Cognitive Psychology 138 (2022) 101516

Contents lists available at ScienceDirect

Cognitive Psychology

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

Integrated diffusion models for distance effects in number memory☆

Roger Ratcliff[1]

The Ohio State University, United States

ARTICLE INFO

Keywords:
Number memory
Diffusion decision model
Response time and accuracy
Distance effects
Overlap and gradient models

ABSTRACT

I evaluated three models for the representation of numbers in memory. These were integrated with the diffusion decision model to explain accuracy and response time (RT) data from a recognition memory experiment in which the stimuli were two-digit numbers. The integrated models accounted for distance/confusability effects: when a test number was numerically close to a studied number, accuracy was lower and RTs were longer than when a test number was numerically far from a studied number. For two of the models, the representations of numbers are distributed over number (with Gaussian or exponential distributions) and the overlap between the distributions of a studied number and a test number provides the evidence (drift rate) on which a decision is made. For the third, the exponential gradient model, drift rate is an exponential function of the numerical distance between studied and test numbers. The exponential gradient model fit the data slightly better than the two overlap models. Monte Carlo simulations showed that the variability in the important parameter estimates from fitting data collected over 30–40 min is smaller than the variability among individuals, allowing differences among individuals to be studied. A second experiment compared number memory and number discrimination tasks and results showed different distance effects. Number memory had an exponential-like distance-effect and number discrimination had a linear function which shows radically different representations drive the two tasks.

Memory for number is a crucial component of general mathematical ability. It is essential for learning how to perform computations with numbers and for performing computations in real-life situations. There has been much research on short-term (working) memory for numbers, often by showing subjects a short list of numbers (e.g., 1–7 numbers) followed immediately by a test number for which subjects are asked to decide whether or not it had appeared in the just-presented list (the Sternberg paradigm, Sternberg, 1966, 1969; Clifton & Gutschera, 1971; Corballis, 1967; De Rosa & Morin, 1970). In a study that called into question all of the simple scanning models that were developed for this task, Monsell (1978) showed that a test item contacted previous lists that occurred as much as 10 min before the current list (in an experiment using words as stimuli). Results showed slowing in a "new" test item could be detected if it had been presented as a "new" test item up to 10 min previously compared with a "new" test item that had not been presented earlier. This suggests that long-term memory is involved in even such a simple task. Despite the interest in this short-term memory task, long-term memory for numbers has been rarely studied in tasks analogous to simple list-learning experiments used to examine primacy and

☆ This work was supported by funding from the National Institute on Aging (Grant No R01-AG041176 and R01-AG057841). Data and code are available from the first author on reasonable request. This study was not preregistered.

E-mail address: ratcliff.22@osu.edu.

[1] The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, United States.

https://doi.org/10.1016/j.cogpsych.2022.101516
Received 19 May 2022; Received in revised form 26 August 2022; Accepted 30 August 2022
Available online 14 September 2022

recency in the relationship between primary and secondary memory.

Three general issues are addressed in this article. One is to examine the similarity structure of the representation of numbers in long-term memory. To address this, I used a recognition memory paradigm in two experiments. Subjects were shown study lists of six two-digit numbers (1500 ms per number), each study list followed by a list of 12 test numbers for which they were to decide if each test number had or had not been in the list they were just shown (the same decision as for the working-memory experiments mentioned above). The manipulation of most importance was the numerical distance of a test number from a number in the study list. For example, if 46 was a studied number, then the numerical distance from it to 47 and 45 would be one, from it to 48 and 44 would be two, and so on. The experiments produced two measures: the extent to which studied ("old") numbers could be distinguished from non-studied ("new") numbers and the extent to which this depended on the numerical distance between a new number and an old number; in other words, whether number representations encode similarity based on numerical distance.

The second issue was whether the two measures are precise enough to provide reliable differences among individuals. If so, the measures could provide a broad view of the range of differences among individuals and they would allow meaningful assessment of correlations between these measures and practical measures such as age and scores on standard mathematical ability tests.

The third issue was whether the representations of number obtained for number memory and those obtained for other number tasks exhibit the same properties. In earlier studies (Ratcliff, 2014; Ratcliff et al., 2015), number discrimination (e.g., is 78 greater or less than 50) showed quite different distance effects from number memory, meaning that it is essential to test hypotheses about representation with more than one task. To address this, the second experiment used the same number memory task as for the first experiment plus a number discrimination task, both tested on the same subjects.

The focus of the study by Ratcliff et al. (2015) was to provide an approach that would allow an understanding of relationships among numeracy tasks and the different dependent variables in the tasks. The tasks used were number memory, number discrimination, and nonsymbolic tasks (e.g., is the number of asterisks in an array greater or less than 50), and examined why there are sometimes correlations among their dependent variables and sometimes not and why there are sometimes correlations between the dependent variables and measures of achievement and sometimes not (e.g., Chen & Li, 2014; Inglis et al., 2011; Maloney et al., 2010; Price et al., 2012). The problem has arisen because of the sometimes arbitrary choices of dependent measures, with some labs using mean response times (RTs), others using accuracy, and others using the Weber fraction (an accuracy-based measure). Ratcliff et al. (2015) addressed these issues by using a model of decision processes (the diffusion model, Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff et al., 2016, described below) that gave a detailed analysis that showed how the different measures are related to each other and how their relations and correlations can be different for different tasks. Ratcliff et al. argued that it is essential to explain how accuracy and RTs jointly affect performance and that a model-based analysis is required to do this.

Other than the experiments in Ratcliff et al., there have been very few studies that have examined the representation of numbers in memory from a list of numbers that had to be remembered for a short time (with longer delays between study and test than working memory span). Similarity (numerical distance) is one aspect of numbers that might be encoded and I have found only one study of numerical distance in long-term recognition memory for numbers. It was an early paper-and-pencil study by Dale and Baddeley

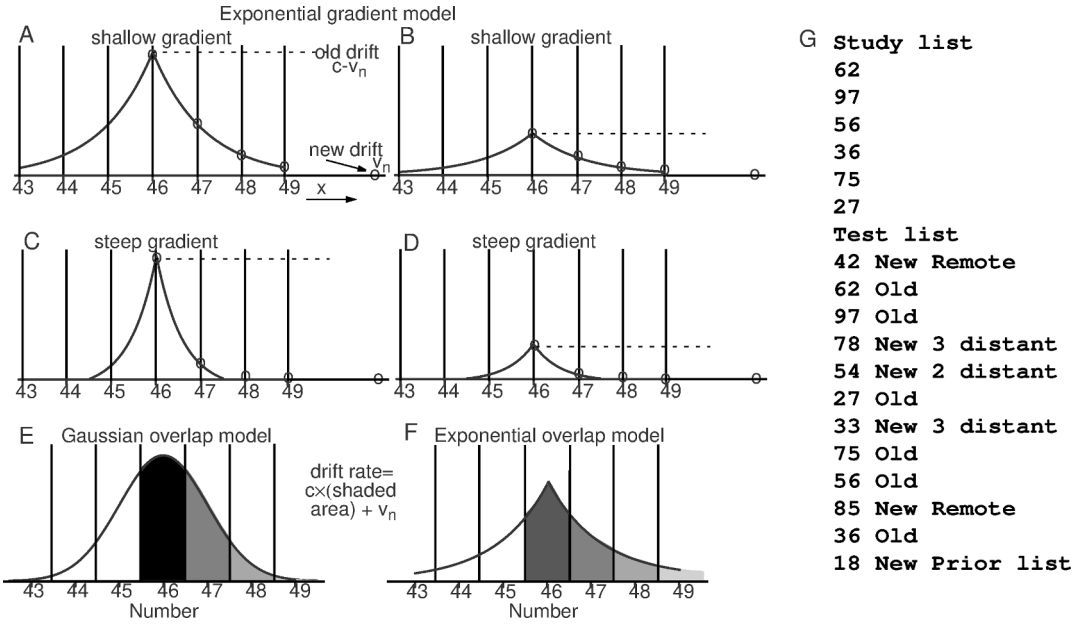

Fig. 1. The three representation models. Fig. 1A-1D show the drift rates for the exponential gradient model (the circles on the functions) when the number 46 is studied. Fig. 1A and 1C show the case when discriminability between "old" and "new" numbers is high and Fig. 1B and 1D show the case when it is low. Fig. 1A and 1B show a shallow gradient (with low discrimination/high confusability) and Fig. 1C and 1D show a steep gradient. Fig. 1E and 1F show the Gaussian and exponential overlap models. Fig. 1G shows a sample study and test list from the experiment.

(1966), who used a paradigm in which subjects were shown a list of 15 two-digit numbers and then asked to recall them. They found systematic intrusions; for example, if 28, 58, 68, 95, and 97 were in the list, 98 might be recalled in error. Then they placed the numbers with high intrusion rates into a recognition memory test and found elevated false alarm rates for those numbers relative to numbers with low intrusion rates.

I tested three models for the representation of numbers in the number memory task used here. The first assumes that new test numbers match old numbers as a function of numerical distance, with the value of the match decreasing with numerical distance as an exponential gradient. The exponential is the simplest function to choose and it has been used extensively in categorization research (Nosofsky 1986; 1987; Nosofsky & Palmeri, 1997) with results that show similarity is related to psychological distance according to an exponential function (Shepard, 1986, 1987).

The other two models assume that the representations of numbers are distributed over numerical distance from an old number such that the representations of new numbers overlap with the representations of old numbers as a function of distance (Ratcliff 1981, 1987; Gomez, Ratcliff, & Perea, 2008; see also models for transposition errors in short-term recall, e.g., Lee & Estes, 1977). For one of these models, the distribution of an old number is assumed to be Gaussian and for the other, it is assumed to be back-to-back exponentials. The three models are illustrated in Fig. 1 and will be described in more detail later.

Throughout this article, I emphasize that the information contained in memory representations cannot be determined directly; rather its effects can be observed only when accuracy and RT distributions are used jointly to model decision processes. I integrated the diffusion model with the representation models and the latter provided the degree to which a new number is similar to an old number which is the information given to the decision process. The decision model then translates that information into RT distributions and accuracy.

1. The diffusion decision model

The diffusion model provides explanations of behavior in two-choice tasks. There are a number of comprehensive reviews of it and how it has been applied in a number of domains (e.g., Forstmann et al., 2016; Ratcliff & McKoon, 2008; Ratcliff et al., 2016), including clinical and neuroscience applications.

The model assumes that evidence from a stimulus or memory is noisy from moment to moment and that this noisy evidence is accumulated over time from a starting point (z) until one of two decision boundaries (a and 0) is reached, at which point a response is initiated. The model accounts for the effects of experimental manipulations on all aspects of two-choice data: accuracy, mean RTs for correct responses and error responses, the full distributions of RTs for correct and error responses, and the relative speeds of correct and error responses. One of the strongest constraints on modeling is the shape of RT distributions (Ratcliff, 2002). For simple two-choice decisions, empirical RT distributions for humans are almost always positively skewed and have roughly the same shape even though the location and spread may differ across conditions that vary in difficulty. It is the variability in the accumulation process that gives rise to the right-skewed RT distributions and variability also gives rise to errors when the accumulation process reaches the wrong boundary.

The model provides a decomposition of data that isolates components of processing. One important component is the settings of the decision boundaries that represent the amount of evidence that must be accumulated for a decision to be made (boundary separation, a). Another is the evidence from a stimulus or memory that drives the accumulation process (drift rate, v). The third is the time taken by processes outside the decision process itself - the time to encode a stimulus and extract decision-relevant information from it and the time to make a response. These last three durations are combined into one parameter of the model, nondecision time (T_{er}).

The model also assumes that these components cannot be set to exactly the same values on successive trials (starting point and nondecision time) or with nominally equivalent stimuli (drift rates). Therefore, the model assumes across-trial variability in drift rate (normally distributed with standard deviation η), in starting point (uniformly distributed with range s_z), and in nondecision time (uniformly distributed with range s_t). Predictions of the model are robust to the precise form of these distributions (Ratcliff 2013). Also, across-trial variability in drift rate and starting point allow the model to account for the relative speeds of correct and error responses (Ratcliff & McKoon, 2008, Fig. 4).

2. Integrated diffusion models

In most past research with the diffusion model, drift rates have been estimated for each condition in an experiment separately, but in recent research, there has been development of models that integrate models of perception or memory with the diffusion model to provide a more complete model of representation and processing. In integrated models, the perceptual or memory models provide drift rates and the diffusion model provides the decision process, which, in turn, provides RT distributions and accuracy. Sometimes models provide values for each condition of an experiment and sometimes they provide values for each individual trial. In the former, the model can be fit with methods that group responses such as by using quantile RTs. In the latter, the choice and RT for each trial can be fit with maximum likelihood (see Ratcliff & Tuerlinckx, 2002).

There are a number of examples of integrated diffusion models in the literature. Ratcliff (1981) proposed an integrated model for a perceptual matching task in which a letter string was studied which was followed immediately by a test string; a subject was to decide whether the two strings were the same or different. The model assumes distributions of letters over position and the overall degree of overlap between the two strings provides drift rates to the diffusion process. If a test string contains only new letters, there is little overlap and this produces a drift rate with a large negative value, leading to fast and accurate “different” decisions. If a test string contains letters transposed from the study string, overlap is larger and drift rate has a smaller negative value, which makes “different”

decisions slow and inaccurate. Because of limitations in computer speed, the representation and diffusion models were not fit jointly, rather drift rates for separate conditions were estimated and the representation model fit to those drift rates. Gomez et al. (2008) applied the model successfully to experiments with combinations of transpositions, replacements, and repetitions of letters but did not model RTs. Similar distributed representations are also assumed in models for accuracy and RT distributions for confidence judgments in recognition memory (Ratcliff & Starns, 2009, 2013).

Smith and Ratcliff (2009) developed integrated models for simple two-choice perceptual tasks such as deciding whether the orientation of Gabor patches is horizontal or vertical. The front-ends of the models encode a representation of visual working memory that produces a drift rate as a function of contrast, stimulus duration, with or without masking, with or without a strong attentional cue to stimulus onset, and whether the stimulus is in an attended or unattended location. Smith and Ratcliff considered four models that crossed single- versus dual-racing diffusion processes with the effect of attention on the growth of the visual trace formation, represented by either a change in drift rate or a delay in the onset of growth. The four models were fit to data for which contrast, masking, stimulus duration, and an attentional cue were manipulated in an experiment by Gould et al. (2007) and one by Smith et al. (2004). The models produced similar accounts of the data and so provided an account of attention processes and their effects on speed and accuracy in these perceptual tasks.

For a recognition memory task, Ratcliff et al. (2016) assumed that drift rate was a linear function of an EEG measure. On each trial of the experiment, the drift rate was computed from the linear function and the likelihood of that stimulus and response (for maximum likelihood fitting). They found that a coefficient (the slope of the linear function) that mapped from the EEG measure to drift rate was reliably different from zero, indicating that the EEG signal provided a measure of the strength of items in memory on an item-by-item basis.

Sewell et al. (2019) fit an integrated learning/diffusion model to accuracy and RT distributions with a simple probabilistic category learning task. In the task, one stimulus is presented and the probability of assigning a stimulus to one or the other category is manipulated for different stimuli. For this task, the model accounted for changes in performance with only drift rate changing.

Pedersen et al. (2016) and Pedersen and Frank (2020) integrated a reinforcement learning model with a diffusion model. In a typical task, two alternatives were presented and the subject had to choose one of them. Feedback was provided that signaled which of the choices should have been made. The expected reward was updated using the delta rule. Drift rate on a trial was a drift rate coefficient multiplied by the difference in evidence for the two choices. Because this model produced a value of drift rate for each trial, it was fit using a Bayesian method (using a likelihood for each trial) and provided a reasonable account of the data. Pedersen and Frank (2020) developed a module for the Bayesian diffusion model fitting package (HDDM, a hierarchical drift–diffusion model, Wiecki et al., 2013) that allows fitting of a hierarchical version of the reinforcement learning model. Other integrated models can be also be fit with the HDDM model fitting package with simple models of drift rates, boundary settings, and nondecision time (e.g., linear models) but the hierarchical part of the package does not allow differences in the across-trial variability parameters across individual subjects. There are two other recent integrated reinforcement learning/diffusion models, one by Fontanesi et al. (2019) and one by Miletic et al. (2021). The former ignores the behavior of RT distributions but the latter provides a detailed analysis of distributions. In contrast to the Pedersen et al. model, neither model has boundaries changing with learning. But critically, neither model would account for large shifts in RT distributions that occur when unrewarded (low reinforced) alternatives are paired at test producing high conflict (Ratcliff & Frank, 2012). This is a fruitful area in which integrated models can be developed and tested. The challenge is to work with clinically-relevant data and to apply the models to more complicated reinforcement learning paradigms.

3. Numerical cognition and diffusion models

In the numerical cognition literature, numerosity discrimination tasks have been used to examine representations of nonsymbolic numerosity. Ratcliff and McKoon (2018; 2020a) integrated linear and logarithmic models of representations of numerosity with the diffusion model (linear and log models have also been applied to brightness and motion discrimination tasks, Ratcliff, Voskuilen, & McKoon, 2018). Ratcliff and McKoon's study (2018; 2020a) used one task for which blue and yellow dots were mixed in a display and subjects decided which color was more numerous. In another task, stimuli were two side-by-side arrays and subjects decided which had the larger number of dots. In the linear model, drift rate was a function of the difference in the number of dots and in the log model, drift rate was a function of the difference in the logs of the number of dots. In the linear model, across-trial SD in drift rate was a linear function of the sum of the squares of the two numbers and in the log model it was constant.

The mixed-display task produced a surprising result: for a small constant difference in the number of dots, as the total number increased (e.g., 10 and 15 to 20 and 25 to 35 and 40), accuracy decreased but counter-intuitively, RT also decreased. For the separate-display task, there was the usual finding: accuracy decreased and RT increased. As was suggested above, different models of representation are sometimes needed for different tasks: the linear model accounted for the mixed-display result and the log model accounted for the separate-display result. A speculation as to why this occurred is that when there is a mixed display, the only information that is available is differences and so the linear model with increasing variability accounted for the data. When the two displays are side by side, separate representations are available and the so the log model accounted for the data. Kang and Ratcliff (2020) extended these models to examine non-numeric (e.g., dot area) variables as well as numeracy ones. These results reinforce the argument that representations used in decision-making in some cases are task dependent.

Two tasks with symbolic (two-digit numbers) have been examined using diffusion models. In a number discrimination task from Ratcliff et al. (2015), subjects were to decide whether a two-digit number was greater or less than 50, one of the tasks used in Experiment 2 below. Drift rate functions were approximately linear with drift rates near 50 lower than those further away numerically (see also Experiments 3 and 4 in Ratcliff, 2014). In a number-line task, Ratcliff and McKoon (2020b) applied a spatially continuous

diffusion model to results from a task in which responses were made on a continuous line. In this task, a two-digit number was presented and subjects had to point to the position on a line (0–100) that reflected that number. Results showed that the analog representation of the symbolic numbers was symmetrically distributed with standard deviations that were quite similar to each other across the range of stimulus numbers.

4. Representation models for numbers in recognition memory

The three models of representation that I evaluated are shown in Fig. 1. One is the exponential gradient model (Fig. 1A-1D). Drift rates are an exponential function of distance from an old number:

$$v(x_i) = v_n + c \times \exp(-(x_i - x_0)/\tau), \tag{1}$$

where v_n is the drift rate for a test number that is far distant from all old numbers, c is the difference in drift rates between an old and a far-distant new number (so the drift rate for an old number, 46 in the figures, is $v_n + c$) and τ is the decay constant. In Figures A-D, the drift rate falls from that for 46 to one-distant numbers, two-distant numbers, and then 3-distant numbers. The drop off is more shallow in A and B than in C and D (i.e., τ is larger in A and B compared with C and D) and the discriminability between old and new numbers is higher in A and C than in B and D. This illustrates the range of values the function can take.

Fig. 1E and 1F show the overlap models, with only one example for each instead of the four for the exponential gradient model. The representation of a number is distributed over position as a Gaussian distribution, 1E, and back-to-back exponentials, 1F. (Only the right-hand portions of the distributions are shown; the left-hand ones would be mirror images of the right-hand ones). Drift rates are proportional to the areas under the functions, the black area between 45.5 and 46.5 for old numbers, the dark-gray areas between 46.5 and 47.5 for 1-distant new numbers, the mid-gray areas between 47.5 and 48.5 for 2-distant new numbers, and so on. The expression for drift rate is given by.

$$v(x_i) = v_n + c \int_{x_i-0.5}^{x_i+0.5} f_{x_0}(x)dx \tag{2}$$

The parameter of the overlap models that corresponds to the decay parameter in the exponential gradient model is the SD of the distributions.

The parameter for the SD in the exponential overlap model is the decay constant τ. The integration in Equation (2) gives an exponential function that is the same as that for the exponential gradient model so the two models appear to be the same, but these differ in the center of the function. In the gradient model, the peak is exponential at 46 in the example, but in the overlap model the area is reduced from a pure exponential because it has areas from the two back-to-back exponentials. Thus it produces a lower drift rate for old numbers (the black area in Fig. 1F) than the exponential gradient model relative to the drift rates for the other distances.

In applications of the model to memory tasks, the across-trial SD in drift rate has been found to be different for old and new items (Ratcliff et al., 1992; Starns & Ratcliff, 2014; Starns et al., 2012). Therefore, for all three models, variability in drift rate across trials was assumed to be normally distributed with different SDs for old numbers (η_o) and new numbers (η_n).

5. Experiment 1

This experiment used the number memory task described above; subjects studied a list of 6 numbers followed by a test list of 6 old and 6 new numbers. The data of interest were RTs and accuracy for old test numbers, new test numbers that were numerically far from all studied numbers (I label these remote numbers), new test numbers that were 1-, 2-, or 3 -distant from an old number, and a number that was new in the immediately prior test list. The three integrated models were tested with the data from this task. Another model was added for which the diffusion model was applied in the usual way, that is, without any representation model to determine drift rates and so the drift rates for each condition were different. I call this the default model.

Designing this experiment was tricky because there are a large number of possible confounds such as repetitions of numbers, numbers that might be particularly memorable, test numbers that cross decades, and so on. As described below, the design of the experiment controlled for as many of these factors as possible.

5.1. Method

Thirty-eight college students, all at Ohio State University and all of whom signed a consent form approved by the IRB, participated in the experiment for one 55-minute session for credit in an introductory psychology class. All the data were collected before they were examined and subject and outlier elimination for all experiments was carried out before any data analysis and modeling was performed. Eight subjects produced large proportions of fast guesses; for these subjects, there were an average of 19.3 % responses with RTs less than 300 ms with accuracy at chance. The data from these subjects were eliminated from the analyses. This was done because these subjects were not following instructions and so other aspects of their data may be suspect.

Study and test numbers were displayed on a PC monitor using local real-time software and responses were collected from the PC keyboard. Subjects were asked to respond to the test numbers as quickly and accurately as they could. They were tested either alone or in pairs.

There were 80 study-test lists. The first was used for practice and the data were discarded. There were 64 lists that manipulated

distance and 16 filler lists. Each list began with an instruction to press the space bar to begin the study list. Each number was displayed for 1300 ms, then the screen cleared for 200 ms, and then the next number was displayed. After the last number, subjects pressed the space bar to begin the test list. The "/" key was used for "old" responses and the "z" key for "new" ones. If a response was shorter than 280 ms, a message "TOO FAST" appeared for 1500 ms. If a response was longer than 1250 ms, "TOO SLOW" appeared for 300 ms. Correct or error feedback (the words "CORRECT" or "ERROR") was given for 300 ms after the TOO SLOW or TOO FAST message (if there was one) or immediately after the response. After this, there was a 300 ms blank screen and then the next test number.

In the 64 lists that manipulated distance, 3 of the study numbers in each list were those to be tested, each with a number that was 1-distant, 2-distant, or 3-distant (randomly chosen). These 3 numbers were studied only in positions 2 through 5 in the study list to avoid beginning or end of list effects. These 3 study numbers were selected randomly from the ranges 14–16, 24–26, …, 94–96, with the restriction that they were at least two decades away from each other. The other 3 study numbers were selected randomly from numbers that were not in the same decade as any of the other five study numbers, had not appeared in the preceding two study-test lists (with the exception of the number that appeared in the prior list), and were not decade numbers (10, 20, ……90).

The 16 filler study-test lists were used for numbers that were under-represented in the 64 experimental lists. The filler lists were separated from each other by between 3 and 7 of the experimental lists. Numbers in the range 11–13 occurred with (relative to the other numbers) frequency 5, numbers in the range 14–16 occurred with frequency 1, numbers in the ranges 17–19 and 21–23 occurred with frequency 5, …, numbers in the range 94–96 occurred with frequency 1, and numbers in the range 97–99 occurred with frequency 5.

In a test list, the first tested number could not be the last number from the study list. The 1-, 2-, and 3-distant numbers appeared in positions 3–9 of the test list, at least three test numbers before their studied number (the study numbers for these test numbers appeared in positions 6–12 in the test list but the other "old" test numbers could appear in any position in the test list). The test number from the previous list appeared in a random position. An example study and test list is shown in Fig. 1G. In the test list, only the numbers were presented; to the right of the number is a description of the condition (which was not shown to subjects).

6. Results

The first study-test block and the first response in the test list in each block were eliminated from the data. Of the remaining data, responses with RTs less than 350 ms and greater than 3000 ms, about 2.4 % of the data, were also eliminated. For responses with RTs between 300 and 350 ms, the accuracy of "old" items was 0.638 and new items not from the prior list was 0.603. For responses with RTs between 350 and 400 ms, the accuracy of "old" items was 0.642 and new items not from the prior list was 0.544. Thus, accuracy started to rise at around the 350 ms cutoff (Because the model was fit using quantile RTs, elimination of a few short RTs hardly affects the model fits, see Ratcliff & Tuerlinckx, 2002).

Table 1 gives the means for accuracy and RTs for "old" numbers, 1-, 2-, and 3-distant numbers, remote numbers, and numbers from the prior list. As would be expected, the probability of a "new" response increased, by about 11 %, as the distance between a test number and its study number increased from 1-distant to remote. Mean correct RTs changed by 28 ms from 1-distant to remote test numbers.

I performed one-way ANOVAs on accuracy values and correct mean RTs for 1-, 2-, and 3-distant numbers, and remote numbers. For accuracy, the effect of distance was significant, $F(3,87) = 17.8$, $p = 4.5 \times 10^{-9}$, $\eta_p^2 = 0.38$ (partial η_p^2) and for mean RTs, the effect was also significant, $F(3,87) = 4.5$, $p = .0055$, $\eta_p^2 = 0.13$.

6.1. Fits of the models

I fit the three integrated models and the default model to the data from each subject separately using the G-square method described in the Appendix. The means over subjects of the parameters that best fit the data and the mean over G-square values are shown in Table 2. The eighth and ninth columns show the values for the parameters of the representation models; the decay constant for the exponential model (τ) or the SD (σ or τ) of the overlap models and the drift-rate multiplying constant (c). The drift rates for remote (new) numbers and new numbers from the previous list are also shown in Table 2.

Table 1
Accuracy, mean RTs, and numbers of observations.

Condition	Experiment 1				Experiment 2			
	P ("old")	Mean RT "old" (ms)	Mean RT "new" (ms)	Mean N per subject	P ("old")	Mean RT "old" (ms)	Mean RT "new" (ms)	Mean N per subject
Old number	0.667	642	675	420	0.716	657	715	345
New 1-distant	0.439	652	707	63	0.429	683	743	51
New 2-distant	0.403	659	688	62	0.397	674	739	52
New 3-distant	0.351	675	697	62	0.350	697	734	52
New remote	0.328	654	679	185	0.320	678	708	154
New prior	0.408	656	678	62	0.383	702	721	52

Table 2
Mean model parameters from fits to data.

Expt.	Model/task	a	T_{er}	η_o	η_n	s_z	s_t	z	σ/τ	c	v_n	v_p	G^2
1	Default	0.098	0.481	0.198	0.117	0.040	0.247	0.054			−0.125	−0.080	66.2
	Exp. grad.	0.098	0.476	0.194	0.115	0.031	0.244	0.054	0.977	0.220	−0.123	−0.080	67.8
	Gaus. overlap	0.099	0.477	0.199	0.122	0.034	0.245	0.054	0.632	0.407	−0.118	−0.081	69.0
	Exp. overlap	0.098	0.478	0.194	0.119	0.035	0.245	0.054	0.800	0.251	−0.123	−0.079	68.4
2	Default	0.108	0.487	0.204	0.145	0.036	0.270	0.062			−0.151	−0.100	64.8
	Exp. grad.	0.108	0.486	0.216	0.151	0.030	0.272	0.063	0.824	0.276	−0.141	−0.094	67.6
	Gaus. overlap	0.108	0.487	0.218	0.158	0.029	0.273	0.063	0.763	0.621	−0.145	−0.101	68.0
	Exp. overlap	0.108	0.484	0.200	0.142	0.024	0.271	0.062	0.860	0.311	−0.140	−0.091	67.2
	Number disc.	0.124	0.338	0.088		0.034	0.115	0.062					68.0

The model parameters are: boundary separation a, starting point z, nondecision time T_{er}, SD in drift rate across trials for studied items η_o and for "new" test items η_n, across trial range in starting point s_z, across trial range in nondecision time s_t, decay constant τ for the exponential gradient model and SDs σ for the overlap models, drift rate multiplying constant c, drift rate for remote new numbers v_n, and drift rate for numbers from the prior list v_p.

The values of the diffusion model parameters were quite close to each other for the three integrated models and the default model and similar to those from recognition memory experiments with words (Ratcliff et al., 2004, 2010, 2011) and the number memory experiment in Ratcliff et al. (2015). Table 3 shows data in the first row and the next rows show drift rates and the probabilities of responses that the drift rates predict: the probabilities of "old" responses, of 1-distant, 2-distant, and 3-distant "new" responses, and of remote "new" responses. The predicted probabilities for the default model are all within about 0.01 of the data, as might be expected.

Of the three integrated models, the predicted probabilities of the exponential gradient model matched the data best; the largest miss was 0.03 for the 2-distant condition. The Gaussian overlap model's predictions matched the data less well; it over-estimated the probabilities for old and 1-distant numbers (both about 0.03 higher than the data) and under-estimated them for 2-distant numbers (about 0.07 lower than the data). In other words, the model predicted a steeper gradient between 1- and 2-distant numbers than the data. The exponential overlap model over-estimated the probabilities for old and 1-distant numbers (about 0.03 and 0.05 higher than the data, respectively) and under-estimated them for 2-distant numbers (about 0.03 lower than the data). For new numbers from the previous test list, accuracy, mean RT, and drift rates were similar to those of 2-distant numbers (Tables 1 and 2), showing some leakage from the previous list as for similar studies with word stimuli (Ratcliff, 1978; Ratcliff, Clark, & Shiffrin, 1990, Experiment 4).

Fig. 2 shows quantile-probability plots in which the 0.1, 0.3, 0.5, 0.7, and 0.9 quantile RTs are plotted vertically against the proportions of responses for each experimental condition for the exponential gradient model (the plots for the other models are quite similar and are not shown). The data and the model predictions are generated for each subject separately then averaged in the same way. Fig. 2A shows "old" responses and 2B shows "new" responses. The x's are the data and the o's and the lines between them are model predictions. Quantile-probability plots show how accuracy and the shapes of the RT distributions jointly change across the conditions of an experiment that differ in difficulty. The shapes of the distributions can be seen (approximately) by drawing equal-area rectangles between the quantile RTs, as shown in Fig. 2B. The 0.1 quantile represents the leading edges of the distributions and the 0.9 quantile the tail of the distributions. The only systematic misses for the model are small and occur in the leading edges of the RT distributions (the 0.9 quantiles have high variability and so misses are not as systematic).

Fig. 2C, again for the exponential gradient model, shows the probabilities of "old" responses for all subjects and all conditions, with the data and predictions from the model plotted against each other. The x's are the points with more than 80 observations (and so less variability) and the o's for those with fewer observations. Fig. 2D, 2E, and 2F show the same plots for the 0.1, 0.5, and 0.9 quantile RTs. All of the plots show good correspondence between theory and data, with very few large deviations for points with more than 80 observations. The plots for the other models are quite similar to those for the exponential gradient model and are not presented here.

Fig. 3A shows a plot of the drift rates produced from the exponential gradient model (the "o" symbols) and the drift rates produced from the default model (the "x" symbols), for old numbers, new numbers 1-, 2-, and 3-distant, and remote numbers (placed at distance 6 because the function had asymptoted by distance 6). The plot shows a close match between the drift rates. Values for drift rates for all

Table 3
Mean drift rates and proportions of "old" responses.

Expt.	Data/model	v_o	v_1	v_2	v_3	v_n	pr_o	pr_1	pr_2	pr_3	pr_n
1	Data						0.667	0.439	0.403	0.351	0.328
	Default	0.101	−0.058	−0.078	−0.115	−0.125	0.678	0.438	0.402	0.339	0.323
	Exp. gradient	0.097	−0.044	−0.095	−0.113	−0.123	0.676	0.461	0.369	0.338	0.321
	Gauss. overlap	0.114	−0.034	−0.114	−0.118	−0.118	0.693	0.475	0.334	0.328	0.328
	Exp. overlap	0.110	−0.027	−0.096	−0.115	−0.123	0.693	0.494	0.371	0.338	0.325
2	Data						0.716	0.429	0.397	0.350	0.320
	Default	0.131	−0.072	−0.086	−0.113	−0.151	0.732	0.429	0.404	0.358	0.299
	Exp. gradient	0.135	−0.059	−0.117	−0.134	−0.141	0.736	0.459	0.360	0.333	0.322
	Gauss. overlap	0.158	−0.001	−0.130	−0.145	−0.145	0.761	0.560	0.342	0.319	0.319
	Exp. overlap	0.134	−0.020	−0.103	−0.128	−0.140	0.740	0.519	0.371	0.328	0.309

Drift rates v are: v_o for old numbers, v_n for remote new numbers, and v_1 - v_3 for numbers 1 – 3 digits separated from a studied number. The pr's are probabilities corresponding to those drift rates.

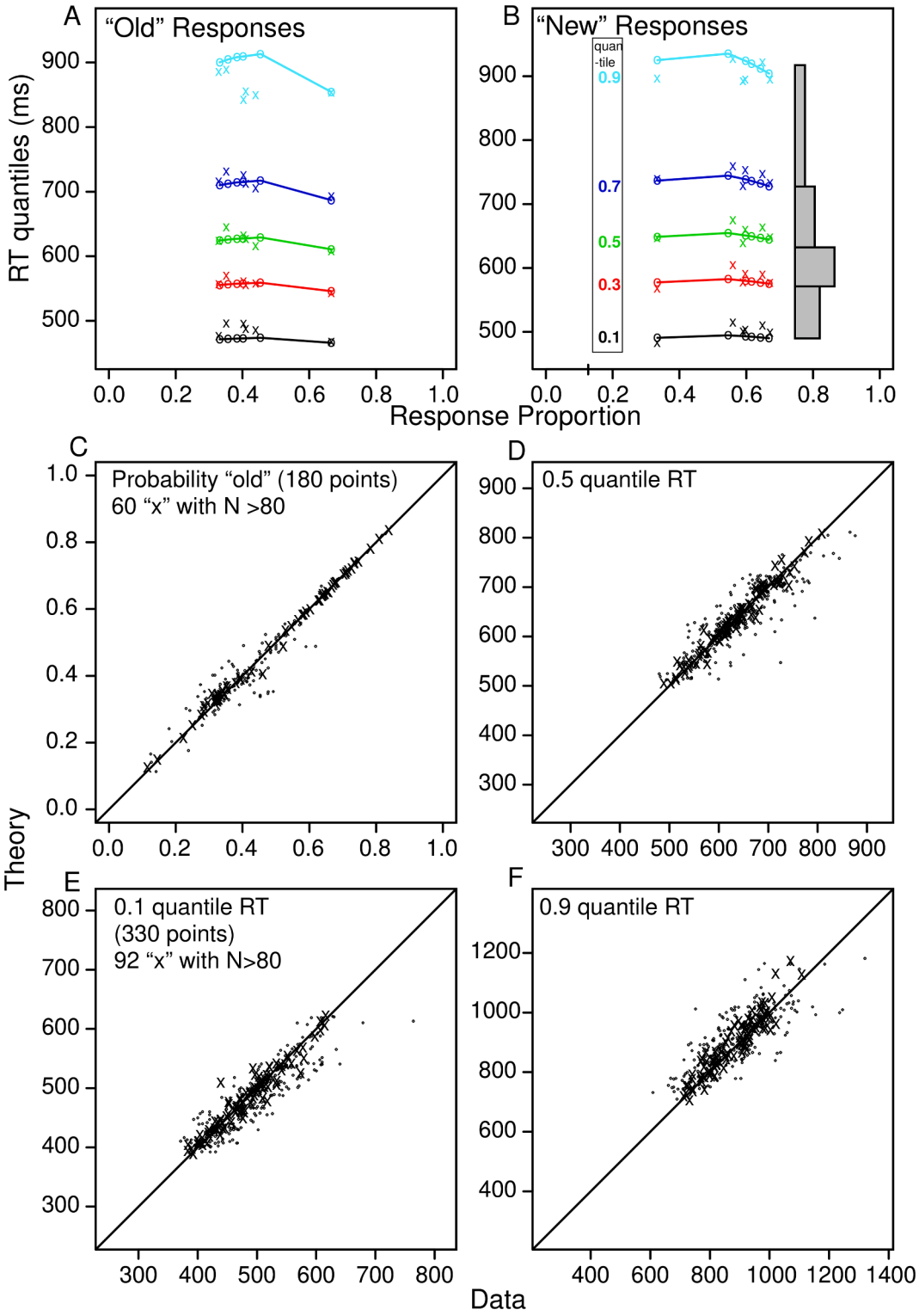

Fig. 2. The top two panels show quantile probability plots for "old" and "new" responses, with x's for the data and o's for the predictions of the exponential gradient model. The lines represent the 0.1, 0.3, 0.5, 0.7, and 0.9 quantile RTs and the horizontal location of the quantiles represents the proportion of those choices. The bottom four panels show the choice proportions and 0.1, 0.5, and 0.9 quantile RTs for theory (exponential gradient model) against data for each condition for each subject of the experiment. The "x" symbols are for conditions with more than 80 observations and the dots for conditions with between 10 and 80 observations. 30 data points from conditions and subjects for which there were too few observations to compute quantiles.

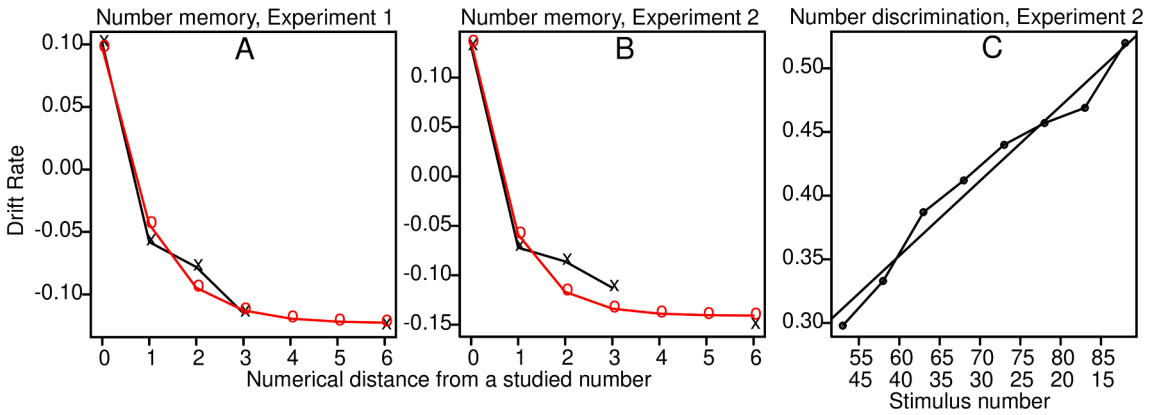

Fig. 3. Fig. 3A and 3B (for Experiments 1 and 2 respectively) show drift rates plotted against numerical distance. The x's are estimated drift rates from the default model with separate drift rates for each condition and o's are those from the exponential gradient model. Distant 0 is for studied numbers, distances 1, 2, and 3 are for test numbers 1, 2, and 3 from a studied number (e.g., when 46 was studied the test numbers might be 47, 48, and 49 respectively), and 6 is for remote numbers. Fig. 3C shows a plot of drift rates as a function of the numerical distance from the referent (50) for the discrimination task in Experiment 2. The straight line is a linear regression line.

the models and their predicted accuracy values are shown in Table 3.

6.2. *Goodness of fit*

For each experimental condition, the RT distributions are represented by 5 quantile RTs, (the 0.1 quantile represents the leading edge, the 0.5 quantile is the median, and the 0.9 quantile represents the tail). This produces 6 bins between and outside the quantiles which produces 11 degrees of freedom for each pair of correct and error RTs (the probabilities have to add to 1 which makes the 12 bins produce 11 degrees of freedom). Accuracy is represented in the RT bins: For accuracy 90 %, 0.9 probability will be spread in the bins for correct responses and 0.1 in the bins for error responses. Thus, for the 6 conditions of Experiment 1, there are 66 independent bins (i.e., 66 degrees of freedom) in the RT distributions, all to be fit with an 11 parameter model. This produces 55 degrees of freedom in the fits of the model to data.

The mean (over subjects) G-square statistics for the three integrated models and the default model are shown in Table 2. Numerically, in terms of G-squared values, the exponential gradient model fit best followed by the exponential overlap model, and then the Gaussian overlap model, but the differences are very small and the fits are quite similar. Numerical goodness of fit for the default model was slightly better than the three integrated models as would be expected because the drift rates are independent.

The goodness of fit statistic G-square is asymptotically distributed as chi-square and for this experiment, with 55 degrees of freedom for the integrated models, the critical value is 73.3 (for the default model with 53 degrees of freedom, the critical value is 71.0) For all of the models, the mean G-square values were less than the critical chi-square value. For the exponential gradient, Gaussian overlap, and exponential overlap models, 8, 9, and 8 subjects, respectively (out of 37) had significant values and for the default model, 6 subjects had significant values larger than the critical value. These show quite good fits of the models to the data like those from fits to other tasks (e.g., Ratcliff et al., 2010).

Because G-square is a multinomial log likelihood, AIC and BIC can be computed and the number of subjects favoring each model can be compared (see the Appendix). The results are presented in Table 4. For the G-square statistic, the default model was preferred by the most subjects, but this is expected based on the relative freedom of the model and the larger number of parameters. For AIC and BIC, around half the subjects preferred the exponential gradient model, next was the Gaussian overlap model, and then the exponential overlap model. This shows that the exponential gradient model fits best for around half the subjects, but because the fits are similar, the small differences between it and the other models do not allow us to decisively discriminate among the models; all do a reasonable job of fitting the data.

It is important to stress the constraints on the models in fitting data. First, as said above, they must explain why RT distributions

Table 4
Number of subjects best fit by the model for three goodness of fit statistics.

Experiment	Statistic	Default separate drift rates	Exponential gradient model	Gaussian overlap model	Exponential overlap model
1	G^2	22	4	1	3
	AIC	2	14	9	5
	BIC	0	16	9	5
2	G^2	26	2	1	0
	AIC	3	13	7	6
	BIC	0	16	7	6

have the shape that they do and why this shape is the same for all the conditions of the experiment, that is, why mean RTs change across conditions but the shapes of the RT distributions do not (see Ratcliff & McKoon, 2008). Second, there are 11 parameters for each of the integrated models and these 11 must jointly account for how accuracy and RT distributions change together across the conditions of an experiment.

The third constraint for all diffusion model applications is that changes in accuracy and RT distributions have to be consistent. Ratcliff (2002) made up several sets of fake but plausible data with RT distribution shape changing, whether the distributions shifted or spread with changes in accuracy, and whether correct and error RTs were shifted relative to each other. For most of the patterns of results, the diffusion model could not fit the fabricated data.

6.3. Individual differences

At this point, I have shown that, for all four models, the values of the parameters that gave the best fit to the data produced drift rates that produced accuracy and RT values that match the data well (although a little better for the exponential gradient model). While these results can tell us how the range of the distance effect varies across individuals, they do not tell us whether drift rates (or other parameters), can be useful in examining individual differences. The following analysis of the variability across subjects relative to the variability in parameter values shows that the model parameters are estimated with enough precision for individual difference applications. For example, a diffusion model analysis of a item recognition task with words and simple machine learning discrimination methods has given a separation of memory-disordered patients from controls with about 83 % accuracy (Ratcliff et al., 2021).

To use the model to examine individual differences, the variability among subjects in the parameter values needs to be larger than random variability from fitting the model to data (as a function of the number of observations). To estimate random variability, I used Monte Carlo simulations. For each model, I generated 64 sets of simulated data using the mean parameter values in Tables 2 and 3 with the same numbers of observations per condition as in the experiment (see Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002, for examples with differing numbers of observations). The model was then fit to the simulated data set generated from the model and the SD in the value of each parameter across the data sets was computed. The question is whether the variability across subjects in the real data reflects only this variability arising from different random samples of data or a combination of this variability and meaningful differences among subjects.

The combination of two sources of variability is computed by taking the square root of the sum of the squares of the SDs from each source (i.e., variances add). As an example, suppose the true SD in a model parameter across subjects was 1.0 and the SD from the Monte Carlo was 0.8, then the combination of the two would be 1.28. In other words, even with a random component (from limited sample size, estimation error, etc.) that is 80 % the size of the variability across subjects, it is possible to measure individual differences reliably. For example, there were correlations in the range of 0.5–0.6 for IQ with drift rate in Ratcliff et al., (2010, 2011).

The key parameters for the integrated models that produce drift rates are the baseline drift rate v_n, the multiplying constant (c), and the decay constant for the exponential gradient model and the SDs for the overlap models. Table 5 shows the SDs in the model parameters from the real data and the Monte Carlo data. The ratios of the Monte Carlos to the real data are small enough for all three parameters for all three models that individual differences can be reliably measured. The ratios for the exponential gradient model for τ, c, and v_n were 0.83, 0.65, and 0.58 respectively. For the Gaussian overlap model, the ratios for σ, c, and v_n were 0.38, 0.63, and 0.68 respectively, and for the exponential overlap model, the ratios for τ, c, and v_n were 0.68, 0.51, and 0.51, respectively.

One point to note is that the SD parameter for the Gaussian overlap model is better estimated than the decay parameters for the exponential gradient and overlap models. This suggests that the Gaussian overlap model might provide a better measure of individual differences than the two exponential models. Even so, these parameters correlate highly over individuals (0.92 for the exponential gradient model and Gaussian overlap models, 0.90, for the exponential gradient and exponential overlap models, and 0.96 for the two overlap models) showing that they provide similar measures of individual differences (as would be expected because they are based on the same data).

For the three diffusion model parameters nondecision time, boundary separation, and across-trial range in nondecision time, the ratios are small enough to measure individual differences (as in Ratcliff et al., 2010, 2011). However, the ratios for across-trial SDs in drift rate and the range of starting points are too large to allow anything beyond the largest differences among individuals to be measured.

Fig. 4 shows correlations among pairs of model parameters, scatter plots for these pairs of parameters, and histograms of each model parameter for the exponential gradient model for the experimental data. The histograms show symmetric or slightly right-

Table 5
SDs in Model Parameters from Fits to Data and Fits to Monte Carlo Simulated ata for Experiment 1.

Model	Source	a	T_{er}	η_o	η_n	s_z	s_t	σ/τ	c	v_n	v_p
Exponential gradient	data	0.010	0.058	0.090	0.068	0.024	0.069	0.639	0.077	0.057	0.060
	monte	0.006	0.011	0.074	0.084	0.023	0.018	0.530	0.050	0.033	0.047
Gaussian overlap	data	0.010	0.056	0.087	0.068	0.025	0.068	0.493	0.299	0.057	0.058
	monte	0.007	0.012	0.089	0.095	0.025	0.020	0.185	0.188	0.039	0.039
Exponential overlap	data	0.010	0.058	0.071	0.063	0.025	0.068	0.749	0.210	0.059	0.059
	monte	0.005	0.011	0.066	0.076	0.022	0.017	0.384	0.120	0.030	0.040

The model parameters are listed in the footnote to Table 3.

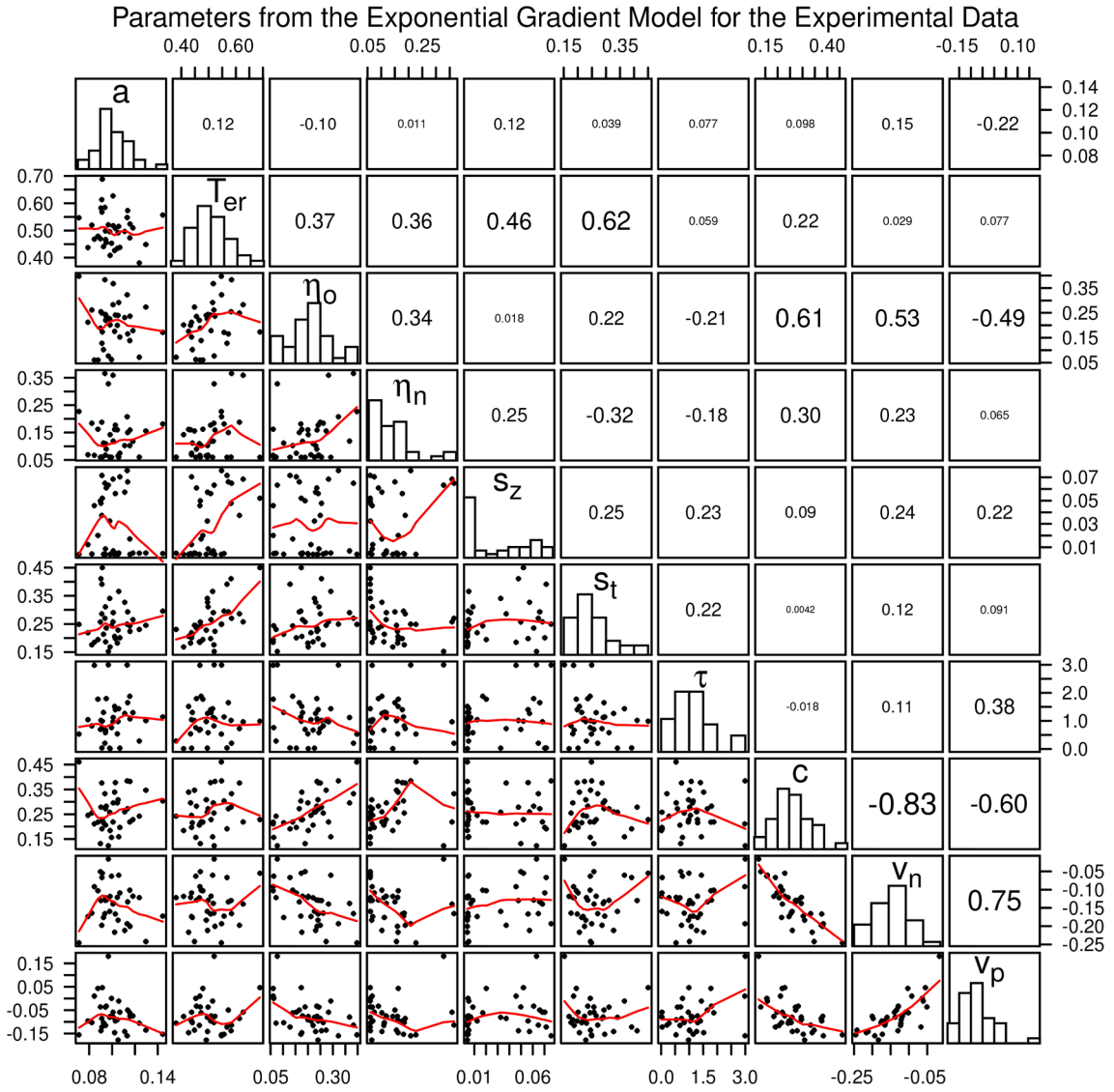

Fig. 4. Scatter plots, histograms, and correlations for model parameters for the fit of the exponential gradient model to the experimental data. Each dot represents the parameter from an individual subject. The identity of the comparison in each off-diagonal plot or correlation is obtained from the task labels in the corresponding horizontal and vertical diagonal plots. The lines in the bottom left of the plots are lowess smoothers (from the R package). The model parameters are: boundary separation a, nondecision time T_{er}, SD in drift rate across trials for studied numbers η_o and for "new" test numbers η_n, across-trial range in starting point s_z, across-trial range in nondecision time s_t, decay constant τ, drift rate multiplying constant c, drift rate for unrelated non-studied numbers v_n, and drift rate for numbers from the prior list v_p. A correlation of 0.31 with 28 degrees of freedom (30 subjects) is significant at the 0.05 level. But there are many correlations in the figure so care must be taken in interpreting correlations around that number.

skewed distributions except for across-trial SD in drift rate and across-trial range in starting point. A correlation to note is the negative correlation between the multiplying constant c and the baseline drift rate v_n. This means that if discriminability increases, c goes up then v_n goes down to make a higher hit rate and a lower false alarm rate. Two other correlations to note are that nondecision time correlates with across-trial range in nondecision time and across-trial SD in drift rate correlates with the drift rate parameters (c and v_n). These suggest a scaling effect in which an increase in η_o is accompanied by an increase in variability (as in Ratcliff et al., 2001, p. 337).

Fig. 5 shows the same plot for the Monte Carlo simulated data for comparison with the individual differences in Fig. 4. As for Fig. 4, histograms show symmetric or slightly right skewed distributions except for across-trial SD in drift rate and across-trial range in starting point. The same negative correlation is observed between c and v_n and these parameters are correlated with across-trial SD in drift rate. There are also correlations between boundary separation and drift rates and between nondecision time and drift rate parameters. All these patterns have been observed and explained before (Ratcliff & Tuerlinckx, 2002, Figure 6 and Table 3) in terms of

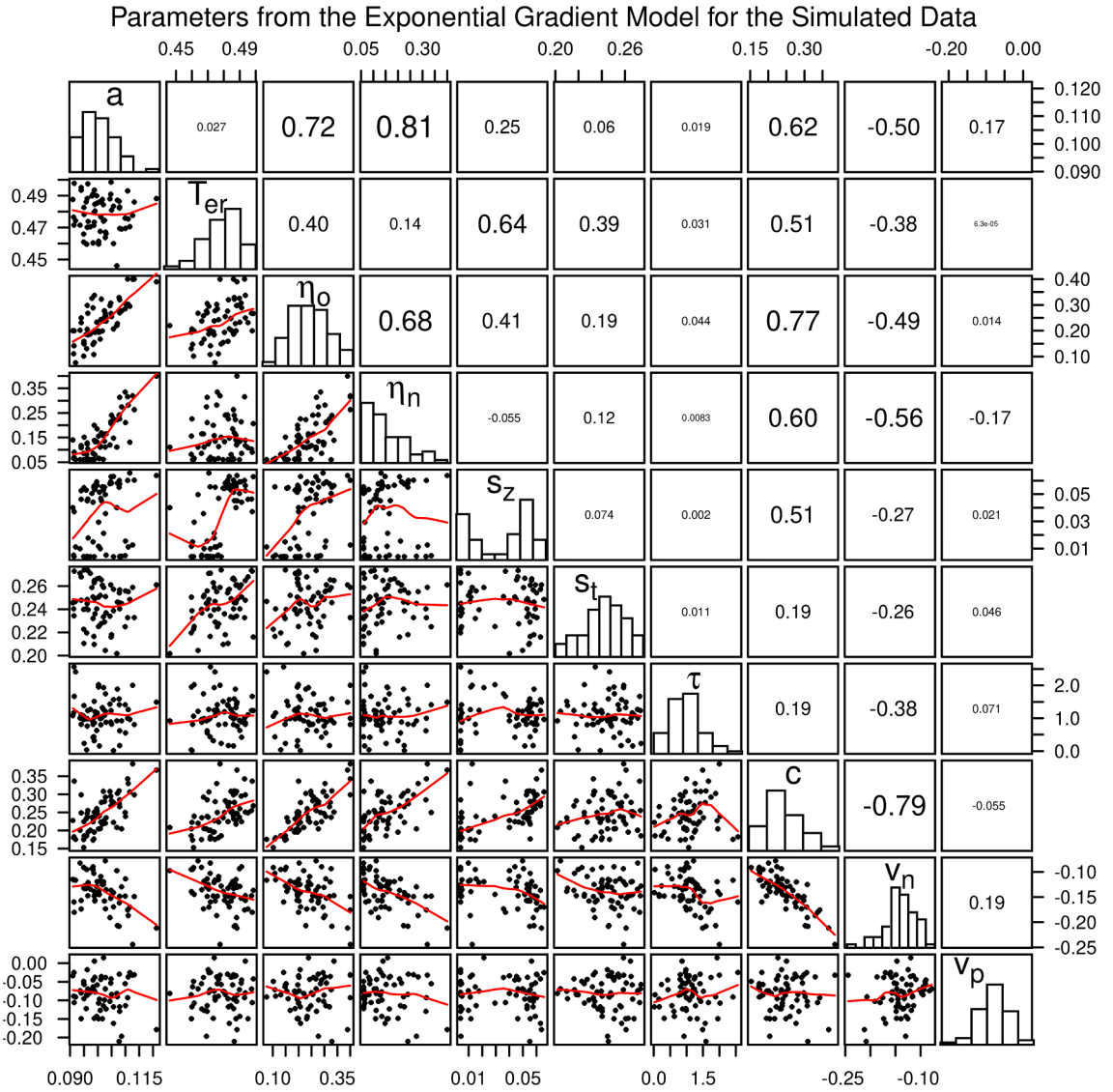

Fig. 5. The same plot as in Fig. 4 for the fits of the exponential gradient model to the Monte Carlo simulated data from the exponential gradient model. There are 64 data points in the figure and a correlation of 0.21 with 62 degrees of freedom is significant at the 0.05 level. There are many correlations in the figure so care must be taken in interpreting correlations around that number.

how model parameters covary to accommodate random variation in data.

Another notable result that occurs for the parameters for both the experimental data and the Monte Carlo data is that the exponential decay parameter is not correlated with the drift-rate parameters c and v_n. This suggests that they represent different individual differences (Fig. 4) and do not tradeoff in estimation (Fig. 5).

The simulated data were generated from all three integrated models and the three models can be fit to each data set to determine whether the models are identifiable based on the numbers of observations per subject in Experiment 1. The aim is to determine whether a model fits data generated from itself better than the other models fit the data. Because the goodness of fit values for the three models are so close to each other (Table 2), I did not expect them to be identifiable. For the simulated data from for exponential gradient model, the exponential gradient, the Gaussian overlap, and the exponential overlap models were best fit for 22, 22, and 20 (respectively) of the 64 data sets. For data from the Gaussian overlap model, the corresponding numbers were 21, 24, and 19, and for data from the exponential overlap model, the numbers were 21, 25, and 18. These results show that the models are not identifiable based on the amount of data collected in one hour of testing using this paradigm.

These model fits can also be used to see if the model parameters are correlated across fits to the data sets. For the drift rate parameters, decay constant/SD in the overlap distributions, τ/σ, the baseline new drift rate v_n, and the multiplying constant c, the correlations across the 9 comparisons (3 pairs of models by 3 simulated data sets) were between 0.71 and 0.97 for the decay constant/SD, between 0.91 and 0.98 for the baseline new drift rate, and between 0.43 and 0.91 for the multiplying constant. For the multiplying

constant, the models operate a little differently because for the overlap models, the height of the distributions is a function of both the SD and the multiplying factor. For the two overlap models the correlations for the multiplying constant were between 0.83 and 0.91. These correlations show that the models are producing estimates of decay in drift rates as a function of distance that are similar to each other and so are describing the data in the same way.

7. Experiment 2

This experiment provides a replication of the memory task used in Experiment 1 and adds a number discrimination task so that the distance effects in the two tasks can be compared. This experiment was conducted in response to a reviewer of a previous version of this article who raised the question of whether the distance effects obtained in the memory task and those obtained in number discrimination tasks (e.g., Moyer & Landauer, 1967; Dehaene et al., 1990) are similar. The discrimination task in Experiment 2 used two-digit numbers and was the same as that used by Ratcliff (2014, Experiment 5) and Ratcliff et al. (2015) and similar to that used by Dehaene et al. (1990). Specifically, subjects decided whether a two-digit number was greater or less than 50.

To anticipate, in number discrimination, drift rate was linear as a function of the distance between a test number and 50. This contrasts sharply with the distance effect found in Experiment 1 where the function was approximately exponentially decreasing (Fig. 3A).

7.1. Method

I aimed for 30 subjects, but one did not finish the number discrimination task and results and analyses are for the data from 29 subjects. They were recruited from The Ohio State University student body and took part in two 45-minute sessions, one for each task, and the order of the tests was randomized. They were paid $12 per session.

The number memory task was the same as that in Experiment 1. For the number discrimination task, on each trial, a white number between 10 and 90 was displayed on a black background in the middle of a laptop screen. The number remained on the screen until a response key was pressed. Then the screen cleared, accuracy feedback consisting of a smiling or frowning face was displayed for 500 ms, there was a 100 ms blank screen, and then the next trial. Subjects were instructed to respond "small" with one key if the number was between 10 and 49 and "large" with another key if it was between 51 and 90 and to do so as quickly and accurately as possible. There were a total of 17 blocks of trials, 80 trials per block, with each of the possible numbers tested once in each block in random order.

For the number memory task, subjects finished an average of 59 study-test lists in the time allotted and for the number discrimination task, an average of 15 blocks. Unlike Experiment 1, none of the subjects performed with fast guesses (because they were paid and also had been tested in other experiments in our laboratory and were found to be reliable) and so none were eliminated. The subjects were tested individually with a research assistant monitoring them.

7.2. Results

For the number memory task, RTs less than 350 ms and greater than 3000 ms were eliminated which removed 2.8 % of the data (the same cutoffs were used as for Experiment 1). For the number discrimination task, RTs less than 250 ms and greater than 3000 ms were eliminated which removed 0.2 % of the data (accuracy was 0.761 for responses between 250 and 300 ms so the lower 250 ms value was chosen as the lower cutoff).

The three integrated models were fit to the data from the number memory task in exactly the same way as for Experiment 1. A summary of the data is shown in Table 1. Accuracy was about 5 % higher for "old" responses than in Experiment 1 and accuracy for "new" responses was about the same. Mean RTs were between 15 and 50 ms longer than those for Experiment 1.

I performed one-way ANOVAs on accuracy values and correct mean RTs for 1-, 2-, and 3-distant numbers and remote numbers. For accuracy, the effect of distance was significant, $F(3,84) = 13.6$, $p = 2.5 \times 10^{-7}$, $\eta_p^2 = 0.33$ (partial η_p^2) and for mean RTs, the effect was also significant, $F(3,84) = 4.0$, $p = .011$, $\eta_p^2 = 0.12$. The distance effects were significant as in Experiment 1.

The values of the diffusion model parameters are shown in Tables 2 and 3. The values of the boundary separation, nondecision time, and across-trial variability parameters differ by less than 10 % from Experiment 1 except that the across-trial SD in drift rate for new numbers is larger by about 25 %. The drift rates in Table 3 are higher for old numbers and more negative for new numbers.

Fig. 3B shows the same plots from Experiment 2 as in Fig. 3A for Experiment 1 with drift rates from the exponential gradient model plotted against those from the default model. The results show a slightly worse match between the exponential gradient model drift rates and the drift rates from the default model compared with Experiment 1. This slight mismatch may be due to a slight misfit

Table 6
Accuracy, Mean Correct and Error RTs, and Drift Rates for the Number Discrimination Task (in Experiment 2).

Stimulus number	Pr(correct)	Mean correct RT (ms)	Mean error RT (ms)	Drift rate *v*
10–19 and 81–90	0.976	491	445	0.495
20–29 and 71–80	0.961	512	464	0.449
30–39 and 61–70	0.936	534	501	0.400
40–49 and 51–60	0.913	550	477	0.315

between accuracy values for unrelated non-studied numbers for the default model (see the first two rows for Experiment 2 in Table 3). Drift rates and predicted accuracy values for all the models for Experiment 2 are shown in Table 3.

The same model comparisons using G-square, AIC, and BIC were conducted as for Experiment 1. The results are presented in Table 4 and are similar to those for Experiment 1 with the default model selected by G-square and both AIC and BIC supporting the exponential gradient model. As for Experiment 1, the differences in goodness of fit for the three models are not large and they provide similar accounts of the experimental data.

For the number discrimination task, accuracy and mean RTs for correct and error responses are shown in Table 6. Accuracy is high and decreases and RTs increase as the test number approaches the referent. Conditions with no error RTs can be fit with the diffusion model as long as some conditions have error RTs because the RTs for correct responses in the conditions with no errors are sufficient to constrain the drift rates (additional discussion is presented in Ratcliff, 2014).

The default model was fit as for the number memory task, with a different drift rate for each of eight groups of data. There is one important difference in the fits compared with the number memory task. Because RTs and accuracy values were reasonably symmetric for "small" responses to small numbers and "large" responses to large numbers, conditions for small numbers were collapsed with conditions for large numbers for ranges equidistant from the referent (see Table 6). The model was then fit to correct and error responses and because these were collapsed, the starting point has to be equidistant from the decision boundary ($z = a/2$). The 8 groups of data were from stimuli 10–14 combined with 86–90, 15–19 combined with 81–85, …, and 45–49 combined with 51–55. These are the same combinations in Ratcliff et al. (2015) and the same fitting program was used. Pairs of these 8 conditions were combined to produce 4 pairs for more compact display of the results in Table 6 though all 8 drift rates are plotted in Fig. 3C.

The model parameters are shown in Table 2 and mean drift rates are shown in Table 6. The parameter values are very similar to those for the number discrimination task presented in Table 2 of Ratcliff et al. (2015). The drift rates are shown in Fig. 3C, plotted as for Fig. 3A and 3B. The plot shows roughly linear functions that closely replicate those in Ratcliff (2014) and Ratcliff et al. (2015). The important point is that drift rate as a function of distance shows quite different patterns for the two tasks. For the number memory task, the functions are decreasing with a drop to baseline when the test number is 3-distant. In contrast, the number discrimination task produces linear drift rate functions as a function of the difference between the test number and the referent.

Because the same subjects were tested on both tasks, individual differences in model parameters can be examined using correlations. First, the correlation between boundary separation for the two tasks was 0.47 and the correlation between nondecision times was 0.53. Correlations between the across-trial variability parameters were all small (although the two values for the memory task, η_o and η_n, correlated 0.42). To compute correlations for drift rates, the values of drift rates were averaged to give one number per subject for the number discrimination task. For the number memory task for the exponential gradient model, the difference between the multiplying constant and the drift rate for unrelated new numbers (i.e., the difference in drift rate between "old" and unrelated "new" drift rates which represents old/new discriminability) was used. These two correlated 0.37. For 29 subjects there are 27 degrees of freedom for the correlations and a value of 0.31 is significant at the 0.05 level. So the four correlations above are reliable.

To compute the correlations in distance effects, the slope of drift rate as a function of distance for the number discrimination task was computed for each subject. The correlation between this slope and the decay constant for the exponential gradient model was -0.24 and the correlations between the slopes of the number discrimination task and the SDs for the Gaussian and exponential overlap models were -0.19 and -0.16 respectively. These low negative values hint that with more power there may be a relationship between the range of the decay in memory as a function of numerical distance and the decrease in drift rate with distance from the referent in number discrimination. However, the effects might wash out with more data and higher power.

The results from Experiment 2 show that the number discrimination task and number memory task produce quite different distance effects. The former is linear with distance and the latter is roughly exponential. Model parameters correlate between tasks and overall memory discriminability is correlated with number discriminability, which replicates results from Ratcliff et al. (2015).

8. Discussion

In this article, three models for the representation of numbers in memory (Fig. 1) were developed and tested against the data from a recognition memory task. The data show a decrease in confusability as a function of the numerical distance between a studied number and a test number. The representation models were combined with the diffusion model to produce integrated models of representation and decision-making that produce predictions for accuracy and RT distributions. The first model was an exponential gradient model in which drift rate decreases as a function of numerical distance. The second and third models were Gaussian and exponential overlap models in which numbers are represented as distributions over number and drift rate is a function of the overlap of the areas between the study and test numbers (Fig. 1).

The exponential gradient model fit the data from both number memory experiments modestly better than the two overlap models. All three gave measures that produced good fits to the data for old/new memory discrimination, c-v_n for the exponential gradient model and c multiplied by the overlap area minus v_n for the overlap models. All three also provided estimates for the decrease in drift rate as a function of distance, the decay constants τ in the two exponential models and the SD σ in the Gaussian model. For the 6 conditions of Experiments 1 and 2, there were 66 independent bins of probability mass for the correct and error RT distributions (i.e., 66 degrees of freedom) that are fit with the 11 parameters of each model.

The Gaussian overlap model had the best properties for measuring individual differences because it had the smallest variability in the estimates of the SD in the overlap distributions (σ) relative to individual differences in the SDs from the data. This is similar to earlier studies (Ratcliff et al. 2010, 2011) in which variability across subjects in parameter estimates were larger than variability in parameter estimates from the fitting process (based on the number of observations), thus allowing individual differences in, for

example, IQ, to be correlated with other variables such as drift rate. Of note, the correlations between old/new discriminability (e.g., c-v_n) and the decay constant in the exponential model and between old/new discriminability and the SD in the Gaussian overlap model are less than 0.1. This suggests that they represent different individual differences.

Overall, the results from fitting the three integrated models to the data do not allow us to decisively choose among these models although the fits of the exponential gradient model are better numerically and more subjects are best fit by it than the two overlap models in both experiments. However, because the decay constant parameter and the SD parameters in the models correlate highly, use of any one of the models in applications will likely produce conclusions similar to those from the other models.

The exponential gradient model is related to the exponential similarity models used in categorization research. The exponential function championed by Shepard (1987) is used in both Nosofsky's (1986) exemplar-based model and Nosofsky and Palmeri's (1997) exemplar-based random walk model. The latter model (for a two-category task) assumes that exemplars race to be retrieved with a rate that is an exponential function of their similarity to a category; in the random walk, a retrieved exemplar increments a random walk by plus 1 for the category to which it belongs and minus 1 for the other category, until one of the two decision boundaries is reached. Unlike diffusion models, the exemplar-based random walk model does not have across-trial variability in the components of decision processes or within-trial variability; all of the variability in processing comes from the random selection of exemplars.

In the tasks used in applications of the Nosofsky (1986) and Nosofsky and Palmeri (1997) models, similarity has two dimensions. Two dimensions for a number memory model that could be examined are numerical distance in the units digit and numerical distance in the tens digit. For example, if 48 was a test number, then it would be 1-distant on the units dimension from the study number 47 and it would be 1-distant on the tens dimension from the study number 58. For the exponential gradient model, a test number's drift rate would be a weighted sum of a test number's distance from a studied number in 1′s units and 10′s units (as in Equation (1)). For the overlap models, drift rate would be the combination of the overlap of a test number's distribution with a studied number's distribution for 1′s and the overlap for 10′s. In addition, it might also be that some numbers are more memorable, for example, 44, 55, 30, and 40, etc., and/or that transpositions have additional similarity to a studied number (e.g., 84 if 48 was studied). The study by Kang and Ratcliff (2020) mentioned in the introduction is an example of how multi-dimensional representations can be used to produce the drift rate for a test item as a function of multiple variables. In that study, drift rate was made a function of both numeric and non-numeric variables and so the relative contributions of the two (or more) sources of evidence used in making decisions could be evaluated.

The distributed representations in the overlap models considered here are based on those used for a perceptual matching task by Ratcliff (1981, also Gomez, Ratcliff & Perea, 2008 and confidence judgments, Ratcliff & Starns, 2009, 2013). The overlap model for number memory in Experiments 1 and 2 adds to the success of these earlier models in that it used the overlap of distributed representations integrated with a model representing decision process to determine responses for number memory.

The results from the experiments presented here and diffusion model analyses of other number and numerosity tasks show that there can be very different representations for different tasks. In the number memory task used here, there is an exponential-like decrease in drift rate as a function of the numerical distance between a test number and the corresponding study number. In the number discrimination task, drift rate is a linear function of the distance between a test number and the referent. This result is important because it shows that the representation used in the task is a function of the task and different representations are used in different tasks. in the number discrimination task, the task requires a judgment based on numerical distance (from 50) and so this makes the discrimination focus on numerical distance. But the number memory task does not require the decision to be made on the basis of numerical distance, but numerical distance does affect performance. Furthermore, examination of raw accuracy or mean RTs alone in the number discrimination and number memory tasks would not provide the insight that the evidence driving the decision process is different for the two tasks.

There are two other studies in numerical cognition that use models of decision processes that find different effects for different paradigms. For a number-line task (Ratcliff & McKoon, 2020b), the representation of a two-digit number on a continuous scale is roughly normally distributed with a standard deviation that changes little as a function of the size of the number (e.g., for 10 to 90). For the numerosity discrimination tasks (Ratcliff & McKoon, 2018, 2020a) described in the introduction, two different representations, linear and logarithmic, were observed.

The important point is that sometimes representations that are not discriminable based on accuracy measures or RTs alone can be discriminable when the representations are embedded in models of decision processing that account for both accuracy and RT distributions for both correct and error responses (as in Ratcliff & McKoon, 2018). Although number memory and number discrimination have different representations in Experiments 1 and 2, it might be possible in the future to discover a small number of underlying representations that give rise to differences such as these.

The distance effects presented here are not the only representational features that are used in long-term memory for numbers. Numbers with repeated digits and numbers at the beginning or ends of decades may be more or less memorable and transpositions of studied digits might be hard to call "new." Similar experiments with designs that examine decay as a function of tens digits would likely show similar effects, but maybe with smaller effects. Perhaps the best way to study these possibilities would be to run an experiment with multiple sessions with many factors recorded, such as distances in terms of tens and unit digits between a test number and studied numbers (and earlier test numbers), repeated digits, whether numbers have digits that match other numbers in the study and test list, and so on. Then models could be developed with all these relationships included and fits would determine which factors were important in number memory.

In the experiments presented here, the effects of these factors were controlled for by averaging them or excluding them from the conditions measuring distance effects. A complete model of the representation of numbers in memory would have to explore these factors and include them in the model.

As I pointed out in the introduction, the ability to use mathematical information, in for example, learning how to do computations

and using them in practical situations, depends on memory for number. With model-based analyses of number memory, simple numerosity, and numeracy tasks, and the relations among them, we can begin to ask whether differences among individuals that are observed in such tasks pinpoint a numeracy ability that is linked to a single representation system (the approximate number system, for example) or to a larger group of skills. The individual difference measures derived from the experiments in this article (old/new discriminability and confusability) might in future research link to individual differences in other tasks or to other aspects of numeracy and memory. It is an open question whether they will map into a specific ability to use number information that depends on the task or simply general numerical ability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research was supported by funding from the National Institute on Aging (Grant numbers R01-AG041176 and R01-AG057841).

Data availability

Data will be made available on request.

Appendix. Fitting the integrated model

The integrated diffusion models were fit to the data for each task and each subject by minimizing a G-square multinomial log likelihood goodness of fit statistic with a general SIMPLEX minimization routine that adjusts the parameters of the model until it finds the parameter estimates that give the minimum G-square value. G-square is asymptotically equivalent to the chi-square method used by Ratcliff and Tuerlinckx (2002) who provide a full description of that method. The data entered into the minimization for each experimental condition are the 0.1, 0.3, 0.5, 0.7, 0.9 quantile RTs for correct and error responses and the corresponding accuracy values. The quantile RTs and the diffusion model are used to generate the predicted cumulative probability of a response by that quantile RT. Subtracting the cumulative probabilities for each successive quantile from the next higher quantile gives the proportion of responses between adjacent quantiles. For the G-square computation, these are the expected values, to be compared to the observed proportions of responses between the quantiles (i.e., the proportions between 0, 0.1, 0.3, 0.5, 0.7, 0.9, and 1.0, which are 0.1, 0.2, 0.2, 0.2, 0.2, and 0.1). G-square is defined as $G^2 = 2N\Sigma O ln(O/E)$, where N is the number of observations in the condition. Summing over this for all conditions gives the single G-square value to be minimized.

The G-square multinomial log-likelihood statistic allows model comparisons to be carried out using AIC and BIC test statistics. These penalize goodness-of fit to different degrees as a function of number of parameters. AIC $= G^2 + 2k$, where k is the number of parameters, and BIC $= G^2 + k\ln(M)$, where M is the total number of observations. The number of parameters for the integrated models are the same and there is one more parameter for the default model. To compute the mean relative AIC and BIC values, the values from Table 2 can be used. For both experiments, for AIC, 2 is added to G^2. For BIC, the mean values of M are 853 and 704 for Experiments 1 and 2 respectively, therefore, 6.75 and 6.52 are added to G^2 for Experiments 1 and 2 respectively.

References

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.
Clifton, C., & Gutschera, K. D. (1971). Hierarchical search of two-digit numbers in a recognition memory task. *Journal of Verbal Learning & Verbal Behavior, 10*(5), 528–541.
Corballis, M. C. (1967). Serial order in recognition and recall. *Journal of Experimental Psychology, 74*(1), 99–105.
Dale, H. C., & Baddeley, A. D. (1966). Remembering a list of two-digit numbers. *The Quarterly Journal of Experimental Psychology, 18*(3), 212–219.
De Rosa, D. V., & Morin, R. E. (1970). Recognition reaction time for digits in consecutive and nonconsecutive memorized sets. *Journal of Experimental Psychology, 83*(3), 472–479.
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.
Gomez, P., Ratcliff, R., & Perea, M. (2008). A model of letter position coding: The overlap model. *Psychological Review, 115*, 577–601.
Gould, I. C., Wolfgang, B. J., & Smith, P. L. (2007). Spatial uncertainty explains endogenous and exogenous cuing effects in visual signal detection. *Journal of Vision, 7*, 1–17.
Fontanesi, L., Gluth, S., Spektor, M. S., et al. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin and Review, 26*, 1099–1121.
Inglis, M., Attridge, N., Batchelor, S., & Gilmore, C. K. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin and Review, 18*, 1222–1229.
Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology, 120*, Article 101288.
Lee, C., & Estes, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning and Verbal Behavior, 16*, 395–418.
Maloney, E., Risko, E., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Pyschologica, 134*, 154–161.
Miletic, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife, 10*, Article e63055.
Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519–1520.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–109.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar based random walk model of speeded classification. *Psychological Review, 104*, 266–300.
Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical Bayesian parameter estimation for reinforcement learning and drift diffusion models: A tutorial and link to neural data. *Computational Brain & Behavior, 3*, 458–471.
Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rules in reinforcement learning. *Psychonomic Bulletin and Review, 24*(4), 1234–1251.
Price, G., 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.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88*, 552–572.
Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in a two choice brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review, 9*, 278–291.
Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review, 120*, 281–292.
Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 870–888.
Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model. *Decision, 2*, 237–279.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
Ratcliff, R., & McKoon, G. (2018). Modeling numeracy representation with an integrated diffusion model. *Psychological Review, 125*, 183–217.
Ratcliff, R., & McKoon, G. (2020a). Examining aging and numerosity using an integrated diffusion model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 2128–2152.
Ratcliff, R., & McKoon, G. (2020b). Decision making in numeracy tasks with spatially continuous scales. *Cognitive Psychology, 116*, Article 101259.
Ratcliff, R., Scharre, D. W., & McKoon, G. (2021). Discriminating memory disordered patients from controls using diffusion model parameters from recognition memory. *Journal of Experimental Psychology: General, 151*, 1377–1393.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. (1992). Testing Global Memory Models using ROC Curves. *Psychological Review, 99*, 518–535.
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Science, 20*, 260–281.
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83.
Ratcliff, R., & Starns, J. J. (2013). Modeling response times, choices, and confidence judgments in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*, 697–719.
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.
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.
Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*, 127–157.
Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*, 46–487.
Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115–136.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating the parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review, 9*, 438–481.
Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review, 125*, 33–46.
Sewell, D. K., Jach, H. K., Boag, R. J., & Heer, V. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin and Review, 26*, 868–893.
Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General, 115*, 58–61.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116*, 283–317.
Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research, 44*, 1297–1320.
Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52.
Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variability and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64*, 1–34.
Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652–654.
Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. In W.G. Koster (Ed.), (pp 276-315). *Attention and performance II*. Amsterdam: North-Holland.
Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics, 7*, 1–10.
Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology. Human Perception and Performance, 16*, 626–641. https://doi.org/10.1037//0096-1523.16.3.626
Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 163–178.
Ratcliff, R., & Frank, M. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24*, 1186–1229.