determine drift rates, the ANS models use a coefficient that multiplies the difference between two numerosities for the linear model and the difference between the logs of two numerosities for the log model. The linear and log models set the means of the Gaussian distributions around each numerosity (Figure 1C). Equal-area stimuli were more difficult than proportional ones and this was reflected in a smaller coefficient for equal-area than proportional-area stimuli, which made the differences between the means of the distributions smaller for equal-area than proportional-area stimuli. For example, the difference between 30 and 40 dots on the *x*-axis in Figure 1C was smaller for equal-area compared to proportional-area stimuli.

Age had a modest effect on drift-rate coefficients, indicating that the evidence on which decisions were made was lower, but only modestly, for the older adults than the young adults. The size of this difference is a little larger than that seen in item recognition, lexical decision, and other numerosity and perceptual tasks (Ratcliff, Thapar, Gomez et al., 2004; Ratcliff et al., 2001, 2003; Ratcliff, Thapar, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2007, 2010, 2011), but smaller than for letter discrimination and associative recognition (Ratcliff et al., 2011; Thapar et al., 2003). The differences between the equal-area and proportional-area coefficients differed with age. For young adults, the equal-area drift-rate coefficient was about half the size of the proportional-area coefficient for the B/Y task and about 80% of the size for the L/R task (replicating Ratcliff & McKoon, 2018). For both groups of older adults, this ratio fell to about 33% and 66% for the B/Y and L/R tasks, respectively, which suggests that the older adults relied on nonnumerosity variables to a greater degree than the young adults in these tasks.

Cappelletti et al. (2014) found that accuracy for proportional-area stimuli in an experiment similar to our B/Y task did not differ with age, but accuracy for equal-area stimuli did. In contrast, our experiments found that accuracy for proportional-area stimuli decreased with age. They also found a difference in mean RTs for proportional-area versus equal-area stimuli for older adults (617 ms vs. 649 ms), but not for young adults (415 ms vs. 427 ms). Our results in Table 2 show a similar interaction. The size of these differences could be the result of scaling: Because older adults were overall slower than young adults, RT differences are magnified because differences become smaller as RT approaches floor. In the diffusion model, this is because boundaries are lower for young adults and this produces smaller differences in RTs among conditions (a scaling effect; Ratcliff, Spieler, & McKoon, 2000, 2004).

The *SD*s in the drift rates across trials are produced by a coefficient that multiplies the square root of the sum of squares of the two numerosity values (Figure 1C) plus a constant. As discussed below, the *SD* coefficient for the linear model must increase

with numerosity in order for the model to explain why both accuracy and RT decrease as numerosity increases (for a small constant difference between two numerosities, e.g., five). The assumption for the log model is that the *SD* is constant, but in fitting the data, we allowed it to change with numerosity in the same way as for the linear model in order to give it the same flexibility as the linear model. As we discussed above, results showed that the model best able to account for the data depended on the task: The linear model fit the data from the B/Y task a little better than the log model, but the log model fit data from the L/R task about the same as the linear model except for the 70–90-year-old group. In the B/Y task, the *SD* coefficient produced large changes in across-trial *SD* in drift rate, and there was no significant effect of age on the *SD* coefficient. In the L/R task, the coefficient was small and produced little difference in across-trial *SD* in drift rate (which is consistent with the assumption of a constant *SD* for the log model).

There is one deviation between theory and data that we can speculate about. This is the somewhat U-shaped functions for differences of five that are shown in Figures 3D–3F for the L/R task. For the 60–69-year-old adults, the functions for differences of five show RTs decreasing as accuracy increases, except for the 15–10 numerosity conditions for both proportional-area and equal-area conditions (labeled in Figure 3E). The same deviation may occur for young adults, but it is very small. For 70–90-year-olds (Figure 3F), for equal-area conditions (Xs), RT decreases as accuracy decreases, which suggests that this group is performing the task in a way more consistent with the linear model. This suggests that the pattern of results might represent a probability mixture of processes consistent with the log model and of processes consistent with the linear model, with the log model dominating for young adults and the linear model dominating for 70–90-year-olds.

In a more comprehensive study of the effects of perceptual and numerosity variables, Kang and Ratcliff (2020) modeled the joint effects of multiple combinations of perceptual and numerosity variables on accuracy and RT. In two experiments using the B/Y task, they collected data from combinations of area and numerosity, and in another experiment, they collected data from an experiment with combinations of area, convex hull, and numerosity from the L/R task. In that L/R task, they found that convex hull had a large effect on accuracy and RT, whereas area had only a small effect (as in Ratcliff & McKoon, 2018). Models of drift rate were examined that included numerosity and perceptual components, and the models that were most successful had more components than the two components used in the models presented in this article (i.e., more than just the drift-rate coefficients for proportional-area and equal-area conditions). Kang and Ratcliff found that when the perceptual variables were in conflict with numerosity, a new conflict effect was obtained, with accuracy less than chance and a delay in the leading edge of the RT distributions that was too large to be accounted for by only changes in drift rate (cf., Ratcliff & Frank, 2012). In order to model below-chance accuracy, Kang and Ratcliff had to include interaction terms between numerosity and perceptual variables and terms representing the ratio of the difference in numerosities over the sum. This approach, like the models described here, allows the effects of numerosity and one or more perceptual variables to be separately estimated. DeWind, Adams, Platt, and Brannon (2015) performed a similar study but only modeled the effects of perceptual variables on accuracy (Kang and Ratcliff compared the two approaches).

The manipulation of the numbers of dots in our B/Y task is similar to manipulations of magnitude in perceptual tasks. Magnitude effects in brightness and motion discrimination tasks have received attention because they have been thought to be inconsistent with the usual assumptions in evidence-accumulation models, including the standard diffusion model. The problem is that as the overall magnitude of two stimuli to be discriminated increases (for a constant difference between the stimuli), accuracy and RT both decrease (Niwa & Ditterich, 2008; Teodorescu, Moran, & Usher, 2016; Teodorescu & Usher, 2013). This pattern cannot be accommodated by the usual explanation in which drift rate changes as a function of difficulty because accuracy decreases as drift rate decreases but RT increases. However, this is the pattern of results that the linear model used here predicts; thus, in order to produce data for modeling, Ratcliff et al. (2018) replicated these results with experiments with brightness and motion discrimination tasks. They found that the linear model, in which drift rate was a function of the difference in magnitude and across-trial *SD* in drift rate was a function of the sum of squares of the two magnitudes, produced reasonable fits to the experimental data. (The log model was not considered because it failed to fit the qualitative patterns of results.)

An alternative hypothesis for decreasing RT with decreasing accuracy in the brightness, motion, and B/Y tasks is one in which within-trial variability (noise in the accumulation process) increases with stimulus strength (e.g., Donkin, Brown, & Heathcote, 2009; Niwa & Ditterich, 2008; Smith & Ratcliff, 2009; Teodorescu et al., 2016; Teodorescu & Usher, 2013). If within-trial variability increased, the evidence accumulation process will hit a decision boundary earlier, resulting in less accuracy and, importantly, in shorter RTs. The intuition for this can be seen in Figure 1D: If the jagged lines had larger steps up and down vertically, then they would hit a boundary earlier and may hit the wrong boundary by mistake. Ratcliff et al. (2018) fit this model to their data and found similar fits as the linear model.

In most applications of the diffusion model, it is assumed that within-trial variability is constant across levels of difficulty. If within-trial variability were to increase with numerosity or stimulus magnitude, then it would be expected to be a property of numerosity or perceptual magnitude, and it should increase in all numerosity and perceptual tasks with similar stimuli. However, model-based analyses showed that this does not occur in the L/R task for numerosity and Ratcliff and McKoon (2018) and Ratcliff et al. (2018) showed that within-trial variability was almost constant in single-item perceptual and numerosity tasks. Thus, we believe that changes in across-trial variability in drift rate should be preferred over within-trial variability changing with stimulus magnitude or overall numerosity.

The results of this study and that of Ratcliff and McKoon (2018) show that the joint behavior of RTs and accuracy depends on the task, the cognitive representations of numerosities on which performance is based depend on the task, and how much perceptual variables affect performance depends on the task. This illustrates the remarkably different ways that the cognitive system deals with numerosity information. Depending on the task, it encodes numerosities on a linear scale or a log scale; it encodes them with variability in their representations changing with numerosity when

a linear scale is used but not when a log scale is used; and it includes information other than number (e.g., area) to a great deal in some tasks (B/Y) but much less in other tasks (L/R and single array tasks).

As discussed above, research on numeracy has been concerned with whether experimental results can be explained by numerosity alone, without some confounding variable such as area, convex hull, or density (e.g., DeWind et al., 2015; DeWind & Brannon, 2012; Feigenson, Carey, & Hauser, 2002; Gebuis, Cohen Kadosh, & Gevers, 2016; Gebuis & Gevers, 2011; Gebuis & Reynvoet, 2012a, 2012b, 2013; Leibovich, Katzin, Harel, & Henik, 2017; Mix, Huttenlocher, & Levine, 2002). Efforts to control for such variables face the problem that controlling for one leaves another confounded. Our results show, first, that we can estimate the contributions of perceptual variables and numerosity separately; second, that the effect of the perceptual variable, area, is task dependent (there is a larger effect of area in the B/Y task than the L/R task), and the difference increases with age; and third, that the effect of convex hull is task dependent. Perhaps the most important of the effects we discussed above is that encoded representations of numeracy differ as a function of the task.

Considerable controversy has arisen about the presence or absence of correlations among dependent variables in numerosity discrimination tasks and between them and individual differences such as IQ and math ability (see the comprehensive analyses and meta-analyses in Chen & Li, 2014; Gilmore, Attridge, & Inglis, 2011; Halberda et al., 2012; Price, Palmer, Battista, & Ansari, 2012). Sometimes RTs are used, sometimes accuracy, and sometimes the slope of a function that relates accuracy or RTs to the difficulty of a test item, and this inconsistency in the empirical measures used has led to inconsistent findings about how differences among individuals affect performance. For example, sometimes correlations are found between symbolic tasks and nonsymbolic tasks and sometimes not (e.g., De Smedt, Verschaffel, & Ghesquière, 2009; Holloway & Ansari, 2009; Maloney, Risko, Preston, Ansari, & Fugelsang, 2010; Price et al., 2012; Sasanguie, Defever, Van den Bussche, & Reynvoet, 2011). Sometimes correlations are found between nonsymbolic number tasks and math ability, and sometimes not (e.g., Gilmore, McCarthy, & Spelke, 2010; Halberda et al., 2012, 2008; Holloway & Ansari, 2009; Inglis, Attridge, Batchelor, & Gilmore, 2011; Libertus, Feigenson, & Halberda, 2011; Lyons & Beilock, 2011; Mundy & Gilmore, 2009; Price et al., 2012). We believe that model-based analyses such as the ANS-diffusion model approach can provide a coherent view of all the dependent variables including how accuracy and RT relate to each other and so provide a unification of the measures. This approach provides tools with which to examine the effects of development, aging, and dysfunction on numeracy abilities and the relationships among numeracy measures as well individual difference measures—for example, achievement scores.

References

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology, 18,* 207–225. http://dx.doi.org/10.1111/j.2044-8317.1965.tb00342.x

Boyle, P. A., Yu, L., Wilson, R. S., Segawa, E., Buchman, A. S., & Bennett, D. A. (2013). Cognitive decline impairs financial and health literacy among community-based older persons without dementia. *Psychology and Aging, 28,* 614–624. http://dx.doi.org/10.1037/a0033103

Bruine de Bruin, W., McNair, S. J., Taylor, A. L., Summers, B., & Strough, J. (2015). "Thinking about numbers is not my idea of fun": Need for cognition mediates age differences in numeracy performance. *Medical Decision Making, 35,* 22–26. http://dx.doi.org/10.1177/0272989X14542485

Cappelletti, M., Didino, D., Stoianov, I., & Zorzi, M. (2014). Number skills are maintained in healthy ageing. *Cognitive Psychology, 69,* 25–45. http://dx.doi.org/10.1016/j.cogpsych.2013.11.004

Charron, C., Fischer, J.-P., & Meljac, C. (2008). Arithmetic after school: How do adults' mental arithmetic abilities evolve with age? *Research in the Schools, 15,* 9–26. Retrieved from http://www.msera.org/publications-rits.html

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. http://dx.doi.org/10.1016/j.actpsy.2014.01.016

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

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

Delazer, M., Kemmler, G., & Benke, T. (2013). Health numeracy and cognitive decline in advanced age. *Aging, Neuropsychology, and Cognition, 20,* 639–659. http://dx.doi.org/10.1080/13825585.2012.750261

De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology, 103,* 469–479. http://dx.doi.org/10.1016/j.jecp.2009.01.010

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. http://dx.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,* 68. http://dx.doi.org/10.3389/fnhum.2012.00068

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16,* 1129–1135. http://dx.doi.org/10.3758/PBR.16.6.1129

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

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging, 25,* 271–288. http://dx.doi.org/10.1037/a0019106

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198. http://dx.doi.org/10.1016/0022-3956(75)90026-6

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67,* 641–666. http://dx.doi.org/10.1146/annurev-psych-122414-033645

Fraenkel, L., Cunningham, M., & Peters, E. (2015). Subjective numeracy and preference to stay with the status quo. *Medical Decision Making, 35,* 6–11. http://dx.doi.org/10.1177/0272989X14532531

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

Gandini, D., Lemaire, P., & Dufau, S. (2008). Older and younger adults' strategies in approximate quantification. *Acta Psychologica, 129,* 175–189. http://dx.doi.org/10.1016/j.actpsy.2008.05.009

Gandini, D., Lemaire, P., & Michel, B. F. (2009). Approximate quantification in young, healthy older adults', and Alzheimer patients. *Brain and Cognition, 70,* 53–61. http://dx.doi.org/10.1016/j.bandc.2008.12.004

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. http://dx.doi.org/10.1016/j.actpsy.2016.09.003

Gebuis, T., & Gevers, W. (2011). Numerosities and space; indeed a cognitive illusion! A reply to de Hevia and Spelke (2009). *Cognition, 121,* 248–252. http://dx.doi.org/10.1016/j.cognition.2010.09.008

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

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

Gebuis, T., & Reynvoet, B. (2013). The neural mechanisms underlying passive and active processing of numerosity. *NeuroImage, 70,* 301–307. http://dx.doi.org/10.1016/j.neuroimage.2012.12.048

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

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,* 394–406. http://dx.doi.org/10.1016/j.cognition.2010.02.002

Griffith, H. R., Belue, K., Sicola, A., Krzywanski, S., Zamrini, E., Harrell, L., & Marson, D. C. (2003). Impaired financial abilities in mild cognitive impairment: A direct assessment approach. *Neurology, 60,* 449–457. http://dx.doi.org/10.1212/WNL.60.3.449

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,* 11116–11120. http://dx.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,* 665–668. http://dx.doi.org/10.1038/nature07246

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,* 17–29. http://dx.doi.org/10.1016/j.jecp.2008.04.001

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience, 15,* 470–476. http://dx.doi.org/10.1038/nn.3017

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

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology, 120,* 101288. http://dx.doi.org/10.1016/j.cogpsych.2020.101288

Kaphingst, K. A., Goodman, M. S., MacMillan, W. D., Carpenter, C. R., & Griffey, R. T. (2014). Effect of cognitive dysfunction on the relationship between age and health literacy. *Patient Education and Counseling, 95,* 218–225. http://dx.doi.org/10.1016/j.pec.2014.02.005

Laming, D. R. J. (1968). *Information theory of choice reaction time*. New York, NY: Wiley.

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

Lemaire, P., & Lecacheur, M. (2007). Aging and numerosity estimation. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 62,* 305–312. http://dx.doi.org/10.1093/geronb/62.6.P305

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for robust parameter estimation in diffusion modeling? A comparison of different estimation algorithms. *Behavior Research Methods, 49,* 513–537. http://dx.doi.org/10.3758/s13428-016-0740-2

Li, Y., Gao, J., Enkavi, A. Z., Zaval, L., Weber, E. U., & Johnson, E. J. (2015). Sound credit scores and financial decisions despite cognitive aging. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 65–69. http://dx.doi.org/10.1073/pnas.1413570112

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science, 14,* 1292–1300. http://dx.doi.org/10.1111/j.1467-7687.2011.01080.x

Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition, 121,* 256–261. http://dx.doi.org/10.1016/j.cognition.2011.07.009

Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica, 134,* 154–161. http://dx.doi.org/10.1016/j.actpsy.2010.01.006

Maylor, E. A., Sheehan, B., Watson, D. G., & Henderson, E. L. (2008). Enumeration in Alzheimer's disease and other late life psychiatric syndromes. *Neuropsychologia, 46,* 2696–2708. http://dx.doi.org/10.1016/j.neuropsychologia.2008.05.002

Maylor, E. A., Watson, D. G., & Muller, Z. (2005). Effects of Alzheimer's disease on visual enumeration. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 60,* 129–135. http://dx.doi.org/10.1093/geronb/60.3.P129

Mix, K., Huttenlocher, J., & Levine, S. C. (2002). *Quantitative development in infancy and early childhood*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195123005.001.0001

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology, 103,* 490–502. http://dx.doi.org/10.1016/j.jecp.2009.02.003

Niwa, M., & Ditterich, J. (2008). Perceptual decisions between multiple directions of visual motion. *The Journal of Neuroscience, 28,* 4435–4445. http://dx.doi.org/10.1523/JNEUROSCI.5564-07.2008

Norris, J. E., Clayton, S., Gilmore, C., Inglis, M., & Castronovo, J. (2019). The measurement of approximate number system acuity across the lifespan is compromised by congruency effects. *The Quarterly Journal of Experimental Psychology, 72,* 1037–1046. http://dx.doi.org/10.1177/1747021818779020

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

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition, 133,* 188–200. http://dx.doi.org/10.1016/j.cognition.2014.06.011

Pertl, M.-T., Benke, T., Zamarian, L., Martini, C., Bodner, T., Karner, E., & Delazer, M. (2014). Do patients with mild cognitive impairment understand numerical health information? *Journal of Alzheimer's Disease, 40,* 531–540. http://dx.doi.org/10.3233/JAD-131895

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. http://dx.doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278–291. http://dx.doi.org/10.3758/BF03196283

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2,* 237–279. http://dx.doi.org/10.1037/dec0000030

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24,* 1186–1229. http://dx.doi.org/10.1162/NECO_a_00270

Ratcliff, R., Huang-Pollock, C., & McKoon, G. (2018). Modeling individual differences in the go/no-go task with a diffusion model. *Decision, 5,* 42–62.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. http://dx.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,* 183–217. http://dx.doi.org/10.1037/rev0000085

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367. http://dx.doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20,* 260–281. http://dx.doi.org/10.1016/j.tics.2016.01.007

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. http://dx.doi.org/10.1177/0963721415596228

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin & Review, 7,* 1–25. http://dx.doi.org/10.3758/BF03210723

Ratcliff, R., Spieler, D., & McKoon, G. (2004). Analysis of group differences in processing speed: Where are the models of processing? *Psychonomic Bulletin & Review, 11,* 755–769. http://dx.doi.org/10.3758/BF03196631

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19,* 278–289. http://dx.doi.org/10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16,* 323–341. http://dx.doi.org/10.1037/0882-7974.16.2.323

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65,* 523–535. http://dx.doi.org/10.3758/BF03194580

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50,* 408–424. http://dx.doi.org/10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75–90 years old. *Psychology and Aging, 22,* 56–66. http://dx.doi.org/10.1037/0882-7974.22.1.56

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60,* 127–157. http://dx.doi.org/10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140,* 464–487. http://dx.doi.org/10.1037/a0023810

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137,* 115–136. http://dx.doi.org/10.1016/j.cognition.2014.12.004

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,* 438–481. http://dx.doi.org/10.3758/BF03196302

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300. http://dx.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. http://dx.doi.org/10.1016/j.cogpsych.2018.02.002

Reike, D., & Schwarz, W. (2019). Aging effects on symbolic number comparison: No deceleration of numerical information retrieval but more conservative decision-making. *Psychology and Aging, 34,* 4–16. http://dx.doi.org/10.1037/pag0000272

Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (2011). The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. *Acta Psychologica, 136,* 73–80. http://dx.doi.org/10.1016/j.actpsy.2010.10.004

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116,* 283–317. http://dx.doi.org/10.1037/a0015156

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 101–117. http://dx.doi.org/10.1037/0278-7393.32.1.101

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25,* 377–390. http://dx.doi.org/10.1037/a0018022

Szrek, H., & Bundorf, M. K. (2013). Age and the purchase of prescription drug insurance by older adults. *Decision, 1,* 104–123. http://dx.doi.org/10.1037/2325-9965.1.S.104

Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: Violations of value invariance in human decision making. *Psychonomic Bulletin & Review, 23,* 22–38. http://dx.doi.org/10.3758/s13423-015-0858-8

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review, 120,* 1–38. http://dx.doi.org/10.1037/a0030776

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging, 18,* 415–429. http://dx.doi.org/10.1037/0882-7974.18.3.415

Thompson, C. A., Ratcliff, R., & McKoon, G. (2016). Individual differences in the components of children's and adults' information processing for simple symbolic and non-symbolic numeric decisions. *Journal of Experimental Child Psychology, 150,* 48–71. http://dx.doi.org/10.1016/j.jecp.2016.04.005

Trick, L. M., Enns, J. T., & Brodeur, D. A. (1996). Life span changes in visual enumeration: The number discrimination task. *Developmental Psychology, 32,* 925–932. http://dx.doi.org/10.1037/0012-1649.32.5.925

Triebel, K. L., Martin, R., Griffith, H. R., Marceaux, J., Okonkwo, O. C., Harrell, L., . . . Marson, D. C. (2009). Declining financial capacity in mild cognitive impairment: A 1-year longitudinal study. *Neurology, 73,* 928–934. http://dx.doi.org/10.1212/WNL.0b013e3181b87971

Tuerlinckx, F., Maris, E., Ratcliff, R., & De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior, Research, Instruments, and Computers, 33,* 443–456.

Vickers, D., Caudrey, D., & Willson, R. J. (1971). Discriminating between the frequency of occurrence of two alternative events. *Acta Psychologica, 35,* 151–172. http://dx.doi.org/10.1016/0001-6918(71)90018-7

Watson, D. G., Maylor, E. A., & Bruce, L. A. M. (2005). Search, enumeration, and aging: Eye movement requirements cause age-equivalent performance in enumeration but not in search tasks. *Psychology and Aging, 20,* 226–240. http://dx.doi.org/10.1037/0882-7974.20.2.226

Watson, D. G., Maylor, E. A., & Manson, N. J. (2002). Aging and enumeration: A selective deficit for the subitization of targets among distractors. *Psychology and Aging, 17,* 496–504. http://dx.doi.org/10.1037/0882-7974.17.3.496

Wechsler, D. (1997). *The Wechsler Adult Intelligence Scale III*. San Antonio, CA: Psychological Corporation, Harcourt Brace.

Wood, S., & Hanock, Y. (2012). The impact of numeracy on Medicare Part D insurance choice in older adults. In D. J. Lamdin (Ed.), *Consumer knowledge and financial decisions: Lifespan perspectives* (pp. 255–267). New York, NY: Springer Science + Business Media.

Zorzi, M., Stoianov, I., & Umilta, C. (2005). Computational modeling of numerical cognition. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 67–84). New York, NY: Psychology Press.

Appendix A

Description of the Diffusion Model

In the diffusion model (and other sequential sampling models; Ratcliff & Smith, 2004), an individual must decide whether to respond more accurately, sacrificing speed, or faster, sacrificing accuracy. The model separates this component of the decision process that sets speed-accuracy criteria from the other main components, namely, the quality of the information upon which decisions are based and from nondecision time, which is the sum of the time taken to encode a stimulus, convert it to decision-relevant information, and the time to execute a response. In this model, accuracy and RTs are explained by a single mechanism (Ratcliff, 1978; Ratcliff & McKoon, 2008) that is the noisy accumulation of information from a stimulus representation over time (Figure 1D). A response is made when the amount of accumulated information reaches one or the other of two criteria, or boundaries, one for each of the two choices. The rate of accumulation, called drift rate, is determined by the quality of the information from the stimulus or memory, depending on the task. The distance between the two boundaries is determined by an individual's speed/accuracy setting—faster, less accurate responses if the distance is small and slower, more accurate responses if the distance is large. The model is required to account for the locations of RT distributions and their characteristic right-skewed shape of the distributions and the effects of experimental variables on RTs and accuracy (e.g., Ratcliff, Smith, & McKoon, 2015).

The model has been applied in a wide range of domains, including aging, aphasia, sleep deprivation, child development, hypoglycemia, anxiety, depression, language, ADHD, dyslexia, Parkinson's disease, and in research fields such as neuroeconomics and neuroscience in humans, monkeys, rodents, and even insect swarms (see reviews in Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff, Smith, Brown, & McKoon, 2016). For aging research, Ratcliff, Thapar, and McKoon (e.g., Ratcliff et al., 2001, 2003; Ratcliff, Thapar, & McKoon, 2004, 2010, 2011) showed that in many tasks, older adults' long RTs relative to young adults' do not come from the information they encode from stimuli being of poorer quality than young adults' or from a general slowing of all (or most) cognitive processes but instead from their concern, much more than young adults, not to make errors (Starns & Ratcliff, 2010).

Figure 1D illustrates the model. The accumulation of information begins from a starting point, z, toward one or the other of the two boundaries, a or 0. The zig-zag lines illustrate noise in the accumulation process. For the example in the figure, the mean rate of accumulation, drift rate (v), is positive, with some processes finishing quickly, some slowly, and some hitting the wrong boundary by mistake. Total RT is the sum of the time to reach a boundary and nondecision time (T_{er}).

The values of the components of processing in the diffusion model are assumed to vary from trial to trial, under the assumption that subjects cannot accurately set the same parameter values from one trial to another (e.g., Laming, 1968; Ratcliff, 1978). Across-trial variability in drift rate is normally distributed with SD η, across-trial variability in starting point (equivalent to across-trial variability in the boundaries) is uniformly distributed with range s_z, and across-trial variability in the nondecision component is uniformly distributed with range s_t. In signal detection theory, which deals only with accuracy, all sources of across-trial variability are collapsed into one parameter, the variability in information across trials. In contrast, with the diffusion model, there are separate sources of across-trial variability. In the integrated diffusion models for numerosity presented here, across-trial variability in drift rate is explicitly represented as in Figure 1C.

Boundary settings, nondecision time, starting point, drift rates for each condition in an experiment that varies in difficulty, and the across-trial variabilities in drift rate, nondecision time, and starting point are all identifiable with enough observations (Ratcliff & Tuerlinckx, 2002). If exact predictions are entered into fitting programs, the generating parameter values are recovered. When data are simulated from the model (with numbers of observations approximately equal to those that would be obtained in real experiments) and the model is fit to the simulated data, the parameters used to generate the data are recovered with variability that is usually several times smaller than individual differences in the model parameters for drift rate, nondecision time, and boundary separation, but not for the variability parameters (Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002). This is examined for these integrated models in Appendix D below. Also, Kang and Ratcliff (2020) presented a parameter recovery study for the integrated model and showed that with enough observations, there were few biases. The success of parameter identifiability comes in part from the strong constraint that the model must account for the full distributions of RTs for correct and error responses over all the conditions of the experiment (for a study on model freedom, see Ratcliff, 2002).

(*Appendices continue*)

Appendix B

Fitting the Integrated Diffusion Models to Data

The values of all eight parameters of the model are estimated together by fitting the model to the data from all the conditions in an experiment simultaneously. The method computes a multinomial likelihood G^2 statistic, and parameters of the model are adjusted using a standard SIMPLEX method to maximize the value of G^2. The data for each subject is fit individually, and the model parameters reported are the means across subjects.

For RTs, the models must explain the shapes of the RT distributions. To represent distributions, we divide the empirical RTs into five quantiles, the .1, .3, .5, .7, and .9 quantiles. The quantile RTs and the proportions of responses in each quantile for each condition in the experiment are entered into a minimization routine, and the diffusion model is used to generate the predicted cumulative probability of a response occurring by each quantile RT. Subtracting the cumulative probabilities for each successive quantile from the next higher quantile gives the proportion of responses between adjacent quantiles. For our G^2 computation, these are the expected proportions, to be compared to the observed proportions of responses between the quantiles (i.e., the proportions between 0, .1, .3, .5, .7, .9, and 1.0, which are .1, .2, .2, .2, .2, and .1). The proportions for the observed (p_o) and expected (p_e) frequencies and summing over $2Np_o\log(p_o/p_e)$ for all conditions gives a single G^2 (a log multinomial likelihood) value to be minimized (where N is the number of observations for the condition).

The number of degrees of freedom in the data is computed as follows: There are six proportions (bins) between the quantiles and outside the .1 and .9 quantiles. These proportions are multiplied by the proportion of responses for that condition and across correct and error responses; these 12 proportions must add to 1, so there are 11 degrees of freedom in the data for each condition of the experiment. For example, for 10 numerosity conditions crossed with an area variable that has two levels, there are 220 degrees of freedom in the data. When the models are fit to data, the number of degrees of freedom is the number in the data minus the number of the model's free parameters (eight for these experiments).

The model was fit to the data using the G^2 statistic in the same way as fitting the chi-square method described by Ratcliff and Tuerlinckx (2002; see also Ratcliff & Childers, 2015; Ratcliff & Smith, 2004). G^2 statistics are asymptotically chi-square, so critical chi-square values can be used to assess goodness of fit. In many applications, we have found that if the value of the chi-square (or G^2) is below 2 times the critical value, the fit is good (Ratcliff, Thapar, Gomez et al., 2004; Ratcliff et al., 2010). This rule of thumb has been found in the less constrained case than the ANS-diffusion model; in the less constrained case, the diffusion model is applied without a representation model, so each condition has its own drift rate.

It is worth reiterating that the fits of the ANS-diffusion models to data are rather good given the small number of degrees of freedom in the model and large number in the data. Any of a number of aspects of the data could have been different and produced poor fits of the model to data. For example, if any groups of quantile RTs for one condition (e.g., a difference in numerosity of 10) were moved to the left or right, changing accuracy, or up or down, changing RT quantiles, the model would fail to fit the data because it is constrained to produce exactly the changes in RT and accuracy shown in the fits. With a model with this few free parameters, there is little room for overfitting data.

There are several alternative fitting methods and packages for fitting the diffusion model to data. First, standard maximum likelihood is a good alternative in the absence of outliers and produces fits with lower *SD*s in parameter estimates than the chi-square method. The limitation is that in the presence of outliers, it can produce estimates a long way off the true values. Comparisons of the chi-square and maximum likelihood methods in Kang and Ratcliff (2020) showed that they produced very similar results when there were no obvious outliers. There are also Bayesian methods, and the HDDM package can produce fits with low *SD*s in parameter estimates, but it can occasionally produce spurious values (Ratcliff & Childers, 2015). The problem with this and the other packages is that they do not allow models of across-trial *SD* in drift rates to be implemented. The hierarchical Bayesian model in HDDM requires across-trial variability in drift rate to be constant across conditions and across subjects.

(*Appendices continue*)

Appendix C

Model Selection

In the article, the two models we implemented have equal numbers of parameters. However, the log model could be argued to have a constant across-trial variability in drift rate across conditions and so have one less parameter than the linear model. In this case model-selection methods need to be used. Here, we examine what happens to the number of subjects that are best fit by the linear and log models using the AIC and BIC. This is plausible because in the fits of the log model to data, there are very modest contributions from the nonconstant component of across-trial variability in drift rate. However, for some subjects, this is not negligible; to address this, we refit the model with $\sigma_1 = 0$, and these additional G^2 goodness-of-fit values are shown in Table 3. The mean model parameters changed by less than 5%, most less than 1%, and the interpretations are the same.

The difference in G^2 values between the log and linear models provides a numerical goodness-of-fit measure from which the models can be compared. As noted above, because the number of parameters for the two models was the same, the G^2 values provide the same results for comparisons of models as do AIC and BIC values (because these are the multinomial likelihood G^2 plus a penalty term based on the number of parameters, which is the same for the two models). In our view, small numerical differences are not enough to be sure that one model better accounts for the data than the other, especially because some subjects will be better fit by each model, which, strictly speaking, means that some subjects use one representation, and some the other. (This provides a problem in interpretation because if all the data were generated from one model, variability in the data would sometimes produce better fits of the other model if the two models were similar.) Ratcliff, Thompson, and McKoon (2015) used G^2, AIC, and BIC to compare models and found that the patterns of model selection changed quite dramatically depending on the statistic used. We strongly prefer to see qualitative differences in predictions between the models as well as numerical differences that are not too small. For each experiment presented here, we report the number of subjects that favors each model using the G^2, AIC, and BIC values. By a binomial test, if 20 (or more) out of 30 subjects favor one model over the other, then the result is significant.

The number of subjects out of 30 that fit the linear model better than the log model is as follows. The first number is for G^2, the second for AIC, and the third for BIC: for the B/Y task, for young adults, 24, 23, and 19; for 60–69-year-old adults, 22, 21, and 19; and for 70–90-year-old adults, 23, 22, and 19. For the L/R task, for young adults, 14, 13, and 11; for 60–69-year-old adults, 17, 17, and 16; and for 70–90-year-old adults, 20, 20, and 19. For the linear model and the log model with across-trial variability in drift rate allowed to vary over conditions (the model in the body of the text), the number of subjects that fit the linear model better than the log model in G^2 is as follows: for the B/Y task, for young adults, 24, for 60–69-year-old adults, 21, and for 70–90-year-old adults, 26; for the L/R task, for young adults, 14, for 60–69-year-old adults, 18, and for 70–90-year-old adults, 21. Because the number of parameters is the same for each model, the penalty terms are the same for the two models and AIC and BIC results are the same as G^2 results.

These results show that for the B/Y task, more of the subjects prefer the linear model by all the measures (less for BIC, of course, because of the larger penalty). For the L/R task, the results are mixed. About half of the subjects prefer the linear model for the young adult and 60–69-year-old groups on all the measures, while for the 70–90-year-old group, more than half the subjects prefer the linear model by all the measures.

A second way of assessing model selection is to examine the qualitative signature in the data that separates the linear and log models. For a numerosity difference of five, we can examine the plots of mean RT versus accuracy (as in Figure 3) as a function of which model was preferred, the linear or log. The argument would be that a preference for the log model might lead to RT increasing as accuracy decreases, but a preference for the linear model would lead to RT decreasing as accuracy decreases.

Figure C1 shows plots of mean RT versus accuracy for the two experiments and three age groups with data divided into groups that favored the linear model and those that favored the linear model (by G^2). Note that the data were collapsed over the equal-area and proportional-area conditions. Results show little difference for the B/Y task for the different groups of subjects, with the functions for subjects with data that preferred the log model being quite similar to the functions for subjects with data that preferred the linear model. For the L/R task, there was a small but consistent separation in which subjects with data that preferred the log model had functions in which RT increased as accuracy decreased, whereas subjects with data that preferred the linear model had functions in which RT decreased or was flat as accuracy decreased.

The strongest conclusion from this is that for the B/Y task, it can be argued that all subjects use differences between the numbers of dots to drive the decision process whether or not the linear or log model is the best-fitting model (which model is preferred could be argued to depend on random variability in the data). In contrast, for the L/R task, some subjects use separate representations of the two arrays, and hence the log model applies, while others seem to rely on differences in numbers (as in the B/Y task), and hence the linear model applies. There is an increasing tendency with age for the linear model to apply, which suggests an age-related change from using separate representations to differences.

These conclusions are an oversimplification and are not strongly supported by the data, but the results hint at different modes of processing for the L/R task. However, we believe that the results support the view that the linear model applies in the B/Y task, as in Ratcliff and McKoon (2018), but there is not strong evidence for the superiority of either model for the L/R task, in contrast to the results

(*Appendices continue*)

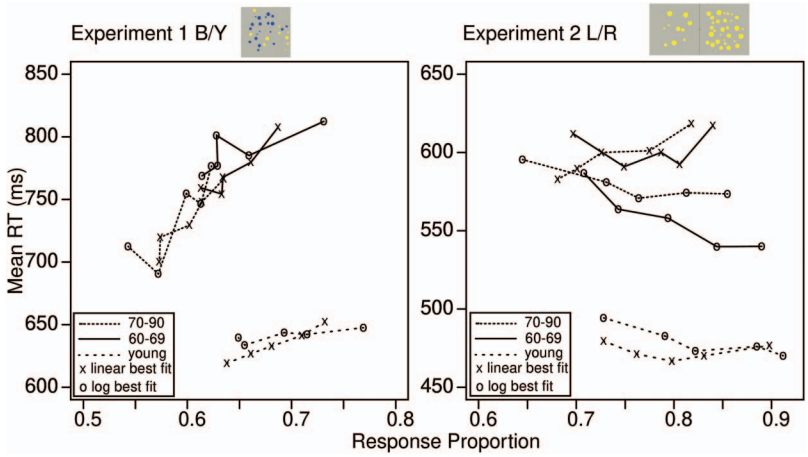

Figure C1. Plots of mean response time (RT) against accuracy for Experiments 1 (B/Y) and 2 (L/R) for groups of subjects in which the linear model fit better than the log model, and vice versa. The plots are only for differences in numerosity of five. See the online article for the color version of this figure.

in Ratcliff and McKoon. Results presented in Ratcliff et al. (2018) provide support for the linear model in tasks with two arrays of patches of bright or dark pixels of grayscale arrays. This suggests that because subjects cannot form separate representations of the two arrays for use in comparison (cf. absolute identification), they use differences between the two arrays. Thus, the argument is that for older adults, there is a greater tendency to rely on differences between the two arrays in the L/R task relative to young adults.

Appendix D

Variability in Recovered Parameter Values Versus Individual Differences

In other studies using the diffusion model, it has been found that for an experiment taking 30–45 min with 1,000–2,000 total observations, *SD*s in parameters from variability in data are typically 3–5 times smaller than individual differences (Ratcliff & Childers, 2015; Ratcliff, Huang-Pollock, & McKoon, 2018; Ratcliff & Tuerlinckx, 2002). This means that individual differences in parameters and differences among groups in parameters can be safely interpreted. This is because the added variability to individual differences is small; for example, if the *SD* in the estimation of a parameter was x and the *SD* across individuals was $3x$, then the combined *SD* would be $\sqrt{(x^2+9x^2)} = 3.16x$, which is a 5% increase in *SD*.

The integrated models are a little different from the standard diffusion model because in the standard model, there is usually one parameter representing across-trial variability in drift rate and separate drift rates for each condition of the experiment. In the integrated models, single coefficients produce drift rate and across-trial *SD*s in drift rates.

To examine the accuracy of parameter recovery both in size and bias, as well as correlations in parameter values, we performed a Monte Carlo parameter recovery study (e.g., Lerche, Voss, & Nagler, 2017; Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002). Mean parameter values for the young adults and 70–90-year-olds for Experiment 1 were used to generate 64 sets of simulated data, and the model was fit to each of these simulated data sets in the same way as the data were fit. To generate the simulated data, the number of observations per condition (for the 20 conditions of the experiment) was 73 for the data from parameters from young adults and 67 for the data from parameters from the 70–90-year-olds. The simulated data were generated using the random walk method (Tuerlinckx, Maris, Ratcliff, & De Boeck, 2001).

Table D1 shows parameters from fits to data (from Tables 3 and 4), the *SD*s in those parameters across subjects, the mean parameter values from the 64 fits to simulated data, and the *SD*s in those parameter values. There are two main results to examine. First, there are small biases in some of the parameters. Boundary separation (a), the range in the starting point (s_z), and the two drift-rate coefficients are estimated to be a little higher for the Monte Carlo fits than for the data used to generate the simulated data (similar biases were obtained in Ratcliff & Tuerlinckx, 2002). The *SD* coefficient (σ_1) was a little biased for the 70–90-year-old parameter set but not for the young adults. The other parameters showed almost no bias (apart from the small constant *SD* in drift rate across trials, η_0).

(*Appendices continue*)

Table D1
Integrated Diffusion Model Parameter Means and Standard Deviations for Data and for Monte Carlo Simulations for the Linear Model for the B/Y Task

Age group	Source and measure	a	T_{er}	σ_1	s_z	s_t	v_p	v_e	η_0	G^2
Young adults	Data mean	0.100	0.466	0.00540	0.046	0.260	0.0319	0.0169	0.023	261.4
	Data *SD*	0.015	0.074	0.00237	0.029	0.088	0.0102	0.0062	0.043	24.1
	MC mean	0.104	0.466	0.00534	0.061	0.260	0.0345	0.0184	0.053	103.0
	MC *SD*	0.004	0.007	0.00147	0.009	0.011	0.0038	0.0021	0.040	10.3
70–90-year-olds	Data mean	0.117	0.518	0.00715	0.052	0.277	0.0233	0.0070	0.022	263.6
	Data *SD*	0.024	0.118	0.00505	0.038	0.103	0.0162	0.0055	0.044	39.6
	MC mean	0.127	0.520	0.00792	0.072	0.279	0.0278	0.0086	0.071	108.0
	MC *SD*	0.007	0.008	0.00244	0.015	0.018	0.0056	0.0022	0.066	12.2

Note. *SD* = standard deviation; MC = Monte Carlo simulation. The parameters were boundary separation a, starting point $z = a/2$, and mean nondecision component of response time T_{er}. The constant coefficient of *SD* in drift rate across trials is η_0, and the coefficient that multiplies the square root of the sum of the squared numerosities is σ_1. Range of the distribution of starting point is s_z, and range of the distribution of nondecision times is s_t. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition. G^2 is the multinomial likelihood statistic.

The size of the *SD*s in parameter values from the Monte Carlo fits were smaller than those from individual differences in data in almost all cases by a factor of over 2.5 (with exceptions: the constant *SD* in drift rate across trials, η_0, and the *SD* coefficient σ_1). This means that differences among individuals dominate variability in the parameter values in these studies.

There are also strong relationships among model parameters in the Monte Carlo data sets, and these are shown in Figures D1A and D1B. (Note that there is a ceiling effect in the s_z parameter because its range is restricted in model fitting so it does not exceed the decision process boundaries). Ratcliff and Tuerlinckx (2002) showed similar patterns of correlations and provided an interpretation in terms of perturbations in data. Suppose one quantile RT for an error response was higher on average than its true value. Then, to compensate, boundary separation and the across-trial *SD* in drift rate would be a little higher to produce the higher RT, and drift rate would also be higher to increase accuracy, which would be needed in across-trial *SD* if drift rate increased. To fully understand and explain these effects, the effects of random changes in data and how the model compensates for these needs to be understood.

(*Appendices continue*)

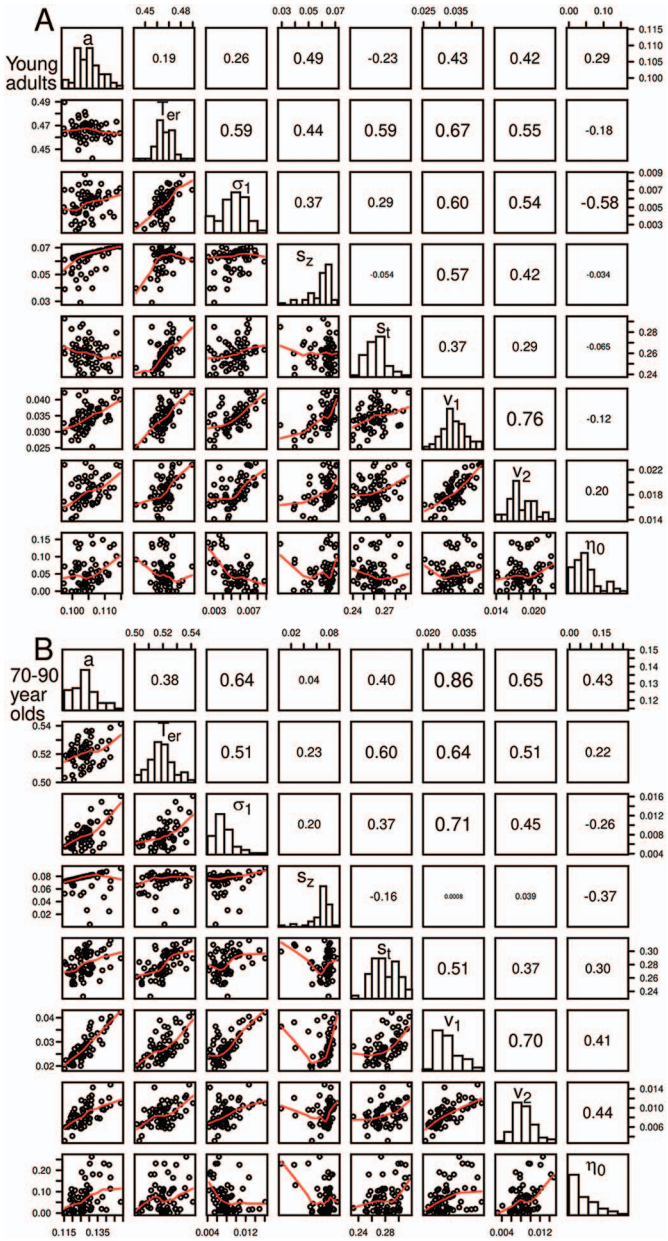

Figure D1. Scatter plots and correlations for ANS-diffusion model parameters for Monte Carlo simulations. In each panel, a single set of parameters (from Table 3, young adult and 70–90-year-old values) was used to generate simulated data, and the model was fit back to the data. The physical size of the correlations (the numbers) represents the size of the correlations. The parameters were boundary separation a, mean nondecision component of response time T_{er}, the constant coefficient of standard deviation in drift across trials is η_0, and the coefficient that multiplies the square root of the sum of the squared numerosities is σ_1. Range of the distribution of starting point is s_z, and range of the distribution of nondecision times is s_t. v_p is the drift-rate coefficient for the proportional-area condition, and v_e is the drift-rate coefficient for the equal-area condition. See the online article for the color version of this figure.

Received November 20, 2019
Revision received May 22, 2020
Accepted May 30, 2020 ■