Individual Differences and Fitting Methods for the Two-Choice Diffusion Model

Roger Ratcliff and Russ Childers

The Ohio State University

Running Head: DIFFUSION MODEL, FITTING METHODS AND INDIVIDUAL DIFFERENCES

Address correspondence to:

Roger Ratcliff,

Department of Psychology,

The Ohio State University,

Columbus, OH, 43210

Phone number: 614 292 7916

Fax number: 614 688 3984

email ratcliff.22@osu.edu

Abstract

Methods of fitting the diffusion model were examined with a focus on what the model can tell us about individual differences. Diffusion model parameters were obtained from the fits to data from two experiments and consistency of parameter values, individual differences, and practice effects were examined using different numbers of observations from each subject. Two issues were examined, first, what sizes of differences between groups can be obtained to distinguish between groups and second, what sizes of differences would be needed to find individual subjects that had a deficit relative to a control group. The parameter values from the experiments provided ranges that were used in a simulation study to examine recovery of individual differences. This study used several diffusion model fitting programs, fitting methods, and published packages. In a second simulation study, 64 sets of simulated data from each of 48 sets of parameter values (spanning the range of typical values obtained from fits to data) were fit with the different methods and biases and standard deviations in recovered model parameters were compared across methods. Finally, in a third simulation study, a comparison between a standard chi-square method and a hierarchical Bayesian method was performed. The results from these studies can be used as a starting point for selecting fitting methods and as a basis for understanding the strengths and weaknesses of using diffusion model analyses to examine individual differences in clinical, neuropsychological, and educational testing.

Research using simple two-choice tasks to examine cognition and decision making has had a long history in psychology. The most successful current models of decision making are sequential sampling models, which assume that decisions are based on the accumulation of evidence from a stimulus, that there are moment-to-moment fluctuations in the evidence during the accumulation process, and that the amount of evidence needed to choose between alternatives is determined by response criteria, one criterion for each of the choices. The time taken to make a decision and which alternative is chosen are jointly determined by the rate at which evidence accumulates and the settings of the response criteria. Trial-to-trial variability in response time (RT) and accuracy is attributed to moment-to-moment variability in the evidence, trial-to-trial variability in decision criteria, trial-to-trial variability in the rate of evidence growth, or some combination of these. In the last 15 years, the number of research domains in which these models are applied has increased dramatically, mainly because the models can separate the settings of the criteria (which in turn determine speed and accuracy tradeoffs) from the quality of the evidence obtained from a stimulus. The model analyses can therefore be used to examine the effects of independent variables on these components of decision making separately.

The most widely studied sequential sampling model for two-choice decision making is the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008). In this model, the rate at which evidence is accumulated toward the response criteria is labeled “drift rate” and the moment-to-moment fluctuations in evidence are labeled “within-trial” variability. The noise in the accumulation process means that there will be variability in the RTs of nominally identical stimuli and that some responses will be incorrect. Total RT is a function of the drift rate, the criteria settings, and the time taken by processes outside the decision process such as stimulus encoding and response execution. The latter processes are combined in the model into one parameter, “nondecision” time. The model explains how choice probabilities and the locations, dispersion, and shapes of RT distributions for correct responses and errors vary as a function of independent variables such as stimulus discriminability, response bias, and speed-accuracy tradeoff settings (e.g., Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Smith, Ratcliff, & McKoon, in press). The model allows the effects of variables such as these to be mapped to the individual components of processing that underlie decisions--drift rates, criteria

settings, and nondecision time.

This article focuses on what the model can tell us about differences between individual subjects, differences between individual subjects and groups of subjects, and differences between groups. Previous studies have examined decision making for children (Ratcliff, Love et al., 2012), sleep-deprived individuals (Ratcliff & Van Dongen, 2009), aphasic individuals, (Ratcliff, Perea, et al. 2004), hypoglycemic individuals (Geddes, et al., 2010), adults and children with ADHD (Mulder et al., 2010; Karalunas & Huang-Pollock, 2013), individuals with dyslexia (Zeguers et al., 2011), and individuals with anxiety or depression (White et al., 2009, 2010a, 2010b). The model has also been used to examine decision processes in neurophysiology with single-cell recording methods (Ratcliff, Cherian, & Segraves, 2003; Ratcliff, Hasegawa, et al., 2007; Gold & Shadlen, 2007), EEG signals (Philiastides, Ratcliff, & Sajda, 2006; Ratcliff, Philiastides, & Sajda, 2009), and fMRI methods (Kuhn et al., 2011).

Research with older adults (60-90 years old) is noteworthy. Older adults respond more slowly than young adults in most cognitive tasks but there is little difference in accuracy compared with young adults. The standard interpretation of this has been that all of older adults' cognitive operations are slower than young adults', 1.5 to 2.0 times slower. Application of the model showed instead that their slower speeds are most often the result of more conservative response boundaries (i.e., the two boundaries set further apart) and slower nondecision times (McKoon & Ratcliff, 2012, 2013; Ratcliff, Thapar, & McKoon, 2001; 2003; 2004; 2010, 2011; Spaniol, Madden, & Voss, 2006). The older adults adopt more conservative decision criteria because they are more concerned about making errors (e.g., Starns & Ratcliff, 2010, 2012). In contrast, for many tasks, drift rates for older adults were not significantly smaller than those for college students.

For individual subjects, the model's estimates of individuals' components of processing provide values that can be correlated with such variables as IQ, age, and severity of diagnosed impairments. For example, Ratcliff, et al. (2010, 2011) found that drift rate correlated with IQ and Schmiedek et al. (2007) found that drift rate correlated with working memory measures.

In the literature on neuropsychological testing, it is difficult to find any test that comes from cognitive research more recent than the 1980's. Current work in cognition and cognitive modeling

has had no impact. For example, a recent edited book, “Information processing speed in clinical populations” (DeLuca & Kalmar, 2008) makes no reference to modeling RTs and a review article of intelligence and speed of processing (Sheppard, 2008) likewise makes no reference to modeling approaches. Given this disconnect between theory and applications, models that deal with speed and accuracy in rapid decision making are a natural domain to explore.

Standard neuropsychological tests often measure performance on several different tasks, with relatively few trials on each task, and then average the results into a single indicator of some ability, such as working memory or speed of processing. From the point of view of cognitive modeling, multi-task tests represent a compendium of disparate processes that may share some common features. However, if there are differences in processing components across the tasks, they might be averaged out. This neuropsychological testing approach is advantageous for many uses, especially when deficits are large, but in some situations, we may want to understand what deficits there are in components of processing with a model-based approach.

One very important issue is practice effects. If performance changes radically over the first tens or few hundreds of trials, and the changes are not consistent across individuals, then the data from these two choice tasks and from model analyses may be of limited use when only a small amount of data is available. Then, the main application of the model would be to cases in which enough data has been collected to provide stable parameter estimates. But this would be in conflict with the limited amount of time that can be given to any one experiment in a neuropsychological testing battery. In looking at deficits using model-based analyses, and if we had an experimental paradigm that was sensitive to the kind of deficit, then we would expect a wide range of parameter values with some in the normal range and with some outside the normal range. The experiments presented below used a homogeneous group of undergraduates (but some may not be particularly motivated) and serve as a benchmark for examining effects of practice and number of observations.

It is important to distinguish between four sources of variability that are relevant to applications of the diffusion model. First, when the model is applied to the data from a group of subjects, the parameters of the model that best fit each subject’s data provide estimates of the differences among individuals in those parameters. This variability can be used to determine if one or more parameter values for an individual are significantly different from the values for the group.

Second, the parameter values can be used to compute standard errors which can then be used as the basis for determining whether one group differs significantly from another group. These standard errors represent the variability in the group means and they can be made smaller by increasing the number of subjects in the group. The third source of variability is the sampling variability in the model parameters, that is, given the number of observations, how close are the parameters recovered to the parameter values that generated the data. For example, if the diffusion model was used to generate simulated data, the parameters recovered from the simulated data would be more variable if there were 100 simulated observations compared with 1000 observations. In typical applications of the diffusion model, this source of variability is typically 3 to 5 times smaller than the variability due to differences among individuals for 45 min. of data collection. The fourth source of variability comes from across-trial variability in model parameters under the assumption that subjects cannot hold components of processing exactly the same from one trial to the next. Across-variability is also needed for model to explain the relative speeds of correct and error RTs.

The focus in this article is on the first three sources of variability, though in all the model fits, estimates are obtained for the across-trial variability parameters.

Our goal for the model is that, in future research, it will give insights for clinical, educational, and neuropsychological testing applications. The highest bar is that the model contributes to diagnoses of cognitive impairments in individuals. The aim is to compare a possibly impaired individual to a matched group of normal individuals, which means that normal ranges of drift rates, criteria, and nondecision times must be estimated and the SDs across individuals calculated in order to determine whether and how far an individual is outside the normal range.

Many of the key questions in this article are about power. First, when the values of the parameters estimated from the model are correlated with variables like IQ, only relatively low precision is needed for individuals' parameter estimates because, for most paradigms in cognitive research, the ranges of the estimated parameters across subjects are quite large. This means that significant correlations can be obtained even when there are small numbers of observations for which the variability in parameter estimates is moderately large. Second, in determining whether two populations differ, variability in the estimation of parameters can be traded off with number of subjects because the SE in parameter estimates decreases as the number of subjects increases.

Third, unlike these first two applications, when trying to detect whether an individual falls outside the normal range of the model parameters, high power and low variability in parameter estimates are needed.

For all these applications of the model just discussed we need to understand how large are SDs in individual differences across subjects as well as SDs in model parameters from fitting the model to data as a function of number of observations in the sample. We may also have to understand how the SDs differ as a function of the values of the parameters, because SDs will differ for different values of the parameters.

We conducted two experiments, one numerosity discrimination and the other lexical decision. In the numerosity task, subjects decided whether a number of asterisks displayed on a computer screen was above or below 50. For lexical decision, there were four conditions: words that occur with high frequency in English, words that occur with low frequency, words that occur with very low frequency, and nonwords. For numerosity, the numbers of asterisks were collapsed into two conditions, easy and difficult. The lexical decision design is the same as that used for many memory tasks, for example, recognition of strongly encoded versus weakly encoded words (for diffusion model applications, see Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar & McKoon, 2004, 2010). The design of the numerosity experiment is the same as that frequently used with perceptual tasks including brightness, letter, motion, number, and length discrimination (Ratcliff, in press; Ratcliff & Rouder, 1998; Ratcliff et al., 2001, 2003; Thapar, Ratcliff, & McKoon, 2003). For each experiment, the model was fit to the individual subjects' data and the means of the best-fitting parameter values were averaged across subjects.

We used the data from the two experiments to address three questions about application of the model. The first was whether there are significant practice effects across trials. The second was what are the smallest numbers of observations needed to estimate the model parameters with reasonably small SE's. Ideally, the parameter estimates for early trials would be consistent with those for later trials and the estimates from relatively small numbers of trials would be consistent with those from larger numbers. If so, the model could be used for the short amounts of times that are often necessary for measuring differences between an individual and a group (e.g., to look for

cognitive deficits). The third question was what size differences relative to SD's in model parameters would allow differentiation of individuals from groups and groups from each other.

In the simulations and experiments below, the number of conditions is smaller than in many but not all experiments in cognitive psychology. This is because the experiments are meant to be representative of the kind that might be performed in a neuropsychological or educational testing context (i.e., with limited testing time with few conditions with low numbers of observations).

Following the experimental study, we tested a number of methods by which parameter estimates are obtained from two-choice data. The first three were two chi-square methods using binned data and the maximum likelihood method (MLH) (Ratcliff & Tuerlinckx, 2002). The other five were recently developed diffusion-model fitting packages, DMAT, with and without correction for contaminant RTs (Vandekerckhove & Tuerlinckx, 2007, 2008), fast-dm (Voss & Voss, 2007, 2008), the non-hierarchical HDDM (Wiecki, Sofer, & Frank, 2013), and EZ (Wagenmakers, van der Maas, & Grasman, 2007). In addition, we tested a hierarchical model from the HDDM package with a more limited set of simulated data.

For Simulation Study 1, the aim was to determine how well each method performed in terms of individual differences, that is, providing the correct ordering of parameters across individuals. A method might produce estimates that are biased away from the true values, those from which the simulated data were generated, but if the ordering is correct, then it can be used as a measure of individual differences. For examining differences among individuals and the relationship of such differences to clinical or educational tests, a more accurate order would be more important than accurate recovery of true values.

We used the best-fitting parameter values from fits to the two experiments to generate simulated data using the designs from the two experiments. For each experiment, different studies used different numbers of observations and each study was performed with and without contaminants. For each parameter, its value for each simulated data set was drawn from a normal distribution with the mean and SD from the parameter values from the fits to Experiment 1 or Experiment 2. Each combination of parameters produced a data set for each simulated subject.

For Simulation Study 2, the aim was to determine how well each method recovered the true

values of the parameters. The fitting methods were evaluated by how much their estimates were biased away from the true values and the variability in these estimates.

For Simulation Study 1, a few of the combinations of parameter values used to generate the simulated data were combinations that are unlikely to exist for real data (because they were drawn from independent normal distributions); for example, a small value of nondecision time may never be associated with a large value of across-trial variability in nondecision time. For Study 2, the combinations were chosen to be representative of what is typically observed in real experiments, 16 such combinations for the numeracy design and 32 for the lexical decision design. For both simulations, data were simulated for 64 subjects, each with and without outlier RTs, for numbers of observations ranging from 40 to 1200.

Estimates of variability in model parameters given the number of observations in a sample could be obtained theoretically from the Hessian matrix, but this is computationally very difficult, and simple Monte Carlo (parametric bootstrap) methods can be used instead (e.g., Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). In these methods, sets of simulated data are produced with some number of observations representing different data sets (replications) from the same subject with the same parameter values. The model is fit to these simulated data sets and then SD's in parameter values can be obtained from these sets of values. This measure represents the variability obtained if repeated measures from the same parameter values were obtained and the model fit to these. This is different from Bayesian posterior measures of the variability in the parameter values given the data. Both are illuminating, but we feel the former is more important in evaluating individual differences.

For Simulation Study 3, we tested the nine-quantile chi square against a hierarchical Bayesian fitting method. Hierarchical methods have been demonstrated to be superior to standard methods when there are low numbers of observations per subject. We compared parameter recovery as a function of number of observations in the numerosity design. The same combinations of parameter values were used as for Simulation Study 1 and data were generated for 64 subjects, with or without outlier RTs, for a range of numbers of observations from 40 to 1000.

The results of the simulation studies are intended to provide methodological guidelines:

which methods of fitting the diffusion model provide the correct ordering of the model's parameters and therefore can be used to determine individual differences for correlational analyses; which methods provide values that are not biased away from the true values and therefore can be used to test differences between groups of individuals; and how many subjects for how many numbers of observations does it take to produce sufficiently useful estimates of parameters. If the design of an experiment is substantially different from the numeracy and lexical decision designs, then the studies here can provide a roadmap for evaluating methods of fitting the model and the power they provide for determining differences among individuals and differences among groups. For the sections below, we begin with a detailed description of the diffusion model.

The Diffusion Model

In the diffusion model, two-choice decisions are made when information accumulated from a starting point, *z*, reaches one of the two response criteria, or boundaries, *a* and *0*. The drift rate of the accumulation process, *v*, is determined by the quality of the information extracted from the stimulus in perceptual tasks and the degree to which a test item matches memory in memory and lexical decision tasks. It is assumed that the value of *v* cannot change during the accumulation of information. The mean time taken by nondecision processes is labeled T_{er}. In separating drift rates, criteria, and nondecision times, the model decomposes accuracy and RTs for correct and error responses into individual components of processing. It explains how the components determine all aspects of data, including mean RTs for correct and error responses, the shapes and locations of RT distributions, the relative speeds of correct and error responses, and the probabilities with which the two choices are made.

The model includes three sources of across-trial variability: variability in drift rate, variability in the starting point of the accumulation process (which is equivalent to variability in the criteria settings), and variability in the time taken by nondecision processes. Variability in drift rate expresses the idea that the evidence subjects obtain from nominally equivalent stimuli differs in quality from trial to trial. Variability in starting point expresses the idea that subjects cannot hold their criteria constant from one trial to the next, and variability in nondecision components from one trial to the next expresses the idea that subjects cannot hold nondecision time constant across

trials. Across-trial variability in drift rate is normally distributed with SD η, across trial variability in starting point is uniformly distributed with range s_z, and across trial variability in the nondecision component is uniformly distributed with range s_t. (Ratcliff, 1978, 2013, showed that the precise forms of these distributions are not critical.)

Subjects sometimes make responses that are spurious in that they do not reflect the processes of interest but instead are due to, for example, distraction, lack of attention, or concern for lunch. There are two assumptions for such contaminant responses, one for generating simulated data and the other for fitting the model to real data. To generate simulated data, on some proportion of trials (p_o), a random delay is added to the decision RT, where the delay time is chosen randomly from a uniform distribution with range from zero to 2000 ms in the simulations here. To fit the diffusion model, a mixture of processes is assumed, one the regular process and another a uniform distribution with a range from the minimum to the maximum RT for each experimental condition (Ratcliff & Tuerlinckx, 2002).

Subjects also sometimes make fast guesses with short RTs (e.g., between 0 and 200 ms). Experimentally these can be minimized by adding a delay (1-2 sec) between the response and the next test item with a message saying “too fast” if the response is shorter than say 200 ms. Often such fast guesses are produced because the subject is eager to leave the experiment. The HDDM package for fitting the model, discussed below, assumes a contaminant distribution that has a minimum value of zero and this can accommodate a small proportion of fast guesses. Other methods for eliminating them are excluding all responses below some cutoff value (e.g. 200 ms) or excluding all responses that occur before accuracy begins to rise above chance (this latter method is formally implemented in the DMAT package discussed below). Another way of looking for fast guesses is to include in the design of the experiment a condition with high accuracy. The proportion of errors in this condition would be an upper limit on the proportion of guesses.

When the model is fit to data, all of its parameters are estimated simultaneously for all the conditions in an experiment. The model is tightly constrained in several ways. The most powerful is the requirement that the model fit the right-skewed shape of RT distributions (Ratcliff, 1978, 2002; Ratcliff & McKoon, 2008; Ratcliff et al., 1999). Another is the requirement that differences

in the data between conditions that vary in difficulty must be captured by changes in only one parameter of the model, drift rate. Boundary settings and nondecision time do not change as a function of difficulty; doing so would require that subjects know the level of difficulty before the start of information accumulation.

The assumption that drift rates do not change as a function of boundary settings has been verified in experiments to date. The one exception is a recent study by Starns, Ratcliff, and McKoon (2012) who found that if subjects are given extreme speed instructions (to respond in a recognition memory task in under 600 ms), then drift rates are lower than with accuracy instructions. This is likely because subjects limit stimulus or memory processing in order to respond quickly; something that they do not otherwise do.

There have been hints in the literature that nondecision time can change as a function of either changes in difficulty or changes between boundary settings. In a few experiments, fits of the model have been modestly better with small differences in nondecision time between speed and accuracy instructions (e.g., Ratcliff & Smith, 2004, p348; Ratcliff, 2006); however, in most other experiments, the difference has been negligible (Ratcliff & McKoon, 2008, p895).

One of the most important functions of the diffusion model is that it maps RT and accuracy onto the same underlying components of processing, drift rates, boundary settings, and nondecision time, which allows direct comparisons between tasks and conditions that show the effects of independent variables in different ways. For example, a study by McKoon and Ratcliff (2012) examined associations between the two words of pairs that were studied for recognition memory. Memory was tested in two ways: either subjects were asked whether two words had appeared in the same pair at study or they were asked whether a single word had been studied. In the latter case, a single word was preceded in the test list by the other word of the same pair (“primed”) or by some other studied word (“unprimed”). For priming, the effects of independent variables showed up mainly in RTs whereas for pair recognition, they showed up mainly in accuracy. McKoon and Ratcliff found that age, IQ, and the semantic relatedness of the words of a pair all affected drift rates in the same ways for the two tasks, from which they concluded that priming and pair recognition depend on the same associative information in memory. This conclusion would not have been possible without the model.

Another important function of the model is that it allows direct comparisons between groups of subjects. For example, older adults are generally slower than younger adults and show larger differences among conditions. For pair recognition, for example, older adults' RTs for same-pair tests might be 1200 ms and for different-pair tests, 1400. For young adults, RTs might be 1000 ms and 1100. Ratcliff et al. (2011) found that the differences in performance between older and younger subjects were due to differences in all three components of the model: drift rates, boundary settings, and nondecision times. This contrasts with item recognition in which older adults (60-74 year olds) show little difference in drift rates compared with young adults.

A problem that comes up for some experimental designs is what we call the "small-n" problem, which is that the number of observations in some conditions of an experiment is less than the number usually needed to use the model, for example, less than 100. In some cases (e.g., neuropsychological testing and clinical applications), this is because only a limited amount of time is available for testing. In other cases, the limit is the number of items that are available, the number that can be constructed, or the number that can be used in an experiment before subjects develop expectations about what kind of test items to expect. As an example, White, Ratcliff, Vasey, and McKoon (2009) investigated differences between mildly depressed (dysphoric) subjects and control subjects in a lexical decision experiment. In the depression literature, there are only about 30 words for which there is good agreement among researchers that they express the negative meanings that are relevant to dysphoric individuals. Experiments on reading comprehension provide other examples. McKoon and Ratcliff (1986; 1992; 2013) tested whether subjects comprehend inferences like "the actress died" from sentences like "The director and the cameraman were ready to shoot close-ups when the actress fell from the 14th story roof" in an item recognition experiment. Constructing items like this is difficult and the number of items needs to be kept small so that subjects do not begin making the inferences on the basis of explicit strategies; McKoon and Ratcliff's 2013 experiments had only 32 such items in their experiments.

Our solution to the small-n problem is to include many filler items in an experiment and use the RTs and accuracy for these items to constrain fitting the model to the items in the experimental conditions. In White et al.'s (2009) experiment, there were 540 filler words and 510 nonwords and in McKoon and Ratcliff's (2013), there were 416 words that had been studied and

416 words that had not. The model is fit simultaneously to all the data with the result that the filler items largely determine all the parameters of the model except the drift rates for the conditions of interest. These drift rates can be estimated with enough precision to compare performance in conditions to each other and to compare performance of one subject group to another.

The degree of precision for drift rates can be considerably greater than what appears in RT and accuracy data. In White et al.'s experiments, there were no significant differences in RT or accuracy between dysphoric and non-dysphoric subjects whereas there were highly significant differences for drift rates (see also White, Ratcliff, Vasey, & McKoon, 2010).

Methods for Fitting the Diffusion Model to Data

The methods used most commonly have been the maximum likelihood method (MLH) (Ratcliff & Tuerlinckx, 2002) and three binned methods: the chi-square (Ratcliff & Tuerlinckx, 2002) method, the multinomial likelihood ratio chi-square (G^2 method, and the quantile maximum likelihood method (Heathcote, Brown, & Mewhort, 2002). The latter two methods are nearly identical, so we do not discuss the quantile maximum likelihood method further.

In all of these methods, from some RT (either a single RT for the MLH method or a quantile for a binned method) and the model parameters, the predicted probability density is computed. The expression for the cumulative density is:

$$G(t, \xi, \zeta) = P(\xi, \zeta) - \frac{\pi s^2}{a^2} e^{-(\zeta\xi/s^2)} \times \sum_{k=1}^{\infty} \frac{2k \sin(k\pi\zeta/a) e^{-\frac{1}{2}(\xi^2/s^2 + \pi^2 k^2 s^2/a^2)\, t}}{(\xi^2/s^2 + \pi^2 k^2 s^2/a^2)}$$

The expression for the response proportion for the choice is

$$P(\xi, \zeta) = (e^{-(2\xi a/s^2)} - e^{-(2\xi\zeta/s^2)})/(e^{-(2\xi a/s^2)} - 1)$$

These equations must be integrated over the distributions of drift rate, starting point, and nondecision time:

$$G(t, v, z) = \int_{(T_{er} - s_t/2)}^{(T_{er} + s_t/2)} \left(\int_{(z - s_z/2)}^{(z + s_z/2)} \left(\int_{-\infty}^{\infty} G(t - \tau, \xi, \zeta) N(\xi; v, \eta) U(\zeta; z, s_z) U(\tau; T_{er}, s_t) \right) \right) d\xi d\zeta d\tau$$

where

$$N(\xi;v,\eta) = \frac{1}{\sqrt{2\pi\eta^2}} e^{-\left(\frac{(v-\xi)^2}{2\eta^2}\right)}$$

and

$$U(x;a,s) = \frac{1}{s}, \quad a - \frac{s}{2} < x < a + \frac{s}{2}$$

To obtain the probability density from the cumulative density, $f(t_i)$, at a time, t_i, the value of $F(t_i)$ and the value for that time plus an increment, $F(t_i+dt)$ is computed, where dt is small (e.g., .001 ms). Then using $f(t)=(F(t+dt)-F(t))/dt$, the predicted probability density at t_i is obtained.

For the MLH method, the predicted probability density ($f(t_i)$) for each RT (t_i) for each correct and error response is computed and the product over all densities for all the RTs t_i is the likelihood ($L=\Pi f(t_i)$). To obtain the maximum likelihood parameter estimates, the value of the likelihood is maximized by adjusting parameter values using a function minimization routine. However, because products of densities can become very large or very small, numerical problems occur and so it is standard instead to maximize the log likelihood, i.e., the sum of the logs of the densities (summing logs of the values is the same as the log of the product of the values, log(ab)=log(a)+log(b)). Summing the logs of the predicted probability densities for all the RTs gives the log likelihood and minimizing minus the log likelihood produces the same parameter values as maximizing the likelihood.

Minus the log likelihood can be minimized using a variety of software routines and we use the robust SIMPLEX routine (Nelder & Mead, 1965). This routine takes starting values for each parameter, calculates the value of the function to be minimized, then changes the values of the parameters (usually one at a time) to reduce minus the log likelihood. This process is repeated until either the parameters do not change from one iteration to the next by more than some small amount or the value to be minimized does not change by more than some small amount.

For the chi-square method, the chi-square value is minimized using the SIMPLEX minimization routine, typically with RTs divided into either 5 or 9 quantiles. For 5 quantiles, the

data entered into the minimization routine for each experimental condition are the .1, .3, .5, .7, .9 quantile RTs for correct and error responses and the corresponding accuracy values. For 9 quantiles, the .1, .2, .3, ..., and .9 quantiles are used. The quantile RTs and parameter values of the model are used to generate the predicted cumulative probabilities 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 chi-square computation, these are the expected values, to be compared to the observed proportions of responses between the quantiles (i.e., for the 5-quantile method, the proportions between 0, .1, .3, .5, .7, .9, and 1.0, which are .1, .2, .2, .2, .2, and .1, and for the 9-quantile method, the proportions between quantiles and outside them, which are all .1). These proportions are multiplied by the number of observations to give the expected frequencies and summing over $(\text{Observed-Expected})^2/\text{Expected}$ for all conditions gives a single chi-square value to be minimized. The SIMPLEX routine then adjusts parameter values to minimize the chi-square value.

For the G^2 method, $G^2 = 2 \Sigma N p_i ln(p_i/\pi_i)$. This statistic is equal to twice the difference between the maximum possible log likelihood and the log likelihood predicted by the model (because $\ln(p/\pi)=\ln(p)-\ln(\pi)$). Every time we have used this method (with several hundred observations per subject) and compared results to the chi-square method, we have found almost identical parameter estimates. This is because the chi-square approximates the multinomial likelihood statistic (see Jeffreys, 1961, p. 197); both are distributed as a chi-square random variable. We fit all the data for each subject in Experiments 1 and 2 and found that boundary separation, nondecision time, and drift rates correlated between the two methods greater than 0.986 for Experiment 1 and greater than 0.958 for Experiment 2. The across-trial variability parameters correlated greater than 0.930 for Experiment 1 and greater than 0.855 for Experiment 2. These correlations show that for the relatively large numbers of observations in Experiments 1 and 2, the asymptotic properties of these estimators are equivalent. When we reduced the numbers of observations, for example to 40, there were some failures such that the two methods did not produce the same values, but we have not pursued this further.

We only applied the chi-square method to the data from Experiments 1 and 2 but we applied

all the methods to the simulation studies. We assumed that every observed RT distribution has contaminants and their probability is p_o, the same probability for all experimental conditions for a subject. The contaminants are assumed to come from a uniform distribution with a range determined by the maximum and the minimum RTs in each experimental condition so that the model fit is a mixture of contaminants and responses from the diffusion process. In generating simulated data from this mixture, contaminants are assumed to involve a delay in processing (but not random guessing). Thus the contaminant assumption in generating simulated data is not the same as in fitting the data. However, the assumption of a uniform distribution of contaminants in all conditions gave successful recovery of the parameters of the diffusion process and the proportion of contaminants (Ratcliff & Tuerlinckx, 2002). In a study of sleep deprivation, Ratcliff and Van Dongen (2009) found it necessary to assume contaminants involved random guessing. They assumed a uniform distribution of contaminants with random guessing in fitting data from a numerosity discrimination task under sleep deprivation.

Fitting the model with the chi-square and G-squared methods is much faster in terms of computer time than the MLH method, especially for large numbers of observations. For example, for Experiments 1 and 2, for each experimental condition, the 5-quantile chi-square required five evaluations of the diffusion model cumulative density for correct responses and five evaluations for error responses, no matter how many observations there were per condition. For the MLH method, the density function must be evaluated for each RT, which means hundreds or thousands of evaluations for each condition (with two evaluations of the cumulative distributions function to produce the density function for each RT). For the studies below, we used the chi-square method because it is what we have been using and because the results would be similar if we used, for example, G-square.

In addition to the chi-square and MLH methods, we tested four diffusion-model fitting packages that are available in the public domain: DMAT (Vandekerckhove & Tuerlinckx, 2007, 2008), fast-dm (Voss & Voss, 2007, 2008), HDDM (Wiecki, Sofer, & Frank, 2013), and EZ (Wagenmakers, et al., 2007). The DMAT, fast-dm, and HDDM packages can all fit all the conditions of an experiment and both correct and error responses simultaneously, but the EZ method fits only one condition at a time and only correct responses or only error responses. For

each method, we used the most straightforward default method and options. The aim was to reproduce what a user might employ in fitting. Simulated data used in the studies used here will be available as a supplement.

The data input to the DMAT package are the RTs and choice probabilities for each quantile of the data, where the number of quantiles is defined by the user (values of the bin limits can also be specified by the user). The values of the model parameters that best generate the quantile data are determined by minimizing a chi-square or

G^2 statistic. The package also allows the user to choose to implement a mixture model for slow contaminants and an exponentially weighted moving average method to eliminate fast outliers (Vandekerckhove & Tuerlinckx, 2007). In the applications below, DMAT was applied both with and without contaminant correction.

Note that DMAT was designed to be used with data with large numbers of observations per condition (hundreds) and was not tuned for smaller numbers and will likely not produce meaningful estimates for numbers of observations per condition in the tens. It also provides warning messages when there may be problems with the fit. Often these are ignored. We operate like the normal user and provide what the package produces, ignoring the warning messages. But then we report the number of them for one of the studies.

Fast-dm uses a Kolmogorov-Smirnov (KS) statistic in which the whole cumulative RT distributions for correct and error responses are generated and then compared with the cumulative for the data. The model parameters are adjusted until the deviation between the two is minimized. Fast-dm, instead of using the expressions in Equations 1-5, solves the partial differential equation (the Fokker-Planck backward equation, Ratcliff, 1978; Ratcliff & Smith, 2004) numerically, which is very fast. It might be possible to use numerical solutions like this for other packages or for the chi-square or MLH methods, but we have not investigated this (but see discussions by Diederich and Busemeyer, 2003, and Smith, 2000). The fast-dm method is robust to contaminants, as demonstrated later.

HDDM uses a Bayesian method that essentially combines a likelihood function with prior distributions over parameters. Boundary separation and nondecision time are Gamma distributed,

drift rate and starting point are normally distributed, across trial variability in drift rate and nondecision time are half normals, and across trial variability in starting point is beta distributed (see Wiecki et al., 2013). We used these informative priors in our fits. The package can fit each subject individually as for the other methods. We used default settings in HDDM, including a 20 sample burn-in, and 1800 samples for the estimation. The proportion of contaminants were estimated and not fixed in the model.

HDDM can also be fit using a hierarchical model in which model parameters are assumed to be drawn from distributions and the parameters of those distributions are estimated along with the parameters for each individual subject. This means that extreme values of parameters that might be produced because of noise in the data are constrained to be less extreme through the distributions over the group of subjects. For the first two simulation studies, we examined separate fits to individual subjects. We also compared parameters recovered from the hierarchical method with those recovered from the chi-square method. The HDDM package implements the same contaminant assumption as Ratcliff and Tuerlinckx (2002) except that the minimum RT is replaced by zero. This provides robustness to a few fast outliers. As for individual fits, we used a 20 sample burn-in, and 1800 samples for the estimation.

The EZ method (Wagenmakers et al., 2007) is based on a restricted diffusion model; there is no across-trial variability in any of the model parameters, the starting point is set midway between the two boundaries, there is no allowance for contaminant responses, and as mentioned above, it can be applied only to correct RTs or only to error RTs, for only one condition of an experiment at a time. Without across-trial variability, it cannot account for relations between correct and error RTs. With the starting point midway between the boundaries, it produces biased parameter estimates if the true starting point is not midway. It also predicts that RTs for correct and error responses will be the same, something that is known to be incorrect for the vast majority of experiments. Without allowance for contaminants, it produces quite biased estimates of parameter values (unless there are no contaminants in the data, which is unlikely; see Ratcliff, 2008).

Wagenmakers et al. (2007) derived expressions to relate the mean RT for a condition, the variability in the mean, and the accuracy for that condition to boundary settings, nondecision time, and drift rate for that condition. Essentially, this transforms three pieces of data into three model

parameters. This means that the model cannot be falsified on the basis of accuracy, mean RT, or variance in RT. The model does make predictions about RT distributions, the same predictions as the standard model and so they can be used to evaluate how well it fits RT distributions.

In our applications of the EZ method, we fit only correct responses. For error responses, the RT means and the variance in them are much more variable (because there are few observations) making parameter estimates less reliable than for correct responses.

van Ravenzwaaij and Oberauer (2009) compared the EZ method to the fast-dm and DMAT methods by generating simulated data and using these methods to fit the model to the simulated data. They found that EZ and DMAT were better at recovering parameter values and that EZ was the preferred method when the goal was to recover individual differences in parameter values. We extend their results by comparing these three methods with the other methods described above and by explicitly introducing contaminants into simulated data.

When faced with a very low number of observations either because existing data sets are being used or it is not possible to collect many observations per subject, it might be tempting to simply pool the observations (e.g., Menz, 2012) if group differences, not individual differences are the focus. This is an incorrect way of grouping data. This is easy to see: if there are only two subjects with distributions that are well separated, then the combination will be bimodal. A better way of combining subjects is to compute quantiles of the distribution and then average them (Ratcliff, 1979). Another way would be to use hierarchical modeling using the HDDM package.

One purpose of the analyses that we carried out for this article was to compare the publicly available fitting packages to home-grown fitting programs such as the one we use (which has been tweaked over the last 15 years). It is often tempting to hand tune home grown programs to the specific data set being fit, and in some cases this may be valid, but in comparisons among different fitting methods, it is important to avoid this. In the worst case, it is analogous to snooping through already-collected data to come up with hypotheses, which the publicly available packages do not allow. To check our home grown fitting programs for the applications in this article, the programs were handed off to the second author who implemented batch scripts for fitting all the data from simulated subjects with the same version of the fitting program (as for the packages). Thus, they

were not tuned to better fit any single data set as a function of an earlier fit.

Refinements of the Chi-square Fitting Method

Over the last several years, we have added four refinements to this method.

1. When the data are divided into 5 quantiles, if the number of observations in a quantile is less than 7 or greater than 1, then a median split is used to form two bins, each with probability .5. When the data are divided into 9 quantiles, the median split for two bins is used when the number of observations is less than 10 (but see qualifications in point 2 below). Numbers of observations as low as these occur mostly for error responses in high accuracy conditions in which there are few errors. If the number of observations is 1, we compute a single value of chi-square ($(O-E)^2/E$, where O is the observed number of observations, 1 in this case, and E is the expected value from the model) for that condition and add this to the sum of the rest of the values for all the conditions and quantiles.

In the DMAT program, if the number of observations (usually errors) is less than 11, then error quantiles are not used in fitting, and drift rates are estimated poorly. DMAT provides a warning when this happens and indicates the fitted value may not be valid. For the other methods, all individual RTs are used, that is, RTs are not binned.

2. When the number of observations in a quantile is less than 7 for 5 quantiles or less than 10 for 9 quantiles and the median is outside the range of the .3 to .7 quantiles for correct responses, then we do not use the median. Instead, we combine the bins into a single bin to produce a single value of chi-square as in point 1 (which of course does not use any information about RTs). We do this because when there are low numbers of errors in conditions with high accuracy, occasionally some of the error RTs can be spurious and (we assume) not from the decision process used in performing the task.

In contrast to binned methods, one extreme long or short RT can produce quite large biases in parameter estimates for methods that use every RT to produce a likelihood. As we show later, the fast-dm method with the KS statistics is robust to such short outliers.

3. Occasionally when there are few error responses, two consecutive error quantile RTs are

the same (e.g., 556 and 556 ms, with RTs with 1 ms resolution) and so the chi-square computation fails because the denominator of the chi-square is zero (because there will be zero probability mass between the two equal quantiles). Also, if quantile RTs are only 1 or 2 ms apart, the chi-square computation produces biased results because the probability mass between the quantiles is small. To address these problems, we added jitter to the raw RTs in the simulation studies described below by adding a uniformly distributed random number between -4 ms and +4 ms to each RT. This eliminated most of the problems that occurred when RT quantiles were too close together.

It might be thought that it is unrealistic to assume RTs are measured accurately to 1 ms. Generally, keyboards are polled and there can be systematic steps of 16 or 33 ms between successive polls of any key. We have measured delays up to 64 ms on older keyboards and found no problems because the variability in the keyboard times is combined (convolved) with the RT and the effect of the granularity in the keyboard response is very small. For example, if the SD in RT for a subject was 150 ms and the time between successive reads of a key was 32 ms, the SD in the keyboard would be 32/sqrt(12) ms (assuming a uniform distribution) so the SD for the combination (the square root of the sum of squares, i.e., sqrt(150^2+9.2^2)=150.3) would be 150.3 ms. Therefore, variability in the keyboard response times is not an issue as long as the SD is much lower than the SD in RTs.

4. Sometimes, accommodating very slow error responses leads to estimates of across-trial variability in drift rate larger than they should be. Thus, because accuracy values have to be fit, drift rates will be estimated to be larger. This means that drift rates covary with across-trial variability in drift rate parameters (e.g., Ratcliff & Tuerlinckx, 2002, Figure 6). An analog to this would be: If the SD in signal detection theory were increased, then to get the same hit rate, for example, the mean would have to be increased to compensate. In practice, with low numbers of observations in human subject data, sometimes long error RTs (some of which may be outliers) can cause the fitting program to produce values of across-trial variability in drift rate that are extremely large, and in compensation, this leads to large values of drift rate. Unless this is addressed, this can lead to drift rates several times larger than the values typically found with large numbers of observations (e.g., drift rates in the range 1.0 to 2.0 when the values should be in the range of .3 to .4).

This problem can be limited by placing upper and lower bounds on across-trial variability in drift rates (e.g., 0.3). The upper bound might be determined by examining the ranges of the variability parameter values from similar experiments with larger numbers of observations. Then upper and lower bounds can be set that are a little larger than the largest value from the subjects in the larger experiment. A lower bound can be set that is a little smaller than the smallest individual value. For fits for the simulation studies below with low numbers of observations, the value of the across trial variability in drift parameter was often estimated to be at the upper or lower bound for our programs that implemented these limits.

Experiments 1 and 2

The numerosity task for Experiment 1 and the lexical decision task for Experiment 2 were chosen because they are representative of commonly used experimental designs and because they are good candidates for tasks that can have practical applications such as diagnosing impaired subjects. Numerosity discrimination has been used to test math abilities for a variety of populations and lexical decision has been used in studies of aphasia and Alzheimer's disease. The design of the numerosity experiment is the same as that frequently used with perceptual tasks including brightness, letter, motion, number, and length discrimination (Ratcliff, in press; Ratcliff & Rouder, 1998; Ratcliff et al, 2001, 2003; Thapar, Ratcliff, & McKoon, 2003). The design of the lexical decision experiment is the same as that used with many memory tasks (e.g., Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar. Gomez, & McKoon, 2004; Ratcliff, Thapar & McKoon, 2004, 2010).

In Experiment 1, on each trial, a 10x10 array filled with spaces and between 31 and 70 asterisks was presented on a PC screen and subjects decided whether the number of asterisks was large (greater than 50) or small (less than 51). This has been used as a baseline task when comparing populations or individuals because the asterisks are displayed until a subject makes a response and so the task imposes no perceptual or memory limitations. The number of asterisks was varied from 31 to 70 to produce a range of levels of difficulty from very easy to very difficult.

In Experiment 2, on each trial, a string of letters was presented and subjects decided whether it was a word or not. There were three levels of difficulty for the words: words that occur

in English with high frequency, low frequency, or very low frequency. The difficulty of the nonwords was not varied.

For both experiments, we examined practice effects by grouping the trials into earlier versus later blocks. We also examined the number of trials that are needed to estimate the diffusion model's parameters with reasonably small SE's. Ideally, the estimated parameters for early trials would be consistent with those for later trials and the estimates from relatively small numbers of trials would be consistent with those from larger numbers. If so, the model could be used for the short times that are often necessary for measuring differences between an individual and a group (e.g., to look for cognitive deficits).

In the numerosity discrimination experiments, subjects started the task immediately after the instructions with no prior practice. In the lexical decision task, subjects were given 30 practice trials before starting the real experiment. This means that we can look at practice effects from the beginning of testing in the numerosity discrimination task, and after a very modest amount of practice in the lexical decision task.

Method

Subjects. Undergraduates at Ohio State University participated in the experiments for course credit, 63 in Experiment 1 and 61 in Experiment 2, each for one 45 min. session. For both experiments, stimuli were displayed on a PC screen and responses were collected from the keyboard.

Numerosity discrimination. On each trial, the asterisks were placed in random positions in the 10x10 array. Subjects were asked to press the "/" key if the number of displayed asterisks was larger than 50 and the "z" key if the number was smaller than 51 and they were asked to respond as quickly and accurately as possible. If a response was correct, the word "correct" was presented for 500 ms, the screen was cleared, and the next array of asterisks was presented after 400 ms. If a response was incorrect, the word "error" was displayed for 500 ms, the screen was cleared, and the next array of asterisks was presented 400 ms later. If a response was shorter than 280 ms, the words "TOO FAST" were presented for 500 ms after a correct response or after the error message for an incorrect response. There were 30 blocks of 40 trials each with all the numbers of asterisks between

31 and 70 presented once in each block.

Lexical Decision. Words were selected from a pool of 800 words with Kucera-Francis frequencies higher than 78, a pool of 800 words with frequencies of 4 and 5, and a pool of 741 words with frequencies 0 or 1 (Ratcliff, Gomez, & McKoon, 2004). There was a pool of 2341 nonwords, all pronounceable in English. There were 70 blocks of trials with each block containing 30 letter strings: 5 high frequency words, 5 low frequency words, 5 very low frequency words, and 15 nonwords. Subjects were asked to respond quickly and accurately, pressing the "/" key if a letter string was a word and the 'z' key if it was not. Correct responses were followed by a 150 ms blank screen and then the next response. Incorrect responses were followed by "ERROR" for 750 ms, a blank screen for 150 ms, and then the next test item.

Results for Experiment 1

Responses shorter than 250 ms and longer than 3500 ms were excluded from the analyses (less than 0.8% of the data). The data for numbers of asterisks less than 51 were collapsed with numbers greater than 50 and they were then grouped into two conditions, easy (numbers 31-40 and 61 to 70) and difficult (41-50 and 51-60. Averaging over all the trials of the experiment, accuracy for the easy condition was .892 (for individual subjects, the highest accuracy was .98 and the lowest .69), mean RT for correct responses was 627 ms, and mean RT for errors was 687 ms. For the difficult condition, accuracy was .676 (for individual subjects, the highest accuracy was .78 and the lowest .57), mean RT for correct responses was 664 ms, and mean RT for errors was 721 ms.

To examine practice effects and the number of trials needed for the diffusion model's parameters to have small SD's, eight groups of data were constructed: trials 1-80, trials 81-160, trials 161-240, trials 1-160, trials 161-320, trials 1-320, trials 321-640, and trials 1-1200. We fit the model to the data with the chi-square method with 9 quantiles. The model fit well, as we detail after presenting results for this experiment and for Experiment 2.

For each of the eight groups of data, Figure 1 plots the mean values over subjects of the parameters that best fit the accuracy and RT data. The means are shown for nondecision time, boundary separation, and two drift rates, one for the easy condition and one for the difficult condition. The wider error bars represent 2 SDs from the mean and the narrower ones, 2 SEs. The

means and SDs are also shown in Table 1.

When the values of the parameters estimated from the first 80 trials (the 1-80 group) were compared to the values estimated from all 1200 trials (the 1-1200 group), there were modest differences. The estimated values from all the trials were only slightly lower than for the early trials for boundary separation, nondecision time, and drift rate for the easy condition. For boundary separation, nondecision time, and drift rates, the SDs and SEs were smaller by one half to two thirds for all the trials than for 80 trials. Results were similar for the 1-160 group and the 1-320 group but with smaller differences. In general, for this subject population (undergraduates) and this task, there is little difference in the parameter values estimated from the first few trials and those estimated from the whole session.

Figure 1 and Table 1 also show further divisions of the data that allow examination of practice effects over the first few blocks of trials. There were small declines in nondecision time and boundary separation from the first block to later blocks. Drift rates for the easy condition were higher for the 81-160 and 161-240 groups than for the 1-80 and 1-160 groups respectively, but these were probably spurious because with lower numbers of observations, there are very few errors and this leads to inflated drift rate estimates. As the amount of data increases, the number of errors increases and so drift rate estimates decrease because the larger numbers of errors allows it to be estimated with more accuracy. The overall decrease in SE's across the groups also reflects increasingly larger numbers of errors. These same trends occur for the difficult condition but with smaller differences.

The SDs across subjects in the estimated parameter values (the larger error bars in Figures 1 and 2) allow examination of power for detecting differences between an individual and our population of subjects. College students are likely to provide the best performance of any population of subjects because they are likely to have the shortest nondecision times and highest drift rates (and perhaps the closest boundary settings, although these vary with task and with instructions that emphasize speed over accuracy or accuracy over speed). To identify an individual as different from our student population, his or her value for any of the parameters would have to lie at least one SD outside the 2 SD confidence interval for the students (e.g., Cumming & Finch, 2005). One SD outside a 2 SD confidence interval gives about 6% false negatives and 6% false

positives. One SD outside 2 SD's for nondecision time is about 500 ms and one SD outside 2 SD for boundary separation is about 0.18. Values of boundary separation and nondecision time larger than these values are often found with older adults which means it is possible to detect age differences through the model parameters. However, it is unlikely that subjects with deficits could be distinguished from the students on the basis of drift rates. This is because the bottom of the two SD range extends to zero or almost to zero and 1 SD lower than this is below zero (i.e., even a drift rate of zero representing chance performance would not be far enough outside the confidence intervals).

In contrast, the model parameters can be used to detect differences between different subject groups. Because standard errors decrease and hence power increases with the number of observations (subjects), even quite small differences can be discriminated. For example, 2 SE's in drift rates in Figure 1 for the easy condition with 1200 observations is about 0.02. This means differences as small at 0.03 could be detected for the population, number of observations, and conditions in Experiment 1 (the SE's can be found from the SDs in Tables 1 and 2 by dividing the SD by the square root of the number of subjects). Thus, there is power to detect even small differences in parameter values between the different groups.

Results for Experiment 2

RTs shorter than 300 ms and longer than 4000 ms were excluded (about 2.7% of the data) from data analyses.

For the data for all the trials, for the high-, low-, very-low frequency words, and nonwords, accuracy was .950, .842, .721, and .893 respectively, mean RTs for correct responses were 609, 714, 767, and 730 ms respectively, and mean RTs for errors were 600, 735, 777, and 792 ms respectively. The minimum and maximum values of accuracy across subjects were 1.0 and .88, .96 and .57, .88 and ,45, and .98 and .79 for the high-, low-, very-low frequency words, and nonwords respectively.

To examine practice effects and the number of trials needed for the diffusion model's parameters to have small SD's, eight groups of data were constructed: trials 1-120, 121-240, 241-360, 1-240, 241-480, 1-480, 481-960, and 1-2100. Figure 2 shows the means across subjects of the

best-fitting parameter values. The wider error bars represent 2 SDs from the mean and the narrower ones, 2 SEs. The means and SD's are also given in Table 2.

Compared to Experiment 1, there is more variability in the estimates of the parameters and this is true even though there were more observations. There are two reasons for this: a larger range of individual differences and more sampling variability because the starting point was estimated from the data; it was not set at $z=a/2$ as it was for Experiment 1.

There were only modest differences between the estimates for the first 120 trials and the estimates from all 2100 trials, mirroring the results from Experiment 1. The estimates from all the trials were only slightly lower than for the early trials for boundary separation and nondecision time, and the estimates were a little higher for drift rates for the low and very low frequency words and nonwords. The estimate for high frequency words was considerably higher due to no error responses for many of the subjects in the first few trials in that condition (leading to spuriously higher estimates). For all six parameters, the SDs and SEs were smaller with all the trials by up to half the value. The pattern of results was similar for all the parameters for the 1-240 group and the 1-480 group, but with reduced differences and smaller SDs and SEs relative to the 1-120 group.

The trends in model parameters as a function of practice (Figure 2 and Table 2) are similar to those from Experiment 1 but with slightly larger differences. Results show that there were small declines in nondecision time and boundary separation from the first block to later blocks. Drift rates for the high-frequency condition are higher for the 1-120, 121-240, and 241-360 groups and are higher than the drift rate for all the data. Also, the drift rates for the 121-240, and 241-360 low frequency groups are higher than the drift rate for all the data. For each of these conditions, 2 SE error bars do not overlap with those for all the data. This is because there are relatively few errors (and zero for some subjects) in these conditions, i.e., for a group with 120 observations, 60 are from word categories and of these, 20 are from each of the three frequency classes and so with a 5% error rate, there will often be zero errors. As discussed earlier, this leads to inflated estimates of drift rates. Two SE error bars for nondecision times and boundary separation for the 1-120, 121-240, and 241-360 groups do not overlap with the 2 SE error bars for all the data (trials 1-2100). But generally, the practice effects are relatively small (especially compared to individual differences).

Following the discussion for Experiment 1, the results of this experiment show that an individual subject could be identified as having a deficit relative to our undergraduate subjects if boundary separation was 1 SD above the 2 SD confidence limit, which, for parameters from fits to all the data, is 0.21. For nondecision time, the value would be 530 ms. Two SD's below the means for drift rate were near zero for all conditions except for high frequency words. Therefore, differences between an individual and the undergraduate group could not be detected unless his or her drift rates were near zero (and there was a relatively large number of trials).

In contrast to detecting differences between an individual and the undergraduates, the SE's on the model parameters have enough power to detect even quite small differences between groups of subjects just as in Experiment 1.

Last blocks of trials. We also examined parameters from fits of the model to the last three blocks of trials, blocks of 80 trials for the numerosity experiment and 120 trials for the lexical decision experiment. This was done as a check to see if there were practice or fatigue effects at the end of the sessions. Boundary separation and nondecision times were a little larger than those for the fit to all the data (by less than .005 and 20 ms respectively) and drift rates were similar to the fits for the first three trial groups in Tables 1 and 2 (i.e., a little larger than those for fits to all the data). These results show that there are no dramatic differences between model parameters estimated from the last few blocks of trials and the first few blocks of trials.

The results of these two experiments are consistent with other studies that have used the diffusion model to examine practice effects. Petrov, Van Horn, and Ratcliff (2011), Ratcliff, Thapar, and McKoon (2006), and Dutilh, Vandekerckhove, Tuerlinckx, and Wagenmakers (2009) all found that boundary separation becomes smaller and nondecision time shorter with increasing amounts of practice but that there is little change in drift rates. These changes are largest between sessions.

Across-trial variability parameters. For both experiments, the across-trial variability parameters were relatively poorly estimated (Ratcliff & Tuerlinckx, 2002). The means of across-trial variability in drift rate, starting point, and nondecision time minus 2 SD's either include zero or are close to it. In fitting the model to the data for Experiments 1 and 2, we constrained the across-

trial variability in drift rate to be in the range that has been observed in other applications of the model to similar experiments. If it were not constrained in this way, one SD in across-trial variability in drift rate would be 2/3 the drift rate. Generally, differences in the across-trial variability parameters between individuals or between populations cannot be determined without extremely large numbers of observations.

Goodness of fit. Tables 1 and 2 show chi-square goodness-of-fit values for Experiments 1 and 2 using the chi-square method with 9 bins per distribution. Chi-square goodness of fit values are often used to assess how well diffusion models (and other two-choice models) fit data. Because the bins are determined by the data, that is, by the values of the RT quantiles, the statistic we calculate is not, strictly speaking, a chi-square. However, the statistic approaches a chi-square asymptotically (e.g., Jeffreys, 1961), and when the standard chi-square is compared to the chi-square based on quantiles, little difference is found between them (Fific, Little, Nosofsky, 2010).

In fitting the model to data, there are two constraints that a fitting method tries to satisfy. First, it needs to adjust the model parameters to adjust proportions (probability mass) across the bins between quantiles within each condition to match the proportions between the quantile RTs in the data. The second is to adjust parameters to move proportions so that the proportion of correct responses and the proportion of error responses match each other.

For each of our data sets, using 5 quantile RTs, there were 12 bins (6 for correct responses and 6 for errors) in each experimental condition and the total probability mass in each condition summed to 1.0, reducing the number of degrees of freedom to 11. For a total of k experimental conditions and a model with m parameters, the number of degrees of freedom in the fit was therefore $df = k(12 - 1) - m$. For Experiment 1, the number of degrees of freedom was 14 and for Experiment 2, the number was 33. With 9 quantiles, there are 20 bins with 19 degrees of freedom per condition. Thus, for Experiment 1, the number of degrees of freedom was 30 and for Experiment 2, the number was 65.

Ratcliff, Thapar, Gomez, and McKoon (2004) examined the effect of moving a .1 probability mass from one quantile bin (so the .2 probability mass became .1) to another adjacent quantile bin (so the .2 probability mass became .3), i.e., the first constraint in the prior paragraph.

They found that the increment to the chi-square was 0.133*N, so for N=100, the increment would be 13.3 and for N=1000, the increment would be 133. Thus, for example, for two conditions and 5 quantiles with 13 degrees of freedom (22 from data minus 9 model parameters), the change in chi square would be over half the critical value (of 22.4) for N=100 and five times the critical value of N=1000. These increments mean that even relatively small systematic misses in the proportions are accompanied by quite large increases in the chi-square. Furthermore, the chi-square statistic is a very conservative statistic so that even small systematic differences between theory and data show large increases in the chi-square as the number of observations increases.

The mean values of the chi-square for the lowest numbers of observations for the two experiments (the first three lines in Tables 1 and 2) are smaller than the mean chi-square from the chi-square distribution. If the data were generated from a chi-square distribution, the mean chi-square would be the number of degrees of freedom. This means that the model is overfitting the data, i.e., the model is producing fits that are accommodating random variations in the data from variability due to small numbers of observations. For all the data from Experiments 1 and 2, the mean values of the chi-square are 50 and 158 respectively, with critical values of 44 and 85. These represent a better estimate of how well the model is fitting the data because the number of observations is large and variations in the data that occur with small numbers of observations are minimized with a thousand observations or more as in these fits. For many data sets from a number of experiments, we have found that with numbers of observations per subject like the ones in the experiments here, the mean values of chi-square over subjects are typically between the critical value and twice the critical value. Thus, the quality of the fits for Experiments 1 and 2 are about the same as for previous experiments (e.g., Ratcliff, Thapar, & McKoon, 2003, 2004, 2010, 2011).

Power analyses for Experiments 1 and 2. Table 3 shows simple power analysis calculations using the means of the parameter values estimated from the data. We used SDs rounded up and down based on those from Tables 1 and 2, values that would correspond to a moderate number of observations, around several hundred. We assumed normal distributions of the populations (the distributions of parameter values across individuals are usually reasonably symmetric, e.g., Ratcliff et al., 2010).

To perform a power analysis, we needed an alternative hypothesis. We assumed another

population (e.g., for which the subjects had deficits) and estimated the value of the mean to obtain 90% and 95% correct classification of individuals. Boundary separations were assumed to be larger, their nondecision times were assumed to be longer, and drift rates were assumed to be lower. In each case, the SD in the model parameters for this population was assumed to be the same as for the non-deficit population (i.e., the undergraduates in our experiments).

Given the means for the non-deficit population and the SDs for both, we found means for the deficit population so that a score selected from either distribution would be classified correctly 90% of the time and another set of means that would produce a 95% correct classification. Results are shown in Table 3 columns 5 and 6. We also did the same analyses with SD's 1.5 times larger than those in column 4 of Table 3 and these are shown in columns 7 and 8.

Results showed that for boundary separation and nondecision time, the differences between the values for the deficit population and the undergraduate population for 90 and 95% classification accuracy were large, but only as large as differences that have been found in previous studies with older adults. This means that these parameters are in the range that might be useful for classifying individuals. For example, for 90% correct classification with the smaller of the two SDs, the means for nondecision time and boundary separation are about the same as those for older adults (Ratcliff et al., 2001, 2004, 2010).

However, for drift rates, the classification would be much more difficult. Few of the conditions had drift rates that would separate the population with deficits from the undergraduate population. For the easy condition in numerosity and the low- and very-low frequency word conditions in lexical decision, the values of the drift rates to provide correct classification were negative, and so these conditions are not shown in Table 3.

For drift rates for the difficult condition in the numeracy design and high-frequency words and nonwords in the lexical decision design, the drift rates to achieve 90% correct classification were low enough that performance would be near, but not quite chance (except for high frequency words in lexical decision).

But for 95% correct classification and for the larger values of the SD in parameters across subjects, performance would be near chance in a Condition to detect a deficit. This suggests that

the range of individual differences in drift rates in these tasks is so large that individuals with a deficit would have a large overlap with the normal range. This analysis is based on each parameter separately and shows that drift rates in single conditions are likely not to be useful detecting deficits in these tasks and designs. However, it may be possible to use multivariate methods to improve classification with combinations of several parameters (e.g., drift rates, boundary separation, and nondecision time) and if subjects were tested on multiple tasks, combinations of measures across tasks.

Correlations among Parameters

If estimates of the model's parameters from small numbers of observations correlate positively with estimates from large numbers, then small numbers can be used to examine individual differences, for example, whether measures such as IQ are correlated with some model parameter or a combination of parameters. Even if the estimates from small numbers are biased, consistency among them (as measured by correlations) would be sufficient for many purposes such as examining whether a depression measure, a reading measure, an IQ measure, etc. was related to model parameters.

The correlations in Table 4 show the consistency of parameter estimates across the various groups of trials and numbers of observations for Experiments 1 and 2. For each parameter, the table shows the correlations between that parameter as estimated for the different groups of trials and that parameter as estimated from all the data.

As would be expected, the correlations increase as the number of trials increases. For boundary separation and nondecision time, the correlations are strong even with smaller numbers of trials, mostly above .5 for both tasks (except the 1-80 group for numerosity).

The correlations for drift rates are lower, ranging from around .25 for smaller numbers of observations to over .5 for larger groups for Experiment 1 and from around .35 for smaller numbers of observations to over .7 for the larger groups for Experiment 2. There was one unexpected result: in Experiment 2, correlations for drift rates for nonwords are below those for low- and very-low-frequency words even though there were more observations for the nonwords. We have no explanation for this.

The conclusion from these correlations is that consistency in parameter values is good from small to large numbers of observations for boundary separation and nondecision time but not drift rates. Thus, if the aim is to examine differences among individuals or populations, this can be done for nondecision time and boundary separation with smaller numbers but more observations would be needed for drift rates.

Summary for Experiments 1 and 2

The data from Experiments 1 and 2 show that there is enough power in the numbers of observations provided by a 20- or 25-min. session of around 200-300 trials to give estimates of boundary separation and nondecision time that are sufficiently precise to detect differences between an individual subject and a population of subjects. However, for drift rates, the ranges of individual differences are too large to provide enough power to detect discrepant individuals unless their performance has fallen to near floor.

For detecting differences between populations of subjects, the data from a 20- or 25-min. session of around 200-300 trials is sufficient to distinguish the populations in boundary separation and nondecision time but more observations would be needed for drift rates.

Simulation Study 1

For this study, we generated simulated data for 64 subjects and evaluated how well each of the eight methods for fitting the diffusion model reproduced the ordering of the values of the parameters across subjects. If the fitted values are strongly correlated with the generating values, then the fitted values can be used to investigate correlations between parameters of the model and subject variables such as age or IQ. The parameters that have been of most interest for correlations are drift rates, boundary separation, and nondecision time. Generally, because the across-trial variability parameters are not well estimated, significant differences in them between individuals or groups are not obtained.

We simulated the data using the values of the diffusion model's parameters that best fit the data from Experiments 1 and 2. For Experiment 1, numerosity, the simulations were performed with 40, 100, and 1000 observations for each condition (easy or difficult). For Experiment 2, lexical decision, the simulations were performed either with 20 observations for each of the word

conditions (high-, low-, and very-low-frequency) and 60 for the nonwords, or 200 for each of the word conditions and 600 for nonwords.

The values of the drift rates for the conditions were correlated (if a subject performed well in one condition, they also performed well on the others). For the numeracy design, a random number was added to the easy condition drift rate and half the same random number was added to the smaller drift rate for each subject. For the lexical decision design, the same random number was added to each drift rate.

For each of the numbers of observations, data were simulated for the 64 subjects, once with 4% contaminant RTs and once without contaminants. For each subject, the value of each parameter was drawn randomly from a normal distribution for which the mean and SD across subjects (bottom of Tables 1 and 2) were rounded versions of those from Experiments 1 and 2. The contaminant RTs were obtained by adding a delay randomly selected from a uniform distribution with range 2000 ms (see Ratcliff & Tuerlinckx, 2002).

For the model to be fit successfully, it is best to have at least one condition with enough errors to provide a modest estimate of the RT distribution for errors. For numerosity, the drift rate for the difficult condition was low enough that there were enough errors (usually greater than 6) to constrain fitting the model. Similarly, for lexical decision, the drift rates were low enough for the low- and very-low frequency words to provide enough errors to constrain the model fitting.

For some of the 64 simulated subjects, the combinations of parameter values were not like those typically observed in practice. For example, the amount of across-trial variability in drift rate is not small when drift rates are large and across-trial variability in nondecision time is not large when nondecision time is small. However, we did not try to assess the plausibility of the various combinations; instead we let them vary independently to provide a wide range of combinations.

The metric for evaluating the fitting methods for this study was the correlation between the recovered parameter values and the parameter values that were used to generate the simulated data. We focus on the boundary separation, nondecision time, and drift rate parameters, parameters that have been used in understanding the effects of, for example, aging, development, and sleep deprivation, on performance.

Results

For each of the eight fitting methods, Tables 5 and 6 show the correlations between the recovered values and the values used to generate the data for each simulated subject. For the numerosity design, Table 5 shows correlations for drift rates for the easy and difficult conditions for 1000, 100, and 40 observations per condition for 0% and 4% contaminants. Table 6 shows them for lexical decision for high-, low-, and very-low frequency words and nonwords, for 200 observations per word condition and 600 for nonwords, and for 20 observations per word condition and 60 for nonwords, each with and without contaminants.

Because the EZ method can fit only a single condition at a time, we fit it separately for each condition of each experiment and then averaged the values of boundary separation and nondecision time.

The numeracy design. For the simulation with 1000 observations per condition and no contaminant RTs, all of the fitting methods produced parameter values that were highly correlated with the generating values, above .9. With only 40 or 100 observations, fast-dm, MLH, the two chi-square methods, and EZ had correlations above .6, but HDDM and DMAT did considerably worse, with correlations dropping nearly to zero when there were only 40 observations per condition. The low DMAT correlations for low numbers of observations are expected because DMAT does not use error RTs to constrain parameter estimates when the number of errors is less than 11.

With 4% contaminant RTs and 1000 observations per condition, the two chi-square methods and the MLH method produced correlations of greater than .83 for all four parameters, fast-dm produced correlations above .78, and EZ produced correlations above .76. DMAT (with and without correction for contaminants) produced correlations above .86 for boundary separation and nondecision time, but the correlations for drift rates dropped substantially to around .6. HDDM produced correlations for boundary separation of .67, for nondecision time .94, and drift rates, .60 and .75. In each of these cases, the correlations were lower than when there were no contaminants.

With 4% contaminant RTs, for 100 and 40 observations per condition, the two chi-square methods and the MLH method produced correlations greater than .63, fast-dm produced correlations above .54, and EZ produced correlations above .57. For HDDM and the two versions

of DMAT some correlations were near zero.

To summarize, the MLH method, the two chi-square methods, and fast-dm were a little superior to the EZ method because they produced higher correlations between recovered parameter values and values used to generate simulated data across all combinations of numbers of observations and did so whether there were contaminants or not. These methods all produced higher correlations than the two DMAT versions and HDDM, each of which produced some correlations near zero.

The lexical decision design. For the most part, the correlation results mirror those for the numeracy design but there is one large difference: HDDM's correlations were competitive and often exceeded the correlations for the chi-square and MLH methods.

For most of the methods, recovered drift rates for the high-frequency words were high even with the largest number of observations (200) per condition and contaminants because there were often no errors. As a result, the recovered values were more variable than for the drift rates for low- and very low-frequency words and nonwords and this led to correlations lower than for the drift rates for the other conditions. In the discussion that follows, we exclude drift rates for v_1, the high-frequency word condition.

With 200 and 600 observations per condition and no contaminant RTs, all of the methods produced parameter values that were highly correlated, above .75, with the generating values. Some of the drift rate correlations were lower than for the numeracy simulations because they used 1000 observations per condition (as opposed to either 200 or 600 used here). With 20 and 60 observations, HDDM, fast-dm, MLH, and the two chi-square methods produced correlations above .5. DMAT did considerably worse, with correlations dropping to the .1-.3 range. Correlations for the EZ were as low as .32.

With 4% contaminant RTs, for 200 and 600 observations per condition, HDDM, fast-dm, DMAT without contaminant correction, MLH, the two chi-square methods, and the EZ method all produced correlations greater than .72. DMAT with correction for contaminants gave lower correlations. The EZ method performed reasonably well, with correlations above .76.

With 4% contaminant RTs and 20 and 60 observations per condition, HDDM, fast-dm, the

two chi-square methods, and the MLH method produced correlations greater than .53. The EZ method's correlations were as low as .26 and the two DMAT methods' were as low as .14.

The results of this study are generally the same as for the numeracy design mirror, but in this lexical decision design, HDDM largely outperformed the other methods.

Fixing across-trial variability parameters. The EZ diffusion method assumes no across-trial variability in model parameters. This would be equivalent to fixing these across-trial variability parameters at zero in the other model fitting methods. By fixing across-trial variability parameters, the model might avoid some of the instability produced when model parameters trade off and become extreme when the number of observations is small. This may be part of the reason the EZ method performs quite well in the simulations presented above and in van Ravenzwaaij and Oberauer (2009).

Rather than fixing these across-trial variability parameters at zero, we fit the model to group data and then fixed these parameters at the values from the group fits for the two chi-square methods for all the combinations used in the numeracy and lexical decision designs as in Tables 5 and 6. We found that in almost every case, the correlations for boundary separation, nondecision time, and drift rates were lower than for the chi-square method with the range of parameter values restricted but free to vary over a plausible range. Thus, fixing the across-trial variability parameter values at their group means did not improve recovery of the parameter values. (We see the same issue with the hierarchical diffusion model in which across trial variability parameters are set to the same value across subjects).

Summary. Overall, the MLH, both chi-square methods, and fast-dm methods were robust to contaminants in the data and low numbers of observations. The correlations between recovered parameter values and generating values were moderate to high.

EZ performed somewhat worse, especially with contaminants (Ratcliff, 2008). It would perform more poorly if some subjects had contaminants and others did not. However, the correlations in Tables 5 and 6 are not that much worse than for the other methods, and so we conclude that EZ might serve as a useful exploratory tool.

DMAT did not perform as well as MLH, both chi-square methods, fast-dm, and EZ. One

reason is that it does use error responses when the number of observations is less than 11. If the tricks outlined in the earlier section were implemented, (e.g., using the median RT to construct two bins when the number of observation is low, excluding error conditions when the data appear spurious, and restricting the ranges of across-trial variability parameters), DMAT without contaminant correction for fast contaminants but with the correction for slow contaminants should perform equivalently to the two chi-square methods.

One striking difference between the results of the numeracy and lexical decision experiments is the behavior of the HDDM method. For numerosity, HDDM performed relatively poorly but in lexical decision, it performed at the top of the list. This may be because of the additional constraint imposed by having four conditions (i.e., four drift rates) in the lexical decision design instead of the two for the numerosity design.

Finally, Figures 3 and 4 show sample correlations for two fitting methods, the chi-square method and the EZ method. The recovered values are plotted against the generating values for the numeracy design with 40 and 1000 observations per condition. We chose the examples presented in Figures 3 and 4 to illustrate what correlations of the sizes presented in Tables 5 and 6 look like and to illustrate biased parameter recovery but with a high correlation (e.g., the EZ method with 4% contaminants N=1000).

Simulation Study 2

In Simulation Study 1, we examined whether the eight fitting methods produced the same ordering of parameter values across subjects as the ordering that were used to generate the simulated data and we did this irrespective of whether the fitted values were biased away from the generating values. In this study, we looked at their accuracy. For each set of parameter values, described below, we generated 64 sets of simulated data and fit the model to them for each of the methods.

The first question was whether and how much the means of the recovered values deviate from the generating values. If the bias for one parameter is consistent across combinations of all the other parameters, then there are few if any negative consequences (e.g., it would not matter if all the boundary separation parameters were estimated to be 80% of the true value; they would all

line up the in the same way). If the bias varies as a function of the values of the other parameters, then it is necessary to assess how large the degree of bias is and do so for different numbers of observations. If bias is reduced as the number of observations increases, then the number of observations appropriate for a specific application must be determined.

The second question was how tightly the recovered parameter values cluster around their mean, that is, what are their SD's. It is especially important to examine this when the number of observations is low in order to detect whether there is sufficient power to test empirical hypotheses.

If the mean of the recovered values is unbiased, then the better estimation methods are those that produce smaller SD's. When the mean is biased, its SD might be smaller than when the mean is unbiased. Sometimes an application may be better served by a biased mean with a smaller SD than an unbiased mean with a larger SD or vice versa (unless the bias changed with the value of the parameter as we see for the EZ method below).

To examine bias and variability in recovered parameters, we simulated data for the numeracy and lexical decision designs using the values of the parameters shown in Table 7. These are representative of the ranges of mean values over subjects that were obtained in Experiments 1 and 2 and in other published sets of data (e.g., Matzke & Wagenmakers, 2009; Ratcliff, 2013). For boundary separation, there were two values: 0.1 is a little more than might be obtained under speed instructions (0.08 would be the smallest) and 0.2 is a little less than might be obtained for older adults (0.25 might be the largest). Nondecision time and across-trial variability in it were held constant across the simulations because nondecision time is simply an additive constant to RTs and variability in nondecision time has effects mainly on the leading edge of the RT distribution (Ratcliff, Gomez, & McKoon, 2004).

Across-trial variability in drift rate had two values and across-trial variability in starting point had two or three values (the larger value differed depending on the boundary separation). The proportion of contaminants was set at zero or .04.

For numeracy, the drift rate for the easy condition was (v=0.2) and for the difficult condition it was (v=0.1). The larger value of drift rate was lower than that in the experimental data in order to produce errors in most of the combinations of parameter values. There were two values

of across-trial variability in drift rate, η=0.1 and 0.2. In this design, the starting point is symmetric between the two conditions because large responses to large stimuli are grouped with small responses to small stimuli (thus $z=a/2$). The range of across-trial variability in starting point is limited by the boundaries. When a was 0.1, the values of s_z were 0.02 and 0.06 and when a was 0.2, the values were 0.06 and 0.08. Altogether, there were 16 combinations of parameter values.

For lexical decision, there were three conditions: one with a relatively high value of drift rate (v=0.2, high-frequency words), one with a lower value (v=0.1 low-frequency words), and one with a negative value (v=-0.2, nonwords). (We did not use a fourth condition as in Experiment 2 and Simulation Study 1 in order to reduce the time it took to fit the simulated data; for some methods, this saved a day of computation time.) In this design, the starting point, z, is a free parameter so it is fit along with the other parameters. In some experiments, the proportion of choices is manipulated and this produces a bias in starting point (Leite & Ratcliff, 2011; Ratcliff, 1985). To represent such manipulations, we used three values of z: halfway between the boundaries, z/a=.5, closer to one of the boundaries z/a=.3, and closer to the other boundary, z/a=.7. Across-trial variability in starting point was the same as for the numeracy simulation, 0.02 and 0.06 when a was 0.1 and 0.06 and 0.08 when a was 0.2. Boundary separation, across-trial variability in drift rate, and proportion of contaminants had the same values as in the numeracy simulation. The total number of combinations was 32.

For each of the 16 combinations of parameter values for the numeracy study, 64 sets of simulated data were generated with 40 observations per condition, 64 sets with 100 observations, and 64 sets with 1000 observations. One thousand per condition would be about the number for a 50 min. session of data collection.

Sixty four sets of data were also generated for each of the 32 combinations for the lexical decision study, once with 30 observations for the first two conditions and 60 for the third, and once with 300 observations for the first two conditions and 600 for the third. A total of 1200 observations is about the number for a 35 min. session.

Our implementation of the MLH method was not efficient and very slow (it would have taken several weeks for 64 sets of fits with large numbers of observations in the two studies). To

rewrite it to be efficient would have taken considerable effort. Also, parameter recovery differed little between the chi-square method and the MLH method in the first simulation study, so the results for this study should be similar to those for Study 1. For these reasons, we did not test the MLH method in this study.

Results

Tables of the means of the best-fitting parameters and their SD's for all 48 combinations of parameter values and all seven fitting methods are given in the online supplement. In the results presented here, we did not separate contaminant versus non-contaminant conditions. When there were systematic differences between them, we noted it in the discussion.

Figures 5 to 8 show, in a concise way, how well the fitting methods recovered the values of the generating parameters. Figure 5 shows the results for 1000 observations per condition for the 16 parameter combinations for the numeracy design and Figure 7 shows them for 40 observations per condition. Figure 6 shows the results for the 300, 300, and 600 observations per condition for the 32 combinations for the lexical decision design and Figure 8 shows them for 30, 30, and 60 observations.

Each panel in the figures shows the results for one parameter for one fitting method. The x- and y-axes represent the parameter's possible values and the vertical line on the x-axis is the value used for generating the data. The circles plot the means of the recovered values across the 64 sets of simulated data, one mean for each of the 16 or 32 combinations of the generating parameters. The means form a horizontal line at the generating value and one SD distance above and below the line is shown by the error bars.

Plotting the results in this way makes the results easy to grasp visually, a necessity because there is a total of 74 panels in the four figures. Consider two examples.

First, for the values of v_1 and v_2 (the numeracy design) recovered by the 9-quantile chi-square method (the top right panel in Figure 5), the vertical lines show the generating values, 0.1 and 0.2. For both, the 16 dots, one for each combination of parameters, lie almost entirely on top of each other, which means that they vary little across the combinations of parameters. They also lie on top of the generating value, which means that they are not systematically biased away from

it. The variability in each of the 16 means (one SD error bars in the y-direction) is small; the one SD error bars cluster tightly around their means.

Second, consider the HDDM method four panels below in Figure 5. Here many of the values are clustered more tightly around the generating mean than for the chi-square example but there are a few outliers that lie above the generating values (shifted to the right) and these have quite large standard deviations (large error bars). Thus, for a few parameter combinations, HDDM has problems.

To preview the results, the chi-square methods and fast-dm did well, with modest biases with low numbers of observations. HDDM did well except that in a few cases, there were large deviations from the values used to generate the simulated data. The EZ method produced biased values of parameters, especially when there were contaminants. With large numbers of observations, DMAT without contaminant correction produced relatively unbiased values, but with SD's larger than the best methods. DMAT performed poorly relative to the other methods with contaminant correction and also with small numbers of observations.

Large numbers of observations for the numeracy design (1000 per condition). The chi-square methods did a good job (relative to the other methods) of recovering parameter values. The recovered values are clustered tightly around the generating values except for a slight bias toward over-estimation of *a*.

The fast-dm method did a little worse than the chi-square methods. Some of the SD's were larger than the chi-squares' and all of the recovered values (except for the smaller value of *a*) were biased toward smaller values than the generating values. Bias was especially large for the drift rates. Inspection of the values used to generate the figure showed that the biases were larger when across-trial variability was the larger of its two values.

The results for HDDM were mixed. For many of the parameter combinations the recovered values were closer to the generating values and had smaller SD's than the chi-square and fast-dm methods. But for the small value of boundary separation, for both it and the two drift rates, there were some large deviations from the values used to generate the simulated data. Some values were much larger than the true values, some double the true value, and their SD's were large. These

biases and SD's occurred mainly when there were 4% contaminants. These deviations are not due to just a few extreme values, but they occur because many of the values for fits to the 64 simulated data sets miss.

After the first draft of this article, we contacted Thomas Wiecki about the anomalous results for this set of data (with 1000 observations per condition). He suggested setting the proportion of contaminants to a fixed 0.05 proportion. Then he suggested using the fixed 0.05 proportion of contaminants plus a longer burn-in (1500 trials). This improved the parameter recovery as is shown in Figure A1, but there are still biases in parameter estimates. These biases occur in the cases in which there are no contaminants in the simulated data. The larger boundary separation is underestimated and in the fits without longer burn-in, drift rates are underestimated. With these fixes, the results do not show the large deviations as in the first version. However, the recovered parameters have somewhat larger variability in larger SDs in many cases than the other methods illustrated in Figure 5.

We also tried fixing the proportion of contaminants at 0.0 and 0.02 with the longer burn-in. For recovery with the proportion of contaminants 0.02, some of the recovered mean drift rates (over the 64 simulated data sets) were 1.5 times the true value used to generate the simulated data. For the proportion of contaminants 0.0, some of the recovered mean drift rates were 2 times the true value. Thus, there is a problem that is largely fixed with the longer burn-in and with the proportion of contaminants fixed at 0.05. However, as we see later, the puzzle is that this problem does not arise with four conditions in the lexical design.

The DMAT method without correction for contaminants produced parameter values that were biased toward larger values than the generating values and the SD's in most of them were larger, some 2 to 3 times larger, than the chi-square and fast-dm methods. The misses in the estimates for DMAT with correction for contaminants were even larger and more variable, so large that it is difficult to see that the correction for contaminants should be used in any situation.

EZ produced some of the smallest SD's (for a, v_1, and v_2), but its recovered values missed the generating values by large amounts. Estimated drift rates were as low as half their generating value and nondecision times were as low as 250 ms instead of 400 ms. Because the SD's are so

small, essentially the EZ model is very sure of the parameter values but very wrong.

Large numbers of observations for the lexical design, 300 observations for each of the word conditions and 600 observations for the nonwords. Generally, the results mirror those from the numeracy design and so provide a replication of its findings. Figure 6 shows the recovered values for the 32 means and SD's for each of the parameters.

The chi-square methods show a little more bias for some drift rates, especially for the 9 quantile method, compared with the numeracy design. Fast-dm shows biases about the same size as for the numeracy design.

HDDM again shows some conditions with extreme biases that result from the lower value (0.1) of boundary separation, and with 4% contaminants. The unbiased values from HDDM are closer to the true values with smaller SD's than for the chi-square methods. DMAT without contaminant correction shows larger SD's in model parameters than the other methods and shows some quite large biases in drift rates. As for the numerosity design, DMAT with contaminant correction show even larger SD's and biases, values so large that 2 SD's for some drift rates include zero.

The EZ method, as for the numerosity design shows quite small SD's, especially in drift rate, but very large biases, so large that some of the 0.1 drift rates are estimated to be zero. Similarly, nondecision time varies from less than 200 ms to 500 ms when the value used to generate the simulated data was 400 ms. This is expected because the EZ method assumes the starting point is halfway between the boundaries, and in a number of the sets of parameter values, the starting point is biased with values $a/3$ or $2a/3$.

Thus, for this design, only the chi-square methods and fast-dm produce acceptable parameter recovery with small bias (relative to the other methods) and with small SD's in recovered parameters. If the spurious values produced by HDDM were fixed, then it would be acceptable also.

Small numbers of observations. For the numerosity design, there were 40 observations per condition and 100 observations per condition, but we only show results for 40 observations per condition (summaries from the 100 observation study are discussed later). For the lexical decision

design, there were 30 observations for each of the word conditions and 60 for nonwords. The recovered values are shown in Figures 7 and 8.

Some of DMAT's recovered values were off the scale of those shown in the figures and plus and minus 1 SD included zero and extended outside the ranges shown in the figures for some combinations of parameters. For this reason, DMAT results were not plotted and it is not possible to recommend DMAT when the number of observations is small. This is not a criticism of DMAT because it was never designed to work with this few observations.

For the numeracy design for the other five methods, the estimated values of the parameters were farther from their true values than for the results with 1000 observations per condition and the SD's were larger. The chi-square methods produced parameter values that were biased toward larger values than the true values and fast-DM produced values that were biased toward smaller values. For all these methods, the SD's were particularly large for drift rates: a plus or minus 2 SD confidence interval in the drift rate with a true value 0.2 often included zero. HDDM produced smaller biases and smaller SD's than the chi-square and fast-dm methods for values of boundary separation and nondecision time, but the estimates of drift rates were biased to values larger than the true values (the biases were similar to those for the chi-square method) but the SD's were somewhat smaller than those for the chi-square and fast-dm methods. EZ's recovered values were again far from the true values.

In all these results for low numbers of observations in the numeracy design, the size of the SDs for boundary separation are quite well estimated for the 0.1 value but for the 0.2 value, some of the SDs are quite large. For nondecision time, 1 SD confidence intervals in the value for the chi-square, fast-dm and HDDM methods have about a 100 ms range.

For the lexical decision design with 30 observations for the word conditions and 60 for the nonwords, the results mirror those for the numeracy design with 40 observations per condition. The SDs in drift rates are a little larger but the biases are of similar size.

The results for HDDM are a puzzle because for some combinations of parameters, biases were less than with larger numbers of observations per condition. This might be because the contaminant model does not work well with larger numbers of observations for which long outliers

are systematically present in the data. With smaller numbers of observations, there would be fewer contaminants and so the correction for contaminants might not be evoked.

EZ parameters were estimated more poorly for the lexical decision design than for the numerosity design as for the large numbers of observations study because of the bias in the starting point for some of the conditions.

DMAT warning messages. There are several warning messages provided by DMAT that indicate problems with parameter estimation. “The last convergence point was still a suspect result” indicates some of the across-trial variability parameter estimates may be wrong, “Hessian is not positive definite”/”Hessian is not of full rank” indicates some parameter is not sufficiently identified by the data, and “Matrix is close to singular .." indicates some of the standard error estimates are likely biased. For the numeracy design with 40, 100, and 1000 observations per condition, the percentage of these three error messages across the 16 set of parameters and 64 simulated data sets per parameter are: 52%, 52%, and 13% for the three messages for N=40 per conditions, 23%, 31%, and 13% for the N=100 per conditions, and 2%, 6%, and 0% for the N=1000 per conditions. For the lexical design, the percentages were: 70%, 57%, and 18% for the smaller number of observations and 36%, 12%, and 2% for the larger number of observations. These error messages indicate that the fits might be invalid (as spelled out in Vandekerckhove & Tuerlinckx, 2008). The results show that the DMAT package is warning that the fits are potentially flawed even with N=100 per condition, but are mainly good with N=1000 per condition.

EZ variants. We evaluated two EZ variants, EZ2 and robust EZ, and found that neither was competitive with any of the methods examined so far. These variants were designed to address the problems with the assumption that the starting point is midway between the boundaries (EZ2, Grasman, Wagenmakers, & van der Maas, 2009) and the problem with contaminants (robust EZ, Wagenmakers, van der Maas, Dolan, & Grasman, 2008).

We began to examine these methods using just a few conditions rather than start with a large comprehensive study. For EZ2, we took 8 conditions from the lexical decision design with z at values $2a/3$ and $a/3$ and with both the larger and smaller numbers of observations. For robust EZ, we used all 16 sets of parameter values with both 0% and 4% contaminants for the numerosity

design for all three sets of numbers of observations. We found that the methods were rather poor at recovering parameter values and so did not pursue them further. We now describe the results from these studies.

EZ2 was designed to produce parameter estimations when the starting point is not half way between the boundaries. Like the EZ method, it assumes no across-trial variability in model parameters. With both the lower and larger numbers of observations, parameter recovery was not good. For a=0.1, the average recovered value was 0.13 and for a=0.2, it was 0.21. For the narrow boundary separation (0.1), when z/a was 0.3, one drift rate was estimated to be 0 and the other 0.23 while when z/a was 0.7, one drift rate was estimated to be 0.23 and the other -0.3 when both were 0.2. Even in the cases in which z/a was 0.6, the mean drift rate was estimated to be 0.114 instead of 0.2. These results showed large deviations in conditions that the model was supposed to address and so we did not pursue the model further.

Robust EZ was developed to estimate the distribution of contaminants and then remove them from the analyses, in effect running EZ on the decontaminated distribution. The method involves first fitting a mixture distribution to the data: an exGaussian for the diffusion model and a uniform for contaminants, and then using the exGaussian distribution to compute the mean and variance of the decontaminated distribution which are then used by the EZ method to produce diffusion model parameters.

We ran this model on the data set for the numerosity task in the second simulation (for the three sets of numbers of observations and 16 sets of parameter values). The results showed that, for example, when the original boundary separation was 0.2, the robust EZ analysis produced a value of about 0.138 for all the conditions with a=0.2. The proportion of outliers estimated did not track the actual number. When there were 40, 100, and 1000 observations per condition, we found the estimates of the proportion of contaminants were 38.4%, about 20.3%, and about 6.1% (even though half the sets of parameter values had 0% contaminants and half 4%). Other model parameters were not as poorly estimated, for example, when a was 0.1, the estimated parameter was 0.088, T_{er} was 0.36 instead of 0.4, and the two drift rates were 0.078 and 0.22 instead of 0.1 and 0.2. But the problem with the wider boundary separation and proportion of contaminants made

this variant uncompetitive.

Short Outlier RTs

Ratcliff and Tuerlinckx (2002) found that short outliers caused serious problems for maximum likelihood fitting methods. The reason is that in order to compute a likelihood for a short RT, there has to be probability density at that point. This means that the estimate of nondecision time has to be below that shortest RT if there is no across-trial variability in nondecision time, and if there is across-trial variability, then T_{er}-s_t/2 has to be smaller than the shortest RT. This produces extreme biases in parameter estimates. In contrast, quantile-based methods are robust to a small percentage of short outliers. For DMAT, contaminant correction methods did not work well although it is robust when the number of short outliers is small because it is quantile (or bin) based. Note that the MLH method without a contaminant model is even less robust (Ratcliff, & Tuerlinckx, 2002).

To complete the picture, we examined the effects of short outliers on the fast-dm and HDDM packages using simulations with a subset of parameters from the simulations above. Fast-dm does not have an outlier or contaminant model, but the Kolmogorov-Smirnov statistic used to evaluate it is robust to a few outliers. HDDM implements the same outlier model as in Ratcliff and Tuerlinckx but sets a lower limit at 0 instead of the minimum RT as Ratcliff and Tuerlinckx did. We now examine how well these methods worked in the face of short outliers.

We performed two studies with a relatively large number of short outliers (though many more than this number can be obtained with uncooperative subjects). In the first, we used the first 4 combinations of parameter values in both the numerosity and lexical decision simulations and randomly replaced 5% of the RTs with a RT of 50 ms. In the second, we replaced the fastest 5% of RTs with 50 ms. The latter would have no effect on quantile methods and the former would affect parameter estimates only modestly (because 5% of RTs are replaced with shorter ones). We used three sets of numbers of observations for the numeracy design and two for the lexical decision design with 64 sets of simulated data for each of the combinations of parameters (just as for Simulation Study 2). The results were only minimally different for the two designs, so we present only the summary of both studies together.

For HDDM, boundary separation was underestimated by up to 15%. Nondecision time was overestimated by up to 60 ms over the 400 ms true value except when boundary separation was 0.20 and there were 1000 observations per condition; in this case, the estimate was too large by about 100 ms. Drift rates were overestimated with 40 observations per condition by up to 50% (estimates of 0.149 and 0.292 instead of 0.1 and 0.2) and they were underestimated by up to 25% with 1000 observations (estimates of 0.084 and 0.163 instead of 0.1 and 0.2).

For fast-dm, there was underestimation of boundary separation by up to 15% and generally underestimation of nondecision time by up to 30 ms (though in some cases there was overestimation by up to 20 ms). Drift rates were underestimated for all but one case by up to 25% (0.073 and 0.146 instead of 0.1 and 0.2). These results show that fast-dm and HDDM are not robust to 5% short outliers.

The fact that fast-dm and HDDM were not robust to 5% short outliers means that more than a very small proportion of short outliers is likely to be a problem. It is important that experimenters examine data and exclude non-cooperative subjects and/or short outliers. Short outliers can also be eliminated with experimental methodologies that punish extra fast responses by inserting an extra (1-2 second) delay between the response and the next test item. If short outliers are eliminated either experimentally or by eliminating responses or non-cooperative subjects, then we do not have to worry about problems with fast-dm and HDDM. In contrast, the quantile based methods have the benefit of being robust for a few percent (under 10%) short outliers and so parameter estimates are not affected even if these outliers are not eliminated.

Simulation Study 3: A Hierarchical Bayesian Diffusion Model

Hierarchical models assume that the values of parameters for individual subjects come from a distribution of values across some group of subjects. Individuals' values and the group values are estimated simultaneously; the group sits hierarchically above the individuals. To the degree that the subjects are similar to each other, variability in the group will be estimated to be small and this will constrain the individuals' parameters to be closer to the mean. The method allows an individual to be distinguished from the group if there are enough data and the difference between individual and group is large enough (see Farrell & Ludwig 2008; Vandekerckhove,

Tuerlinckx, & Lee, 2011; Vandekerckhove, Verheyen, & Tuerlinckx, 2010; Wiecki, Sofer, & Frank, 2013).

Hierarchical Bayesian methods for fitting the diffusion model have been developed by Vandekerckhove, Tuerlinckx, and Lee (2011) and Wiecki, Sofer, and Frank (2013). The latter authors have made available an HDDM (Hierarchal Drift Diffusion Model) software package that fits the diffusion model to data and we used that package for the simulations below. In their implementation, the parameter values for the group are drawn from specific distributions: normal for drift rates and nondecision times and Gamma for boundary separation.

Hierarchical methods have an advantage over non-hierarchical methods in that they can be used when the number of observations per condition is small, as small as the number of parameters in a model plus 1, but there are also other advantages and disadvantages. One is that the distributions of parameters assumed for the group (e.g. normal or Gamma) may not be their true distributions and so estimates of parameters can be biased. This is not a problem for non-hierarchical methods for which the parameter values of a group are simply averages of the individuals' values. Another advantage or disadvantage, especially with low numbers of observations, is that the estimates of parameters for individuals can be pulled toward the values for the group; this is termed "shrinkage." This is an advantage when extreme parameter estimates are due to variability arising from small numbers of observations (Efron & Morris, 1977) and the hierarchical method pulls them back towards the true values. This could be a disadvantage when the aim is to identify sub-groups of individuals within a group, for example, when individuals with a deficit are mixed with individuals without deficits. A problem might occur if shrinkage pulls the estimates of parameter values for the sub-groups toward the mean for the whole group and so makes it difficult to distinguish between them. As we see below, except for very small numbers per condition, this is not a problem in our study.

In HDDM, across-trial variability in drift rates, nondecision time, and starting point and the proportion of contaminants are set the same for every subject because with small numbers of observations, the differences produced by different values of these parameters in the function being minimized are small and Wiecki et al. (2013, p.3) suggest that this might lead to spurious results. In the simulations here, we allowed differences in across-trial variability among subjects in the

simulated data, which results in misspecification of the HDDM hierarchical fitting method. Such differences among subjects in the across-trial variability parameters would certainly occur in real data.

For Simulation Study 3, we compared the HDDM to the 9-quantile chi-square method using the numerosity design (two conditions differing in difficulty) with 40, 100, or 1000 observations per condition with no contaminants or 4% contaminants and 64 subjects. With a number of observations as large as 1000, hierarchical and non-hierarchical methods usually produce similar estimated parameters (e.g., Farrell & Ludwig, 2008). We used the same method of producing combinations of parameters as in Simulation Study 1, where the parameter values for each simulated subject were drawn from a normal distribution centered on a parameter's mean. For each combination, we simulated 64 subjects and fit their data by the two methods under three different distributions of parameters across subjects. One was that the distributions of drift rates, boundary separation, and nondecision times across subjects are all normal. The second was that they are all uniform with the same SD's as for the normal case. The third was that 54 of the subjects had one set of parameter values drawn from normal distributions and the other 10 had values drawn from normal distributions with larger boundaries and nondecision time and lower drift rates than for the other 54 subjects.

In the last part of this section we present results from the hierarchical model fit to simulated data for the three distributions across subjects above with only five observations per condition (this was suggested by Joachim Vandekerckhove). Using very few observations has been a selling point of Hierarchical models, and this set of simulations shows what happens with the diffusion model with very few observations. As noted earlier, if all that was available was a few trials with no practice, then results would be suspect because performance could change radically over 10 or 20 trials with practice and warm up effects. Note, this is not an issue with the hierarchical model, it is a problem with the data, i.e., the hierarchical model would not solve the problem of a very few trials. To avoid this problem, it would be possible to test subjects on some related task for a few minutes with the same response keys.

Normal Distributions Across Subjects

The means and SD's in the parameter values were those in the fourth and third lines from the bottom of Table 1 for which individual subjects varied from each mean according to a normal distribution.

Figure 9 shows the recovered values of the two drift rates, boundary separation, and nondecision time for the two methods for 40 observations per condition and 1000 observations per condition for the two methods (we used the results for the 100 observations condition later, but because the figures look the same as the ones presented, they are not shown). We present the plots in a different way from the other plots to highlight the issue of shrinkage of the parameters (regression to the mean).

Plotted on the x-axis are the values of the parameters from which the simulated data were generated. Plotted on the y-axis are the recovered values minus the value used to generate the simulated data (i.e., the residuals). If there is no shrinkage, the plots will be horizontal, but if there is shrinkage, the lines will have negative slope. Figures 9, 10, and 11 plot the results from the hierarchical method and from the chi-square method, but if we plotted residuals, the functions would largely lie on top of each other and they would be hard to compare. To spread the values apart, they were offset by adding a constant for the chi-square method (the circles in the figure) and subtracting a constant for the hierarchical method (the x's in the figure). Any other biases in the recovered values will be seen as systematic deviations between the points and the horizontal solid lines. Also, the spread in the recovered parameter values, the residuals, can be used to compute the SD in the difference between the recovered parameter values and those used to generate the data. Thus, the vertical spread about the horizontal line visually represents this variability.

The top solid horizontal line in the figures is the mean offset for the chi-square method, the bottom solid line is the mean offset for the hierarchical method, and the dotted lines are regression lines. The correlations between the recovered values of the parameters (not the residuals) and the values used to generate the simulated data are shown in the headings of the panel.

For 40 observations per condition, correlations for the hierarchical method were higher than for the chi-square method for boundary separation and drift rates, but lower or the same for nondecision time. For 1000 observations per condition, the correlations for the chi-square method

are higher for all the parameters.

The hierarchical method (but not the chi-square method) consistently underestimates nondecision time, with most of the points falling below the horizontal line. The hierarchical method also overestimates boundary separation with 4% contaminants (as does the chi-square method to a lesser degree), but does not do so with no contaminants.

For drift rates, there is moderate shrinkage in HDDM's recovered parameters because the dotted regression line for all the drift rates for all the conditions has negative slope. The drift rates for the chi-square method do not show systematic shrinkage. Perhaps surprisingly, there is no shrinkage in the boundary separation and nondecision time parameters for the HDDM model.

We discuss possible reasons for these results and the number of observations at which the chi-square method begins to outperform the hierarchical method, i.e., the cross-over point after the next two studies.

Uniform Distributions Across Subjects

For this study, the distributions for drift rates, boundary separation, and nondecision time came from uniform distributions with the same means and SD's as for the simulation just discussed. This means that each of these distributions across subjects is modestly mis-specified relative to assumed parameter distributions in the hierarchical model.

Figure 10 shows the results in the same way as for Figure 9 with normally distributed parameter values and results showed patterns that were qualitatively similar. The only major difference was that the shrinkage in drift rates was reduced relative to normal distributions in Figure 9, especially for the lower drift rates.

Two Groups of Subjects

Simulated subjects were drawn from two well-separated populations. For 54 of the 64 subjects, we used the means as in the two studies just discussed but with smaller SD's, namely those in the bottom row of Table 1. For the other 10 subjects, we used the means and SD's in the last two rows of Table 1.

We used smaller SD's in the parameter values across subjects relative to those from the

normal and uniform distribution simulations to provide clear visual separation between the two groups. (We performed a study before this one with the values of the SD's the same as those from the normal and uniform distribution simulations, but the groups were not as well separated. The results were similar to those for this study).

The means of the generating parameters over subjects for the two groups were 0.11 and 0.20 for boundary separation, 375 ms and 530 ms for nondecision times, 0.3 and 0.1 for the easy condition and 0.1 and 0.0 for the difficult condition. These are large differences between the two groups, but not out of line with differences in age and IQ (e.g., the 54 subjects might come from a high IQ young adult group and the 10 subjects might come from a low IQ older adult group, Ratcliff et al., 2010).

The pattern of results (Figure 11) is similar to those for the normal and uniform distributions in Figures 9 and 10. The first thing to note is that for both the chi-square and HDDM methods, the recovered parameters for the two groups of simulated subjects were as well separated in the recovered parameters as they were in the generating values, i.e., the two groups were visually separate. For boundary separation and nondecision time, there are groups of 10 points at the right ends of the plots and for drift rate they are at the left ends of the plots.

We noted earlier the concern that the HDDM method would lead to shrinkage, pulling the recovered values for individual subjects toward the values for the group. If this was extreme, the parameters for the two groups of subjects would merge so they could not be discriminated. If this occurred, then this would make the hierarchical model not useful for applications in which discrepant individuals were to be identified. However, for these simulations this did not occur and the groups were separated. There was some shrinkage particularly in the higher drift rate condition as in Figures 9 and 10 (the dashed line has negative slope), but this does not seem to affect the ordering of the drift rates and does not cause the two groups to merge.

For 40 observations per condition, the correlations between the recovered model parameters (boundary separation, nondecision time, and drift rates) and the parameters used to generate the simulated data are higher than for the chi-square method which shows that HDDM provides superior parameter recovery for all the parameters. But for 1000 observations per

condition, the chi-square method provides the higher correlations and better parameter recovery.

Summary of the Hierarchical Model Results

Figure 12 plots the correlations between the best-fitting recovered values and the true values for boundary separation, nondecision time, and the two drift rates as a function of the number of observations. The first column shows the correlations from the case when the distributions of the parameters for the group were assumed normal, the second column for when they were assumed uniform, and the third column for when the distributions were assumed to be normal and the simulated subjects came from two groups with different mean values of each parameter for the two groups. For Figures 9, 10, and 11, we reported results for 40 and 1000 observations per condition; here we include the correlations for 100 observations per condition.

The figure shows approximate cross-over points between the two methods, that is, the points at which correlations switch from having higher values for the HDDM method to higher values for the chi-square method. The oblique arrows in the figure point to the cross-over points. For boundary separation, the cross-over is somewhere around 100 observations per condition, for drift rates, the cross-over is several hundred observations, but for nondecision time, there is no cross-over and the chi-square method produces higher correlations for all sample sizes. These conclusions about the points at which the cross-overs occur are qualified by having only three numbers of observations per distribution type. However, the important finding is a qualitative one, that the correlations do cross-over.

Very Low Numbers of Observations per Subject

Joachim Vandekerckhove suggested running simulations with the hierarchical method with the number of observations equal to the number of parameters plus 1. We used the numerosity design and 10 observations, 5 for each of the easy and hard conditions and the number of parameters for each subject was 9.

For each condition, we fit the hierarchical method to the first five trials in each condition from the data sets with 1000 observations per condition with the three distributions of parameter values from Figures 9, 10, and 11. Five observations per condition would be impossible to fit with the non-hierarchical methods because the low numbers of observations would produce estimates

with very high variability and there would not be enough observations to produce meaningful RT distributions for both correct and error responses. Also, it is likely that the fitting programs would not converge on a solution most of the time.

The hierarchical fits for the three distributions of subject parameters from the prior simulations for 0% contaminants are shown in Figure 13 and the plots are in the same format as for Figures 9, 10, and 11. The correlations in parentheses are from the condition with 4% contaminants (the plots for 4% contaminants are very similar to those presented in Figure 13).

The first thing to note is that there is extreme shrinkage in the drift rates, especially for the normal and uniform distributions of parameters across subjects. In the plots, if the recovered parameters match those used to generate the data, the plot will be horizontal. Instead, the plots are diagonal for drift rates with a slope close to -1 which means that the drift rates are estimated to be almost the same across conditions. For the most extreme case, for the uniform distribution of parameters for v_1 (the higher drift rate), the mean value of the drift rate parameters used to generate the simulated data is 0.284 and the mean recovered value is 0.247, but the SD in the drift rate parameters used to generate the simulated data is 0.094 and the SD in the recovered parameter values is 0.0054, i.e., only 6% of the variability used to generate parameters is recovered. For T_{er} for the two-population plot, there appears to be considerable shrinkage and the SDs in recovered values are only about half that of the values used to generate the simulated data.

One explanation for the extreme shrinkage in drift rates is in terms of what features of the data determine drift rates. Drift rates are most related to accuracy rates, and because there are only 5 observations per condition, there are not enough observations to constrain drift rates. Thus the hierarchical model constrains them to be nearly the same. In contrast, even with 5 observations, the RT differences from differences in boundary separation and nondecision time across subjects are large enough to produce moderate to large differences in recovered values of these parameters.

Despite the shrinkage, the correlations between the parameters recovered by HDDM and the parameters used to generate the simulated data are quite high. This is especially the case when the model is misspecified with two populations (leading to a larger spread of values) rather than one normal population. Thus one could use the recovered parameter values to examine individual

differences within the group. But because of the amount of shrinkage, the recovered parameter values could not be used to compare groups unless a study was done to examine biases for the different sets of parameter values for the groups and individual differences within groups.

Even though the hierarchical method with very few observations might recover individual differences relatively well, there are serious limitations with experiments based on this few observations. These include warmup and practice effects. For example, when undergraduate subjects begin an experiment, it is usual in our laboratory to ignore the first block of trials. If we are testing with a limited number of materials, we may test them for a few minutes on a related task, such as lexical decision or a perceptual task in order to familiarize them with the response requirements. In the first few trials, subjects may working out which fingers to use, still talking to the experimenter and so on. For older adults or adults who might have deficits, the warm up period might be significantly longer. We often train older adults, for example, for a few minutes on a different task to familiarize them with the stimulus presentation and response recording methods. An extreme case of warm up effects were experiments involving speed-accuracy instructions: Older adults required two or three full 45 min. experimental sessions before their performance was stable (e.g., Ratcliff et al., 2001, 2003; Thapar et al., 2003). Therefore, even thought the hierarchical model may recover parameters reasonably well with very few observations, it would be a mistake to assume that the data were of adequate quality if this were the first reaction time task they encountered. Note that this is a problem with data and applies to all methods.

In all these simulations with the hierarchical model (HDDM) used here, the model is misspecified because it assumes all the across trial variability parameters and the proportion of contaminants have the same value across subjects. This misspecification might explain why the standard chi-square method outperforms the hierarchical method for larger numbers of observations. If this misspecification were eliminated by fitting a hierarchical model with all parameters free to differ, the hierarchical model may well outperform standard methods in all situations.

More generally, for educational or neuropsychological testing, if the hypothesis involves drift rates (evidence from the stimulus or memory), and there are a low number of observations (on the order of 100 or 200), then the hierarchical HDDM would be the superior method. But if there

were larger numbers of observations, and all parameters were of interest, than standard methods might be just as good or even superior. It is important to note however that shrinkage is likely not to be a problem in identifying subjects with deficits in hierarchical HDDM in the design studied here (with the caveat that results may differ in different designs).

General Discussion

The studies in this article were designed to examine how well parameters can be recovered when the diffusion model is applied to data of the kinds that might be obtained from studies in practical domains such as neuropsychological, clinical, and educational testing. We explored under what circumstances differences between individuals and groups can be detected and under what circumstances differences between groups can be detected. We also conducted three simulation studies to examine the ability of various fitting methods to recover individual differences and parameter values.

Experiments 1 and 2 provided estimates of the parameters for individual subjects for a numeracy task and a lexical decision task. We used the parameters from fitting all the data and subsets of the data to examine power. Results showed that differences in nondecision time and boundary separation might be large enough to detect discrepant individuals, but individual differences in drift rates were so large that drift rates would have to be near zero for most of the conditions to detect that an individual was discrepant. The wide range of drift rates is a result of the wide range of performance of undergraduates in these experiments. For example, with 61-70 and 31-40 asterisks, the best-performing subject had 97% correct and the worst had 69% correct (other subjects had accuracy values in the low 70% correct range). Either the latter subject had extremely poor numeracy skills or was not trying. Because drift rates are highly related to accuracy values, the fact that the bottom of the confidence intervals on drift rates are close to zero is due to these poor performing subjects. Thus, for any practical applications, a group of control subjects is needed that matches the experimental group on motivation as well as other characteristics, but the key point is that our experiments provide an illustration of how power might be assessed.

We also examined whether the recovered parameters changed with practice from early to late blocks of data and from smaller numbers of observations to larger numbers. Moving from a

few observations (80 or 120) to a large number (over 1000) changes the SDs in individual differences by about a factor of 2. The SDs in model parameters for each fit decrease approximately with the square root of the number of observations, but the SDs across subjects (in Tables 1 and 2 and Figures 1 and 2) decrease less because there is still a large range of individual differences, but the parameter values are better estimated and so values that were estimated to be more extreme become less extreme.

In the first simulation study, simulated data were generated with a random selection of parameter values from the ranges seen in the two experiments. Then the various fitting methods were applied to those simulated data and correlations between the recovered parameter values and generating values were used to examine how well the methods recovered individual differences. We found that the two chi-square methods, fast-dm, and the maximum likelihood methods produced the highest correlations for the simulations that differed in the number of observations and the presence or absence of contaminants. HDDM was the best performing for the lexical design, but produced some poor correlations for the numeracy design. DMAT performed less well than the other methods when there were low numbers of observations (because it does not use error RTs when the number is less than 11) and it performed poorly when the contaminant correction methods were used. The EZ method was competitive with the other methods when there were no contaminants, but with contaminants, correlations were lower.

The correlations in Simulation Study 1 show the upper limit on correlations from measurement error with the range of individual differences observed in the two experiments. For experiments with different designs, the results from this study could be used as a guideline if the designs were not too different, otherwise a new study would be needed. For the better fitting methods, if there are a couple of hundred observations and 2 conditions (with starting point symmetric between the two boundaries), as in the numeracy design, correlations between recovered model parameters and those used to generate the data were generally above .9 for boundary separation and nondecision time and above .75 for drift rates, even in the worst case with 4% contaminants. For the lexical design, even with only 120 observations, correlations for boundary separation and nondecision time were above .85 and for drift rates above .6. These values are large enough so that with 10-15 minutes of data collection, measurement error in model

parameters will be low enough so that any relationships between other measures (IQ, reading or numeracy ability measures, etc.) will be able to be detected.

We have performed studies with similar designs that have produced large individual differences and correlations in the .5 to .6 range across tasks in model parameters (McKoon & Ratcliff, 2012, 2013; Ratcliff, et al., 2010, 2011; Ratcliff, Thompson, & McKoon, submitted). These reinforce the claim above that with larger numbers of observations, recovery of model parameters is good enough (i.e., the relative order as reflected in high correlations) to produce strong meaningful relationships (correlations) even across tasks.

The results from Simulation Study 2 show the means and SDs in model parameters recovered from 64 sets of simulated data with the same parameter values for a number of different parameter sets for two designs with five different numbers of observations. These show the biases in model parameters (i.e., bias from the generating value) and the SD in the recovered parameter values. Results showed that with large numbers of observations, model parameters were recovered by the better methods with low bias and low SDs. Some of the methods (DMAT with contaminant correction, EZ, and HDDM) had discrepant values for some parameter combinations. Some of the methods that were better performing for individual differences produced consistent biases in model parameters, and these differed as a function of the parameter values. For example, fast-dm for the lexical design with the large numbers of observations had boundary separation vary between .15 and .2 when the value used to generate the simulated data was .2. At the same time, some of the drift rates had a bias to low values (e.g., for nonwords, the correct value was .2 and the recovered values ranged from .1 to .2. The problem that this raises is that if the other model parameters differ, no difference in boundary separation between two groups of subjects could be interpreted as a difference if the other parameters differ systematically as in the simulations in Figures 5 and 6.

For low numbers of observations, HDDM provided better parameter recovery than the other methods and did not show the discrepant values that occurred with the larger number of observations. The other methods showed similar biases as with large numbers of observations and as before, the DMAT method was not competitive. EZ showed some extreme biases with 4% contaminants.

We performed comparisons between the chi-square method and the hierarchical method for three populations of individual differences on model parameters and for 3 numbers of observations. We used normally and uniformly distributed populations and a population with two subpopulations (representing a normal group and a group with deficits in the main model parameters). For low numbers of observations, the hierarchical model outperformed the chi-square method, but when there were a large number of observations, the chi-square method outperformed the hierarchical method. This is likely because the HDDM implementation of the hierarchical method assumed that the across trial variability parameters and proportion of contaminants were the same across subjects whereas the simulated data were generated with differences across subjects (as in real data).

There were two main concerns with the hierarchical model. First, when the distribution of model parameters was not the same as the distribution assumed by the hierarchical model, individual differences may be compressed or distorted. In fact, this did not matter at all and the simulations with two populations showed nice separation of the two groups. Second, there was a worry that the recovered parameters would be drawn to the mean (termed shrinkage). But this did not happen to a serious degree even with only 40 observations per condition. Drift rates showed shrinkage, but the model parameters were well enough recovered so this was not a problem.

We then performed simulations with 5 observations per condition with the hierarchical model. Note that none of the other methods would have produced anything meaningful. Individual differences in boundary separation and nondecision time were recovered to some degree, but there was extreme shrinkage in drift rates (the model tended to recover values of drift rates that differed little from the mean). With two populations nondecision time shrunk a lot but the drift rates for the two populations were recovered to some degree. Note that we would not recommend using data with so few observations, especially in clinical or neuropsychological populations, because of practice and warm up effects including getting appropriate instructions to the patients, for example. In such populations, the first few dozen trials may involve orienting to the task to learn how to perform it.

Contaminants and Outliers

There are two kinds of contaminants in the tasks to which the diffusion model is usually applied, fast guesses and slow responses. Slow responses can just involve a delay in processing or could be guesses. In some paradigms, it is easy to see if there is a moderate proportion of fast random guesses if there is an easy condition in which accuracy is at ceiling. If accuracy in that condition is 98% correct, then no more than a few percent of the responses could be guesses. It might be assumed that the proportion of guesses differs as a function of experimental condition (for example, with more guesses for difficult conditions), but such an assumption is rarely made.

Another way to eliminate fast guesses from data for some paradigms is to exclude responses that occur before the point at which accuracy starts to rise above chance (this is similar to the fast error correction method in the DMAT package). In our laboratory, we use such a cutoff, but if a subject has more than just a few fast guesses, we treat them as non-cooperative (with the instructions) and eliminate them from the experiment. This occurs sometimes with undergraduates participating for course credit but rarely for paid subjects. We reduce such non-cooperation experimentally by presenting a message, for example for 2 seconds, saying “too fast” if the RT is less than 200 ms. Often subjects that are fast guessing are trying to get out of the experiment quickly, and the delay slows them down relative to regular responses.

Slow outliers that are not guesses can be the result of a moment’s inattention, a scratch of the head, an interruption from a cell phone call, etc. A problem with such outliers is that some will not lie outside the normal range, those with a short delay and a fast decision process (that is why we call these contaminants rather than outliers). For fitting the model to data, it is assumed that processes with these delays can be represented by a uniform distribution that has a range from the minimum and maximum RT. Therefore RTs are a mixture of this uniform distribution and regular diffusion process responses. Ratcliff and Tuerlinckx (2002) generated simulated data with a random delay added to some small proportion of RTs produced from the diffusion model and showed that the mixture model recovered model parameters and the proportion of contaminants well. In an experiment with sleep-deprived adults, Ratcliff & Van Dongen (2009) found that under sleep deprivation, a moderately high proportion of responses for a few subjects were random guesses (because conditions with high accuracy in non-sleep deprived conditions dropped from say 97% correct to 85% correct in the sleep deprived condition). They modeled this case with a

proportion of random choices with RTs represented by a uniform distribution with a range from the minimum and maximum RT.

The HDDM method, the fast-dm method, and the chi-square methods are quite robust to slow contaminants. Parameter recovery with 4% contaminants is good and the effects of a very few fast outliers are minimal. When there are 5% fast outliers, the recovered parameter values do not differ much from the parameter values used to simulate the other 95% of the data for the chi-square methods, but there are modest deviations of model parameters for fast-dm and HDDM. DMAT is less robust to slow contaminants and it performed more poorly than the other methods on most of the simulations from studies 1 and 2 above. The EZ method is not robust to slow contaminants (Ratcliff, 2008) and, because it is based on the SD in RTs, it is also not robust to fast contaminants.

Biases, Parameter Tradeoffs, and Over-Fitting

When applying a fitting method to data to uncover differences between individuals, success means that the recovered values are not biased away from their true values or are at least equivalently biased across subjects and conditions. For example, if the estimate of boundary separation were underestimated by the same amount for all subjects in an experiment, say by 20%, then questions about individual differences can be answered in the same way as if the parameter was unbiased. However, if the bias changed from one subject to another as a function of the other model parameters (or other things such as the proportion of contaminants) then systematic biases in estimation might be interpreted as individual differences. This would become important to investigate if parameter estimates were being used as an index of, for example, cognitive deficits.

One important problem with fitting any model with low numbers of observations is that the relatively high variability in data will lead to overfitting. Because model parameters are adjusted to get as close to the data as possible, extra high or low values of accuracy or RTs will be accommodated by some combination of extra high or low values of the parameters. For example, Ratcliff & Tuerlinckx (2002, Figure 7) examined the correlations among model parameters when one RT quantile was high by chance. Results showed that several model parameters were adjusted in fitting to accommodate the misfit in the data. This resulted in model parameters covarying with each other; across trial variability in drift rate, boundary separation, and drift rate correlated

between .35 and .66. Such overfitting is signaled by low values of chi-square relative to the significance level (as, for example, for the first three trial groups with low numbers of observations in Tables 1 and 2). However, usually these tradeoffs are much smaller in magnitude than differences in parameter values across individuals.

Because there are correlations between model parameters for the fits, we attempted to see if combinations of them provided a more compact description of the results. We ran exploratory factor analysis on the model parameters (for both low and high numbers of observations in the numeracy design) to see if some combination of parameter values produced better correlations between fitted parameters and those used to generate the simulated data. Although some of the factors made sense, the loadings were weak and did not support the attempt to extract factors from combinations of the model parameters (for these experiments).

Individual Differences

Experiments 1 and 2 provided estimates of model parameters and individual differences in them as a function of the number of observations used in model fitting. Figures 1 and 2 (and Tables 1 and 2) show that the SD's in the parameter values for boundary separation and nondecision time decrease by less than a factor of 2 going from 80 to 1200 observations or from 120 to 2100 observations. Drift rate SD's decreased by about a factor of 2. This can be attributed to a decrease in the SD in model parameters with more accurate recovery from the greater number of observations. Figures 7 and 8 show the SD's in model parameters for the lowest numbers of observations for simulated experiments similar to Experiments 1 and 2. The SDs in boundary separation and nondecision time even for the lowest numbers of observations in the simulations are less than the SDs for parameters recovered for fits to all the data in each experiment. Therefore, the SDs in model parameters from variability in data given the number of observations is smaller than the individual differences in model parameters.

To provide a concrete example of the effect of variability in parameter estimates relative to individual differences, a set of 1000 normally distributed random numbers was generated with SD *s*. From these, three sets of random numbers (*u*, *v*, and *w*) were generated with means *x* and SDs either *s*, .75*s*, or .5*s*. These represent examples in which the *x* values represent individual

differences, and the *u*, *v*, and *w* values represent estimates with SDs in the estimate either *s*, .75*s*, or .5*s*. The correlations between *x* and *u*, *v*, and *w* are .72, .80, and .89 which shows that even if the SD in estimation is the same value as the SD in individual differences, the ceiling on the correlation would be reduced from 1 to only .72.

To examine individual differences in model parameters across subjects, the choice of method (fast-dm, HDDM, chi-square) is not critical unless the method produces a few spurious results (e.g., HDDM, Figures 5 and 6). The loss in precision (larger SD) due to a less efficient but more robust fitting method could be easily outweighed by adding a few subjects. However, in most situations, the difference between 40 and 50 subjects or 100 and 120 subjects will be not be large.

The conclusions are different for the task of attempting to identify individuals that may show a deficit relative to the normal range of processing. In these experiments, quite large differences would be needed in boundary separation and nondecision time, differences as large as the mean difference between young and old adults (e.g., Ratcliff et al., 2010). However, we should expect deficits to occur in drift rates, for example, poorer memory or perceptual processing for an individual relative to a control group. Results from Experiments 1 and 2 showed that drift rates would have to be near zero for a deficit to be detectable in these two tasks. There is one caveat to this and that is to collect data from a control group that has the same motivation as the experimental group. All the undergraduates in Experiments 1 and 2 are not guaranteed to be highly motivated and it is quite possible that some of the lower-performing subjects were non-cooperative at least on a proportion of trials. Therefore, for detecting deficits, it is important to collect as many trials as possible given constraints of accessibility to the population and testing time (e.g., patients of some kind) and to use the best fitting methods, i.e., those with the lowest SD's and that do not produce spurious results.

In Experiments 1 and 2, differences in model parameters between the first few trials and all the data and also practice effects in model parameters from test block to test block were small. There is no guarantee that such results will be obtained for other populations. Undergraduate students probably adapt to a new task about as quickly as any group. Therefore, practice and training effects would have to be examined for any neuropsychological, clinical or educational population and an appropriate control group, especially if there is a limit on the number of

observations that can be collected.

In some applications such as using emotional or threat words, or applications involving specific psycholinguistic constructions in text processing, there may be limitations on the number of critical materials that are available. In such applications, these critical materials are embedded in much longer lists of filler materials and there is usually no constraint on the number of trials from these fillers. (Often they are used to reduce the proportion of trials with critical materials and/or disguise the hypothesis.) Fitting the diffusion model simultaneously to the critical materials plus the filler materials increases the power on the critical materials (see McKoon & Ratcliff, 2012, 2013; Ratcliff, 2008; White et al., 2010a, 2010b) because boundary settings and nondecision times are largely determined by the larger number of filler materials. This is an important methodological advantage in such applications.

There are some other features of the fitting methods that should be mentioned. First, HDDM has an application in which drift rate can be made a function of, for example, a brain measure such as EEG, fMRI, or some eye tracking measure. In this, drift rate is assumed to be a function of, for example, a EEG regressor and if the function is linear, the method allows the slope of drift rate versus the regressor to be estimated. Second, there are other Bayesian diffusion model tools, specifically a WinBUGS implementation in Vandekerckhove, Tuerlinckx, and Lee (2011) and a JAGS implementation in Wabersich and Vandekerckhove (2014). These allow diffusion model fitting to be implemented with minimal programming, but they do not provide a point and click interface. Third, the latest version of fast-dm has implemented both chi-square and maximum likelihood estimation methods, but this was made available after the simulations and fitting was done.

Recommendations for Software to use in Diffusion Model Fitting

When the number of observations is large, most of the methods produce parameter estimates that are reasonable. The exceptions are the EZ method that can be extremely biased in the presence of contaminants (though it is useful for exploration), DMAT with contaminant correction, and HDDM fit to separate subjects (though this package is still under development and upgrades appear regularly). With smaller numbers of observations, the quantile methods and fast-

dm were reasonably robust, but HDDM was a little better. We found that the hierarchical diffusion method performed very well, and is the method of choice when the number of observations is small. But we would not recommend blindly applying any model to experiments with just a few dozen observations because the data may not be reliable enough because of start up issues, the potential for spurious observations, and practice effects. In general, for any application of the diffusion model, the quality of the data needs to be evaluated.

Author Note

Preparation of this article was supported by NIA grant R01-AG041176 and DOE-IES #R305A120189. We would like to thank EJ Wagenmakers, Joachim Vandekerckhove, Andreas Voss, Thomas Wiecki, and Gail McKoon for comments on this article.

References

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

DeLuca, J., & Kalmar, J.H. (Eds.) (2008). *Information processing speed in clinical populations*. New York, NY: Taylor & Francis.

Diederich, A., & Busemeyer, J.R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time and simple response time. *Journal of Mathematical Psychology, 47*, 304-322.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin and Review, 16*, 1026-1036.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*, 119-127.

Farrell, S., & Ludwig, C. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review, 15*, 1209-1217.

Fific, M., Little, T.D., & Nosofsky, R.M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117*, 309-348.

Geddes, J., Ratcliff, R., Allerhand, M., Childers, R., Wright, R. J., Frier, B. M., & Deary, I. J. (2010). Modeling the effects of hypoglycemia on a two-choice task in adult humans. *Neuropsychology, 24*, 652-660.

Gold, J.I., & Shadlen, M.N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574.

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53*, 55-68.

Heathcote, A., Brown, S. & Mewhort, D.J.K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin and Review, 9*, 394-401.

Jeffreys, H. (1961). *Theory of probability* (3rd edition). Oxford: Oxford University Press.

Karalunas, K. & Huang-Pollock, C. (2013). Integrating evidence of slow reaction times and impaired executive function using a diffusion model framework. *Journal of Abnormal Child Psychology, 41*, 837-850.

Kuhn, S., Schmiedek, F., Schott, B., Ratcliff, R., Heinze, H-J., Duzel, E., Lindenberger, U., and Levden, M. (2011). Brain areas consistently linked to individual differences in perceptual decision-making in younger as well as older adults before and after training. *Journal of Cognitive Neuroscience, 23*, 2147-2158.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*, 651-687.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin and Review, 16*, 798-817.

Menz, M. M., Buchel, C., & Peters, J. (2012). Sleep Deprivation Is Associated with Attenuated Parametric Valuation and Control Signals in the Midbrain during Value-Based Decision Making. *The Journal of Neuroscience, 32*, 6937-6946.

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 82-91.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440-466.

McKoon, G., & Ratcliff, R. (2012). Aging and IQ effects on associative recognition and priming in item recognition. *Journal of Memory and Language, 66*, 416-437.

McKoon, G., & Ratcliff, R. (2013). Aging and predicting inferences: A diffusion model analysis. *Journal of Memory and Language, 68*, 240-254.

Mulder, M.J., Bos, D., Weusten, J.M.H., van Belle, J., van Dijk, S.C., Simen, P., van Engeland, H., & Durson, S. (2010). Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry, 68*, 1114-1119.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*,

7, 308-313.

Petrov, A. A., Van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual learning mechanisms revealed by diffusion-model analysis. *Psychonomic Bulletin and Review*, *18*, 490-497.

Philiastides, M.G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience, 26*, 8965-8975.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446-461.

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 212-225.

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. (2006). Modeling Response Signal and Response Time Data, *Cognitive Psychology*, *53*, 195-237.

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin and Review*, *15*, 1218-1228.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review, 120*, 281-292.

Ratcliff, R. (in press). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance.*

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. *Journal of Neurophysiology, 90*, 1392-1407.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical-decision

task. *Psychological Review, 111,* 159-182.

Ratcliff, R., Hasegawa, Y.T., Hasegawa, Y.P., Smith, P.L., & Segraves, M.A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97,* 1756-1774.

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses, *Child Development, 83,* 367-381.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873-922.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain & Cognition, 55,* 374-382.

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences, 106,* 6539-6544.

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science, 9,* 347-356.

Ratcliff, R. & Smith, P.L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333-367.

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.

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. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception and Psychophysics, 65,* 523-535.

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. (2006). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology and Aging, 21*, 353-371.

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. (submitted). Modeling differences among individuals in numeracy.

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. & Van Dongen, H.P.A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin and Review*, *16*, 742-751.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261-300.

Schmiedek, F., Oberauer, K., Wilhelm, O., Suβ, H-M., & Wittmann, W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414-429.

Sheppard, L.D. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences, 44*, 533-549.

Smith, P.L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408-463.

Smith, P.L., Ratcliff, R., & McKoon, G. (in press). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2013). *Psychological Review*.

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.

Starns, J.J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 5,* 377-390.

Starns, J.J., & Ratcliff, R. (2012). Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. *Psychonomic Bulletin and Review, 19,* 139-145.

Starns, J.J., Ratcliff, R., & McKoon, G. (2012). Modeling single versus multiple systems in implicit and explicit memory. *Trends in Cognitive Sciences, 16,* 195-196.

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.

van Ravenzwaaij, D. & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53,* 463-473.

Vandekerckhove, J. & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011-1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40,* 61-72.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16,* 44-62.

Vandekerckhove, J., Verheyen, S, & Tuerlinckx, F. (2010). A crossed random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica, 133,* 269-282.

Voss, A. & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767-775.

Voss, A., & Voss, J. (2008). A Fast Numerical Algorithm for the Estimation of Diffusion-Model Parameters. *Journal of Mathematical Psychology, 52,* 1-9.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46,* 15-

28.

Wagenmakers, E-J., Ratcliff, R., Gomez, P. & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28-50.

Wagenmakers, E-J., van der Maas, H. L. J. & Grassman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin and Review*, *14*, 3-22.

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*, 1229-1235.

White, C., Ratcliff, R., Vasey, M. & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion*, *23*, 181-205.

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*, 39-52.

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion*, *10*, 662-677.

Wiecki, T.V., Sofer, I. and Frank, M.J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 1-10.

Zeguers, M.H.T., Snellings, P., Tijms, J., Weeda, W.D., Tamboer, P., Bexkens, A. & Huizenga, H.M. (2011). Specifying theories of developmental dyslexia: A diffusion model analysis of word recognition. *Developmental Science*, *14*, 1340-1354.

Figure Captions

Figure 1. Plots of the mean values of model parameters across subjects along with the SDs and SEs across subjects for several divisions of the data from Experiment 1. The larger error bars are 2 SD confidence intervals in model parameters and the smaller error bars are 2 SE confidence intervals.

Figure 2. Plots of the mean values of model parameters across subjects along with the SDs and SEs across subjects for several divisions of the data from Experiment 2. The larger error bars are 2 SD confidence intervals in model parameters and the smaller error bars are 2 SE confidence intervals.

Figure 3. Plots of parameter values recovered from fitting the diffusion model to simulated data from Simulation Study 1 plotted against the model parameters used to generate the simulated data. Plots are for boundary separation, nondecision time, and the highest drift rate for the numeracy design with 1000 observations per condition. The different columns are for the chi-square method with 9 quantiles and the EZ method with either 0 or 4% contaminants. The correlations are shown in the top right corner of each plot and the diagonal line has slope 1 and intercept zero.

Figure 4. Plots of parameter values recovered from fitting the diffusion model to simulated data from Simulation Study 1 plotted against the model parameters used to generate the simulated data. Plots are for boundary separation, nondecision time, and the highest drift rate for the numeracy design with 40 observations per condition. The different columns are for the chi-square method with 9 quantiles and the EZ method with either 0 or 4% contaminants. The correlations are shown in the top right corner of each plot and the diagonal line has slope 1 and intercept zero.

Figure 5. Plots of boundary separation, nondecision time, and two drift rates for the numeracy design with 1000 observations per condition in Simulation Study 2. Each row shows a different fitting method. On the x-axis is plotted the mean of the parameter values and on the y-axis, in a horizontal row at the value of the parameter used to generate the simulated data is plotted 1 SD error bars. The thin vertical line represents the values used to generate the simulated data. Movement away from the vertical line on the x-axis represents bias in the recovered parameter

values and a large spread of the error bars represents high variability in the recovered parameter values. There are two values of boundary separation and two values of drift rate, hence the two vertical lines and the vertical separation of the points.

Figure 6. The same plot as in Figure 5, but for the lexical decision design with 300, 300, and 600 observations for the high and low frequency words and nonwords respectively for Simulation Study 2.

Figure 7. The same plot as in Figure 5, but for the numeracy design with 40 observations per condition in Simulation Study 2. The DMAT results had SDs that overlapped zero and so are not shown.

Figure 8. The same plot as in Figure 5, but for the lexical decision design with 30, 30, and 60 observations for the high and low frequency words and nonwords respectively for Simulation Study 2. As for Figure 7, the DMAT results had SDs that overlapped zero and so are not shown.

Figure 9. Plots of the recovered values of parameters from the hierarchical fitting method and for the chi-square method for the numeracy design with normally distributed population parameters for 40 and 100 observations per condition and for 0 and 4% contaminants. Plotted on the y-axis are the recovered parameter values minus the values used to generate the simulated data (i.e., the residuals) offset by the amount represented by the thick horizontal lines. The circles are for the chi-square method and the crosses are for the hierarchical method. The dashed lines are regression lines for the two methods (shrinkage of model parameters results in a negative slope). The numbers at the top of each plot are the correlations between the recovered values and those used to generate the simulated data for the two methods.

Figure 10. Plots of the recovered values of parameters from the hierarchical fitting method and for the chi-square method for the numeracy design with uniformly distributed parameters. Other details of the plots are the same as for Figure 9.

Figure 11. Plots of the recovered values of parameters from the hierarchical fitting method and for the chi-square method for the numeracy design with parameters drawn from two distributions. Other details of the plots are the same as for Figure 9.

Figure 12. Plots of correlations between the recovered values of parameters and the values

used to generate the simulated data as a function of the number of observations for the three distributions in Figures 9, 10, and 11, for the two methods, and for simulated data with 0% and 4% contaminants. The oblique arrows represent the approximate average number of observations at which the correlation for the chi-square method becomes larger than that for the hierarchical method.

Figure 13. Plots of the recovered values of parameters from the hierarchical fitting method for the numeracy design with normally distributed population parameters for 5 observations per condition. The plots are otherwise the same as for Figure 9.

Figure A1. Plots of boundary separation, nondecision time, and two drift rates for the numeracy design with 1000 observations per condition in Simulation Study 2 as in Figure 5. The different rows show the original HDDM parameter recovery and two suggested fits (which help a lot).

Table 1: Mean Parameter Values and SD's Across Subjects for the Numerosity Discrimination Experiment.

	Trial group	a	T_{er}	η	s_z	p_0	s_t	v_E	v_D	χ^2
Mean	1-80	0.126	0.409	0.171	0.050	0.010	0.161	0.321	0.106	25.1
	81-160	0.114	0.411	0.157	0.051	0.006	0.158	0.401	0.128	22.2
	161-240	0.118	0.400	0.138	0.051	0.009	0.172	0.383	0.127	24.5
	1-160	0.119	0.402	0.162	0.050	0.007	0.161	0.314	0.100	29.6
	161-320	0.114	0.398	0.145	0.051	0.009	0.171	0.340	0.115	33.0
	1-320	0.118	0.395	0.155	0.053	0.008	0.182	0.317	0.114	34.1
	321-640	0.114	0.382	0.145	0.063	0.008	0.174	0.324	0.108	34.6
	1-1200	0.112	0.374	0.126	0.064	0.007	0.180	0.282	0.107	50.4
SD	1-80	0.032	0.058	0.076	0.036	0.023	0.087	0.113	0.058	10.5
	81-160	0.036	0.052	0.073	0.037	0.015	0.091	0.147	0.076	10.6
	161-240	0.032	0.043	0.070	0.043	0.018	0.080	0.150	0.081	11.4
	1-160	0.026	0.047	0.075	0.036	0.019	0.078	0.107	0.039	11.0
	161-320	0.031	0.042	0.078	0.034	0.021	0.076	0.126	0.052	15.0
	1-320	0.027	0.039	0.067	0.031	0.018	0.064	0.093	0.038	13.1
	321-640	0.029	0.041	0.061	0.027	0.015	0.058	0.104	0.035	12.0
	1-1200	0.023	0.035	0.047	0.021	0.011	0.049	0.076	0.033	20.7
Mean	Simulation Parameters	0.11	0.375	0.13	0.06		0.16	0.30	0.10	
SD		0.03	0.05	0.08	0.01		0.10	0.10	0.05	
Mean	Simulation parameters: hierarchical study with two subject groups	0.20	0.53	0.13	0.06		0.16	0.10	0.00	
SD		0.02	0.03	0.08	0.01		0.10	0.05	0.025	

Note. a=boundary separation, z=starting point, T_{er}=nondecision component of response time, η =standard deviation in drift across trials, s_z=range of the distribution of starting point (z), s_t = range of the distribution of nondecision times, v_E and v_D are the drift rates for the easy and difficult conditions respectively, and χ^2 is the chi-square goodness of fit measure. For the data, the number of degrees of freedom for the fits to data was 30 and the critical value of chi-square at the .05 level was 43.8. In the simulations, the drift rates were correlated. The same random number was used to generate both drift rates with half the value added to the lower drift rate. The following were lower limits on parameter values: a, 0.07; η, 0.02; T_{er}, 0.25; s_t, 0.04, s_z, 0.9a. Separate sets of simulations were conducted with p_o 0 or 0.04.

Table 2: Mean Parameter Values and SD's Across Subjects for the Lexical Decision Experiment.

	Trial group	a	T_{er}	η	s_z	p_0	s_t	z	v_H	v_L	v_V	v_N	χ^2
Mean	1-120	0.153	0.476	0.133	0.038	0.001	0.171	0.081	0.660	0.271	0.159	-0.240	43.0
	121-240	0.162	0.462	0.150	0.042	0.002	0.175	0.085	0.626	0.323	0.186	-0.259	41.2
	241-360	0.153	0.457	0.122	0.062	0.003	0.153	0.081	0.618	0.290	0.165	-0.233	36.7
	1-240	0.148	0.471	0.139	0.040	0.003	0.182	0.075	0.581	0.272	0.147	-0.219	63.3
	241-480	0.144	0.454	0.125	0.051	0.002	0.173	0.083	0.516	0.244	0.135	-0.204	61.0
	1-480	0.143	0.458	0.132	0.045	0.001	0.193	0.083	0.502	0.237	0.122	-0.201	82.3
	481-960	0.138	0.445	0.144	0.054	0.002	0.167	0.076	0.431	0.206	0.104	-0.221	83.2
	1-2100	0.135	0.436	0.123	0.066	0.002	0.178	0.080	0.399	0.203	0.114	-0.216	158.
SD	1-120	0.041	0.062	0.052	0.037	0.003	0.078	0.028	0.185	0.113	0.087	0.092	20.7
	121-240	0.054	0.052	0.057	0.039	0.006	0.084	0.040	0.176	0.150	0.114	0.100	18.2
	241-360	0.035	0.044	0.045	0.039	0.009	0.084	0.039	0.176	0.136	0.122	0.068	14.0
	1-240	0.040	0.055	0.055	0.036	0.011	0.070	0.028	0.175	0.101	0.077	0.073	20.4
	241-480	0.037	0.040	0.048	0.040	0.005	0.071	0.027	0.164	0.099	0.080	0.055	24.1
	1-480	0.033	0.040	0.047	0.033	0.001	0.059	0.028	0.141	0.087	0.070	0.047	26.1
	481-960	0.030	0.033	0.049	0.030	0.006	0.062	0.024	0.106	0.078	0.069	0.052	25.0
	1-2100	0.025	0.029	0.036	0.025	0.003	0.049	0.023	0.096	0.063	0.052	0.041	51.5
Mean	Simula-tion parame-ters	0.14	0.44	0.115	0.04		0.17	0.080	0.38	0.19	0.11	-0.20	
SD		0.04	0.04	0.06	0.02		0.06	0.002	0.06	0.06	0.06	0.06	

Note. a=boundary separation, z=starting point, T_{er}=nondecision component of response time, η =standard deviation in drift across trials, s_z=range of the distribution of starting point (z), s_t = range of the distribution of nondecision times, v_H, v_L, v_V, and v_N are the drift rates for high, low, and very low frequency words and for nonwords respectively, and χ^2 is the chi-square goodness of fit measure. The number of degrees of freedom for the fits to data was 65 and the critical value of chi-square at the .05 level was 84.8. In the simulations, the drift rates were not correlated. The following were lower limits on parameter values: a, 0.07; η, 0.02; T_{er}, 0.25; s_t, 0.04, s_z, 1.8 the distance from the starting point to the nearest boundary. Separate sets of simulations were conducted with p_o 0 or 0.04. Each drift rate had the same random number added to it so if a subject had a high drift rate in one condition they had a high drift rate in the other conditions.

Table 3: Power analyses showing the value of the parameter needed to detect a score outside the normal range 90% and 95% of the time

Task	Para-meter	Parame-ter value	SD in paramet-er value	Paramet er for 90% correct	Paramet er for 95% correct	SD in paramet -er value	Paramet er for 90% correct	Paramet er for 95% correct
Numer-osity	a	0.110	0.030	0.187	0.208	0.045	0.225	0.258
	T_{er}	0.400	0.045	0.515	0.548	0.068	0.573	0.622
	v_D	0.110	0.035	0.020	0	0.053	0	0
Lexical decision	a	0.150	0.040	0.253	0.282	0.060	0.304	0.347
	T_{er}	0.450	0.040	0.553	0.582	0.060	0.604	0.647
	v_H	0.460	0.122	0.147	0.058	0.183	0	0
	v_N	0.200	0.050	0.072	0.036	0.075	0.008	0

Note. Parameter refers to the population parameter value, SD is the standard deviation in the population parameter value. a is boundary separation, T_{er} is nondecision component of response time, v_H, and v_N are the drift rates for high frequency words and for nonwords respectively. v_L and v_V (drift rates for low and very low frequency words) were not included because there was no value greater than 0 for either 90% or 95% correct detection.

Table 4: Correlations between model parameters for fits to small numbers of trials with model parameters from fits to all the trials.

Task	Trial block	a	T_{er}	v_1	v_2	v_3	v_4
Numerosity	1-80	0.278	0.675	0.247	0.292		
	81-160	0.735	0.578	0.304	0.304		
	161-240	0.687	0.532	0.260	0.306		
	1-160	0.617	0.718	0.377	0.404		
	161-320	0.773	0.628	0.546	0.438		
	1-320	0.733	0.783	0.538	0.467		
	321-640	0.877	0.852	0.725	0.543		
Lexical Decision	1-120	0.704	0.515	0.357	0.526	0.474	0.467
	121-240	0.715	0.487	0.347	0.568	0.476	0.497
	241-360	0.809	0.670	0.359	0.584	0.654	0.447
	1-240	0.795	0.597	0.424	0.625	0.611	0.438
	241-480	0.871	0.760	0.363	0.609	0.731	0.394
	1-480	0.860	0.777	0.525	0.711	0.730	0.485
	481-960	0.889	0.802	0.408	0.722	0.844	0.473

Note: $v_1=v_E$ and $v_2=v_D$ for numerosity (v_E and v_D are the drift rates for the easy and difficult conditions respectively), and $v_1=v_H$, $v_2=v_L$, $v_3=v_V$, and $v_4=v_N$ for lexical decision (v_H, v_L, v_V, and v_N are the drift rates for high, low, and very low frequency words and for nonwords respectively). a is boundary separation and T_{er} is nondecision component of response time.

Table 5: Correlations between parameters used to generate simulated data and recovered parameters for 8 fitting methods and three sets of numbers of observations for the numerosity discrimination design.

		N=1000,1000				N=100,100				N=40,40			
	Method	a	T_{er}	v_1	v_2	a	T_{er}	v_1	v_2	a	T_{er}	v_1	v_2
0% contaminants	HDDM	0.991	0.972	0.952	0.960	0.782	0.877	0.356	0.571	0.426	0.791	-0.129	0.154
	Fast-dm	0.971	0.971	0.932	0.926	0.893	0.910	0.843	0.774	0.878	0.851	0.573	0.605
	DMATout	0.913	0.888	0.810	0.921	0.755	0.808	0.265	0.591	0.493	0.719	-0.041	0.366
	DMATno	0.980	0.957	0.902	0.916	0.798	0.854	0.586	0.651	0.379	0.650	0.091	0.337
	MLH	0.987	0.977	0.937	0.894	0.954	0.923	0.835	0.647	0.923	0.890	0.743	0.665
	Chi9q	0.982	0.961	0.908	0.934	0.943	0.942	0.829	0.743	0.899	0.852	0.668	0.643
	Chi5q	0.979	0.961	0.921	0.951	0.948	0.917	0.822	0.746	0.899	0.871	0.650	0.636
	EZ	0.976	0.928	0.926	0.929	0.896	0.804	0.866	0.798	0.821	0.745	0.741	0.670
4% contaminants	HDDM	0.667	0.939	0.602	0.752	-0.126	0.880	0.181	0.360	0.478	0.790	0.089	0.332
	Fast-dm	0.946	0.968	0.786	0.838	0.876	0.859	0.783	0.684	0.873	0.847	0.629	0.546
	DMATout	0.930	0.930	0.674	0.838	0.525	0.771	0.221	0.494	0.352	0.666	-0.026	0.471
	DMATno	0.862	0.941	0.649	0.745	0.619	0.765	0.597	0.626	0.333	0.626	0.115	0.405
	MLH	0.975	0.975	0.844	0.830	0.894	0.872	0.755	0.636	0.763	0.862	0.599	0.613
	Chi9q	0.987	0.963	0.910	0.932	0.941	0.935	0.848	0.765	0.867	0.866	0.629	0.516
	Chi5q	0.967	0.938	0.854	0.903	0.930	0.907	0.807	0.786	0.835	0.858	0.691	0.594
	EZ	0.804	0.760	0.766	0.849	0.675	0.640	0.814	0.778	0.643	0.575	0.611	0.623

Note: DMATout was the DMAT method with contaminant correction, DMATno was the DMAT method with no contaminant correction, MLH was the maximum likelihood method, chi9q and chi5q were the chi-square methods with 9 and 5 quantiles respecitvely. a is boundary separation, T_{er} is nondecision component of response time, v_1 and v_2 are the drift rates for the easy and difficult conditions respectively.

Table 6: Correlations between parameters used to generate simulated data and recovered parameters for 8 fitting methods and two sets of numbers of observations for the lexical decision design.

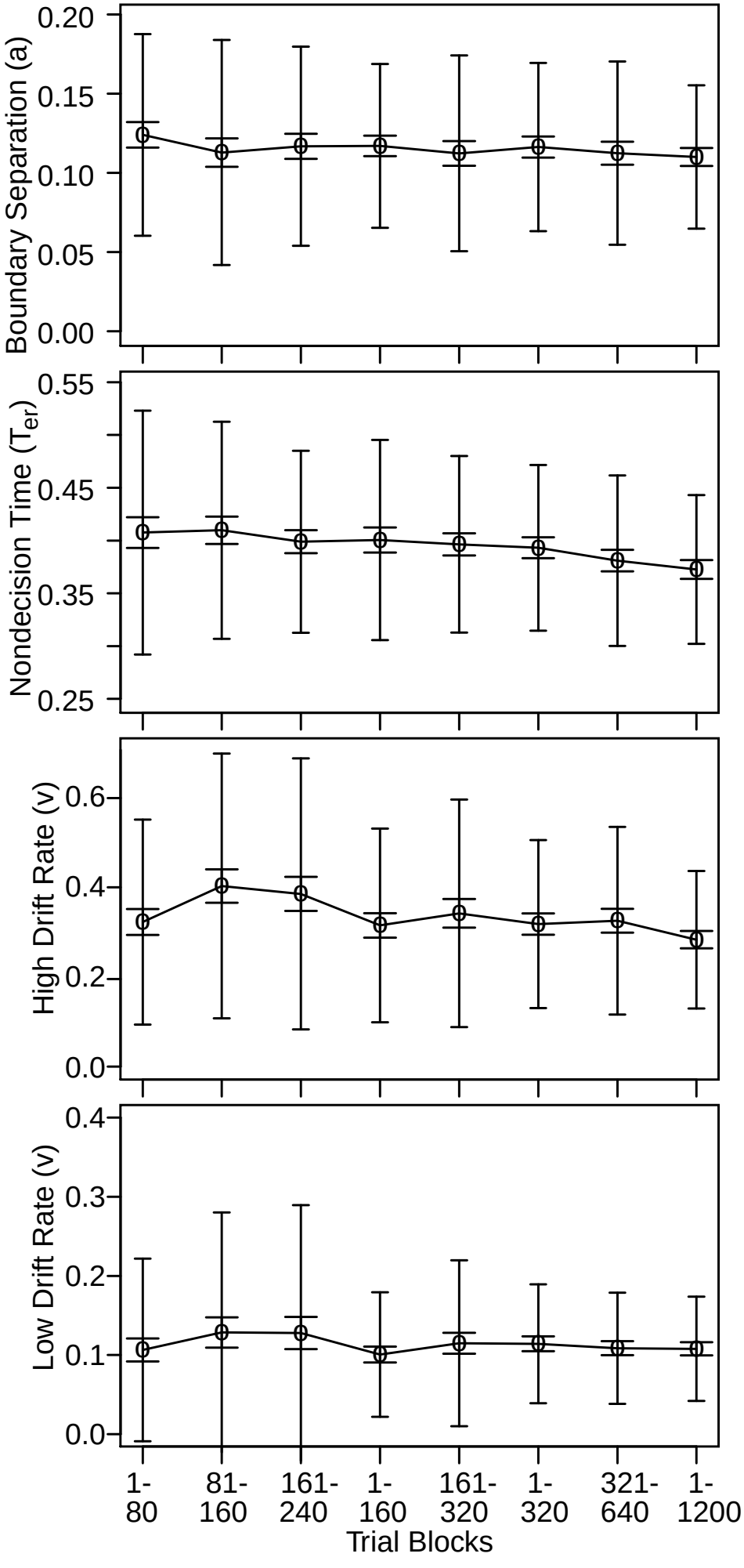
		N=200,200,200,600						N=20,20,20,60					
	Method	a	T_{er}	v_1	v_2	v_3	v_4	a	T_{er}	v_1	v_2	v_3	v_4
0% contaminants	HDDM	0.988	0.980	0.829	0.924	0.951	0.947	0.887	0.844	0.379	0.726	0.633	0.652
	Fast-dm	0.946	0.955	0.682	0.858	0.920	0.878	0.757	0.509	0.362	0.554	0.565	0.564
	DMATout	0.826	0.853	0.395	0.793	0.913	0.757	0.429	0.428	0.129	0.144	0.224	0.066
	DMATno	0.918	0.913	0.533	0.767	0.908	0.793	0.253	0.333	0.086	0.202	0.263	0.375
	MLH	0.979	0.965	0.692	0.805	0.850	0.913	0.898	0.761	0.385	0.566	0.589	0.535
	Chi9q	0.983	0.964	0.768	0.902	0.931	0.929	0.757	0.730	0.375	0.682	0.573	0.666
	Chi5q	0.979	0.953	0.752	0.910	0.937	0.915	0.814	0.751	0.366	0.663	0.599	0.674
	EZ	0.952	0.834	0.608	0.809	0.920	0.750	0.646	0.346	0.405	0.680	0.740	0.325
4% contaminants	HDDM	0.950	0.971	0.721	0.847	0.948	0.841	0.937	0.869	0.517	0.735	0.661	0.711
	Fast-dm	0.911	0.920	0.604	0.844	0.881	0.830	0.838	0.633	0.471	0.540	0.536	0.598
	DMATout	0.889	0.853	0.287	0.648	0.829	0.343	0.446	0.446	0.110	0.150	0.215	0.416
	DMATno	0.940	0.933	0.573	0.790	0.916	0.727	0.273	0.410	0.142	0.136	0.309	0.437
	MLH	0.975	0.975	0.588	0.801	0.779	0.905	0.922	0.811	0.493	0.562	0.712	0.571
	Chi9q	0.978	0.969	0.662	0.908	0.962	0.902	0.838	0.847	0.375	0.590	0.598	0.678
	Chi5q	0.977	0.969	0.717	0.926	0.962	0.924	0.893	0.855	0.388	0.597	0.613	0.715
	EZ	0.892	0.743	0.352	0.756	0.897	0.791	0.789	0.393	0.505	0.716	0.719	0.262

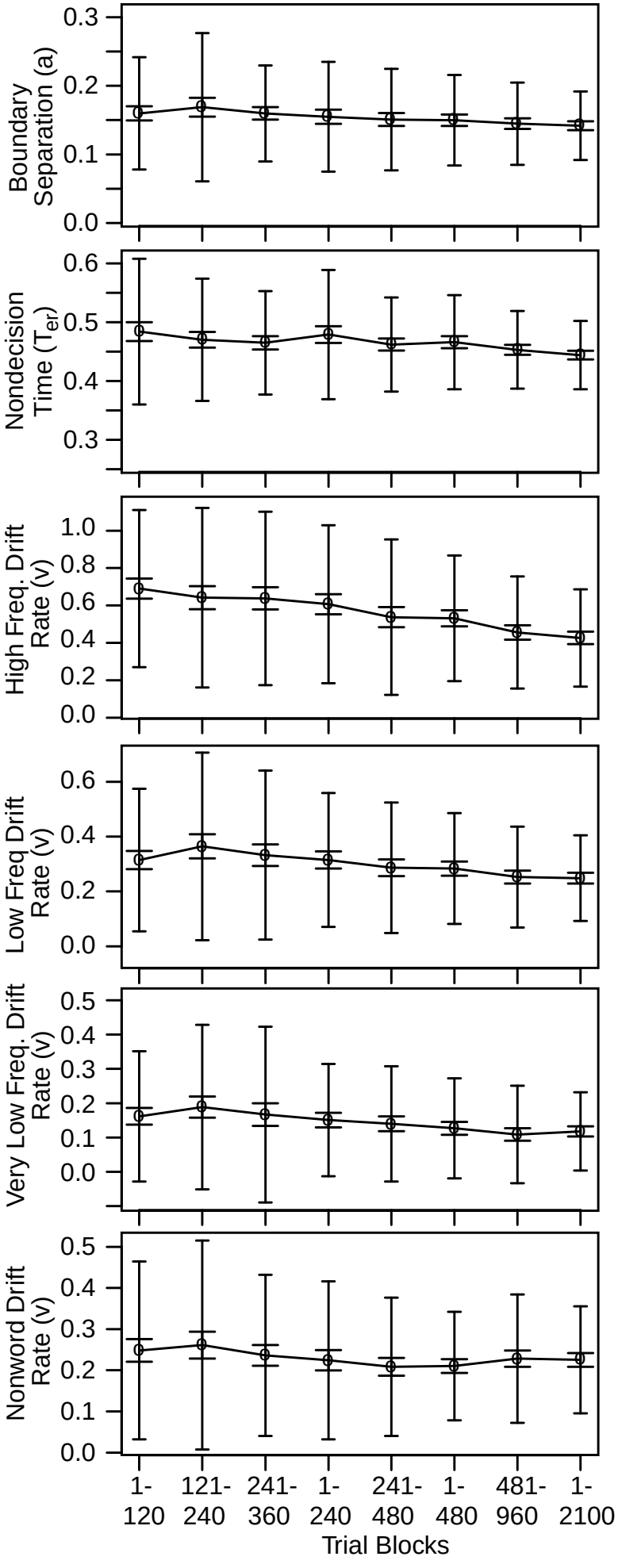
Note: DMATout was the DMAT method with contaminant correction, DMATno was the DMAT method with no contaminant correction, MLH was the maximum likelihood method, chi9q and chi5q were the chi-square methods with 9 and 5 quantiles respecitvely. a is boundary separation, T_{er} is nondecision component of response time, v_1, v_2, v_3, and v_4 are the drift rates for high, low, and very low frequency words and for nonwords respectively.

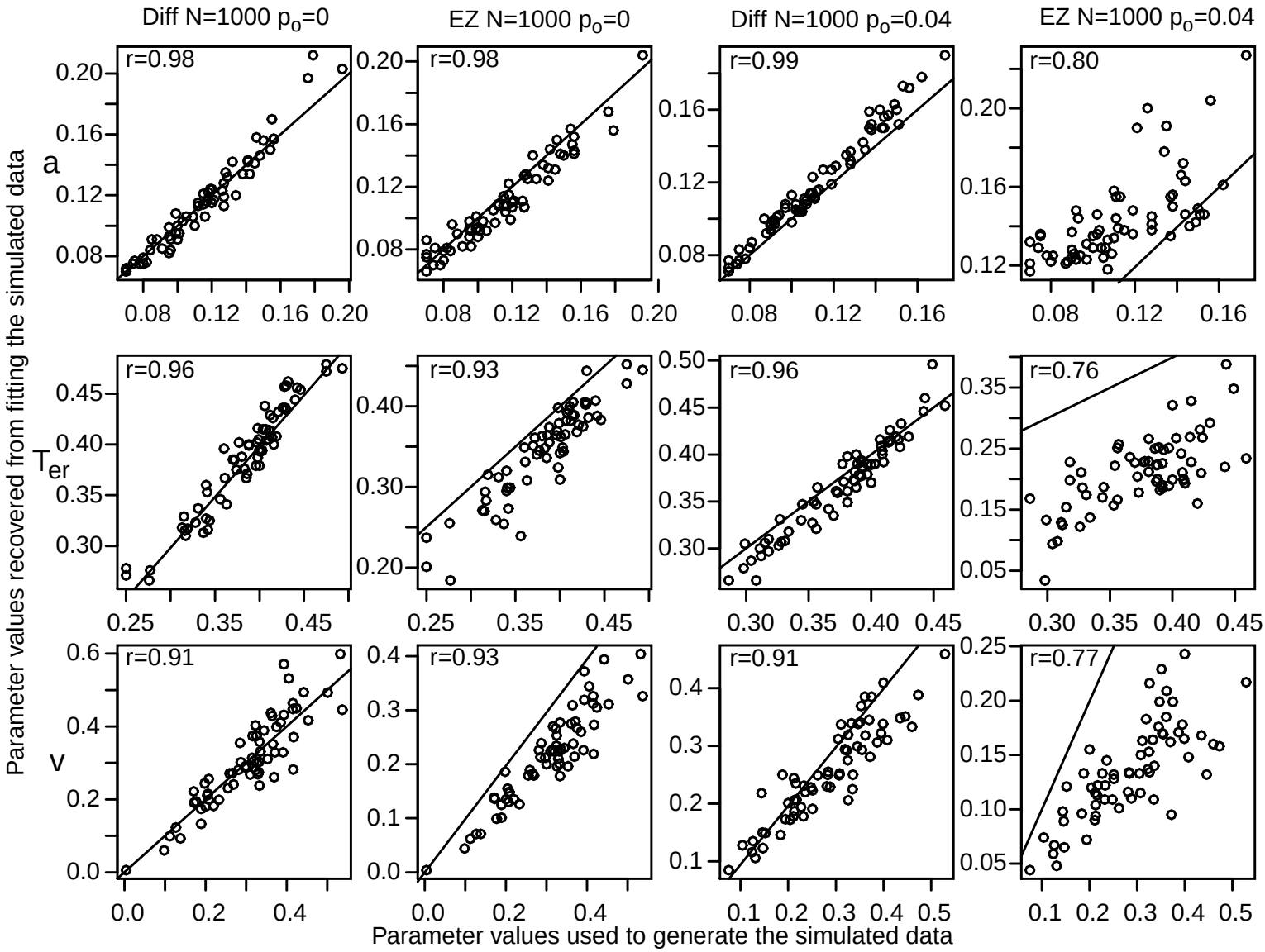
Table 7: Parameter values used in simulations to examine accuracy and bias in parameter recovery.

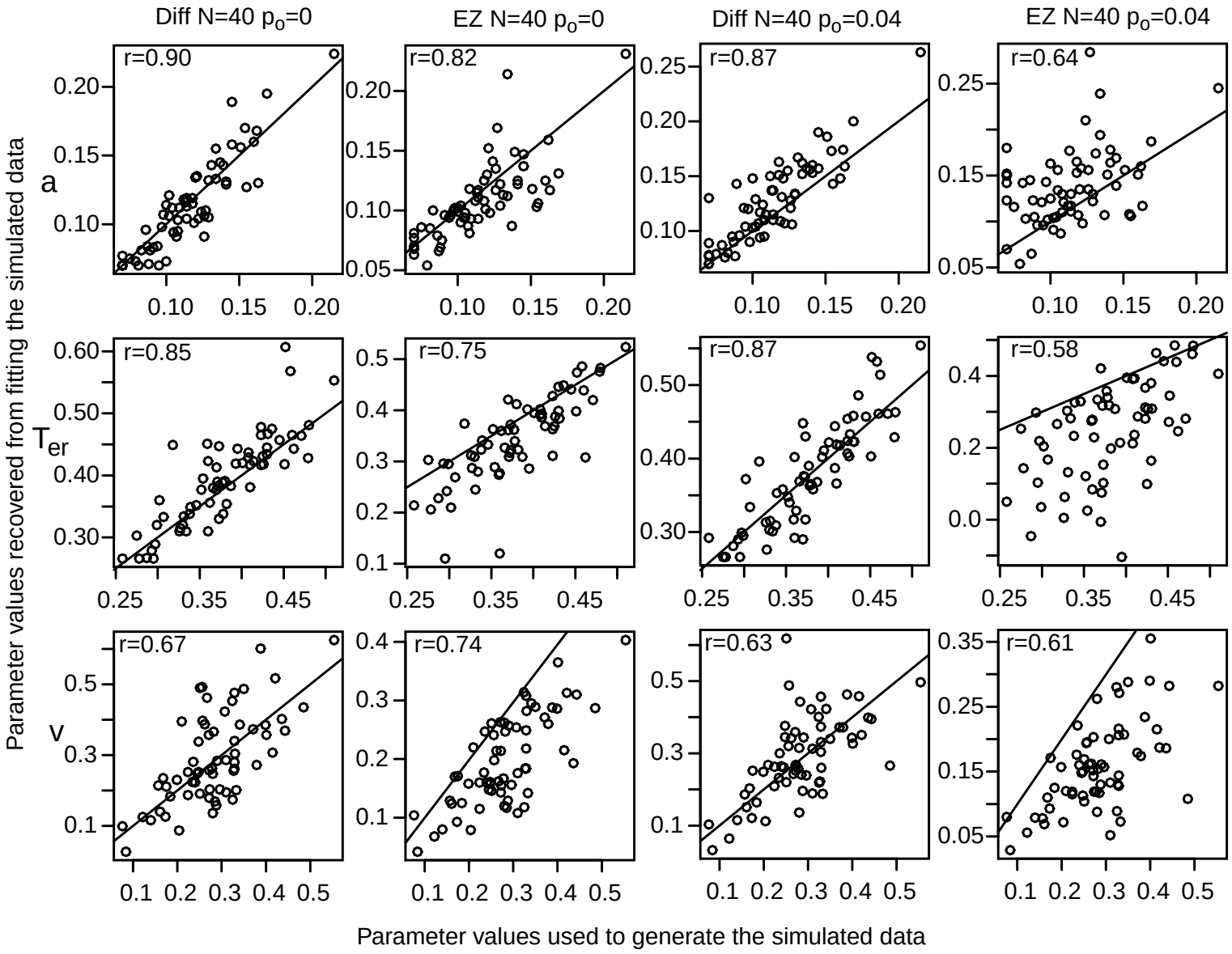
Simulation	a	T_{er}	η	s_z	s_t	z/a	v_1	v_2	v_3	p_o
Set 1	0.1	0.4	0.1	0.02	0.15	0.5	0.2	0.1		0.00
	0.2		0.2	0.06						0.04
				0.08						
Set 2	0.1	0.4	0.1	0.02	0.15	0.3	0.2	0.1	-0.2	0.00
	0.2		0.2	0.06		0.5				0.04
				0.08		0.7				

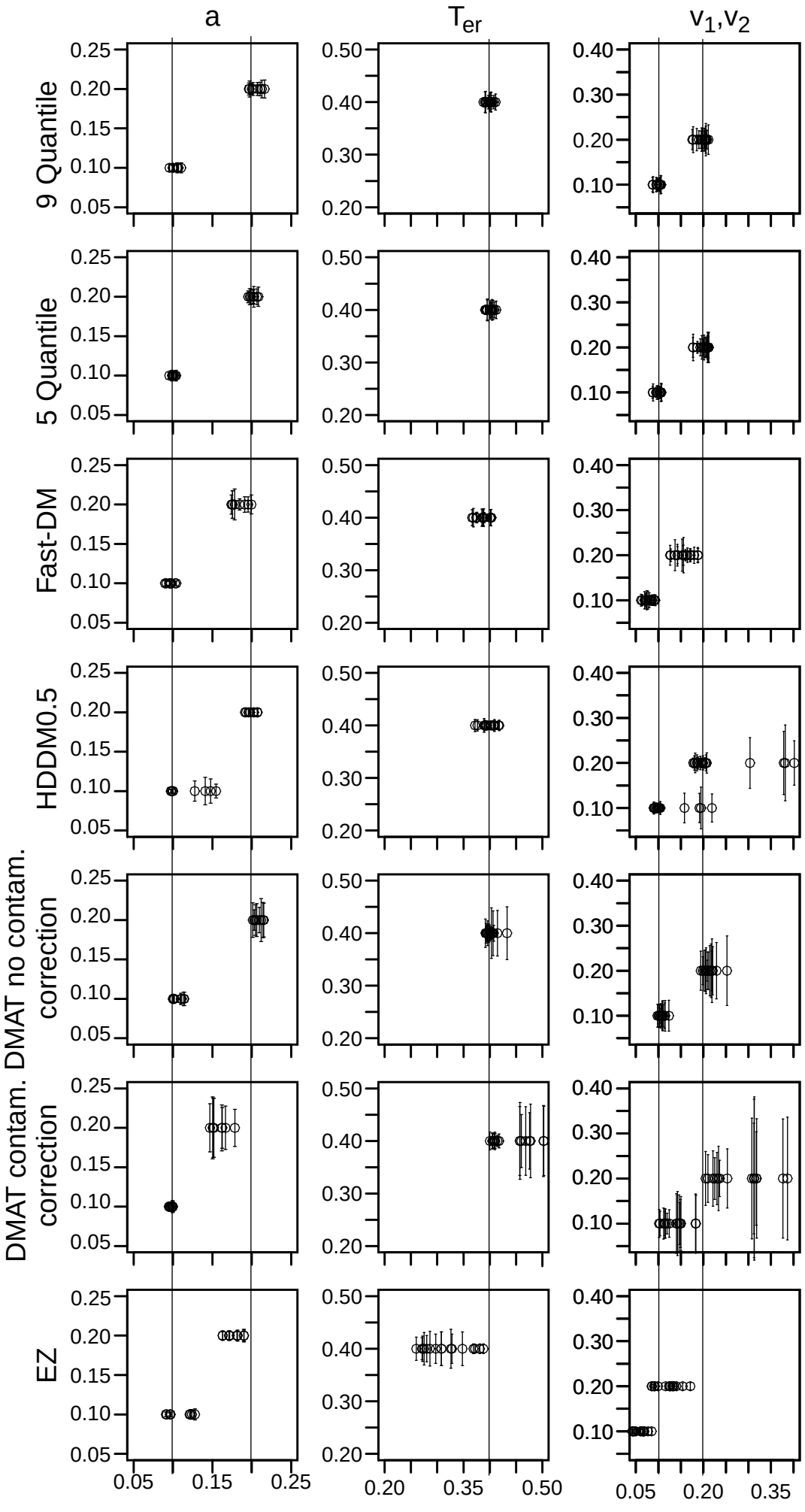
Note. The parameter sets are produced by the combinations of the parameters, a, η, z/a, and p_o. The combinations involving s_z depend on the values of a and z/a. For parameter set 1, when a is 0.1, s_z is 0.02 or 0.06, when a is 0.2, s_z is 0.02 or 0.08. For parameter set 2, for all combinations of the other parameters, one set has s_z=0.02. When z/a=0.5 and a=0.1, one parameter set has s_z=0.08 and when z/a=0.5 and a=0.2, another parameter set has s_z=0.08. When z/a=0.3 or z/a=0.7, the only value of s_z is 0.02.

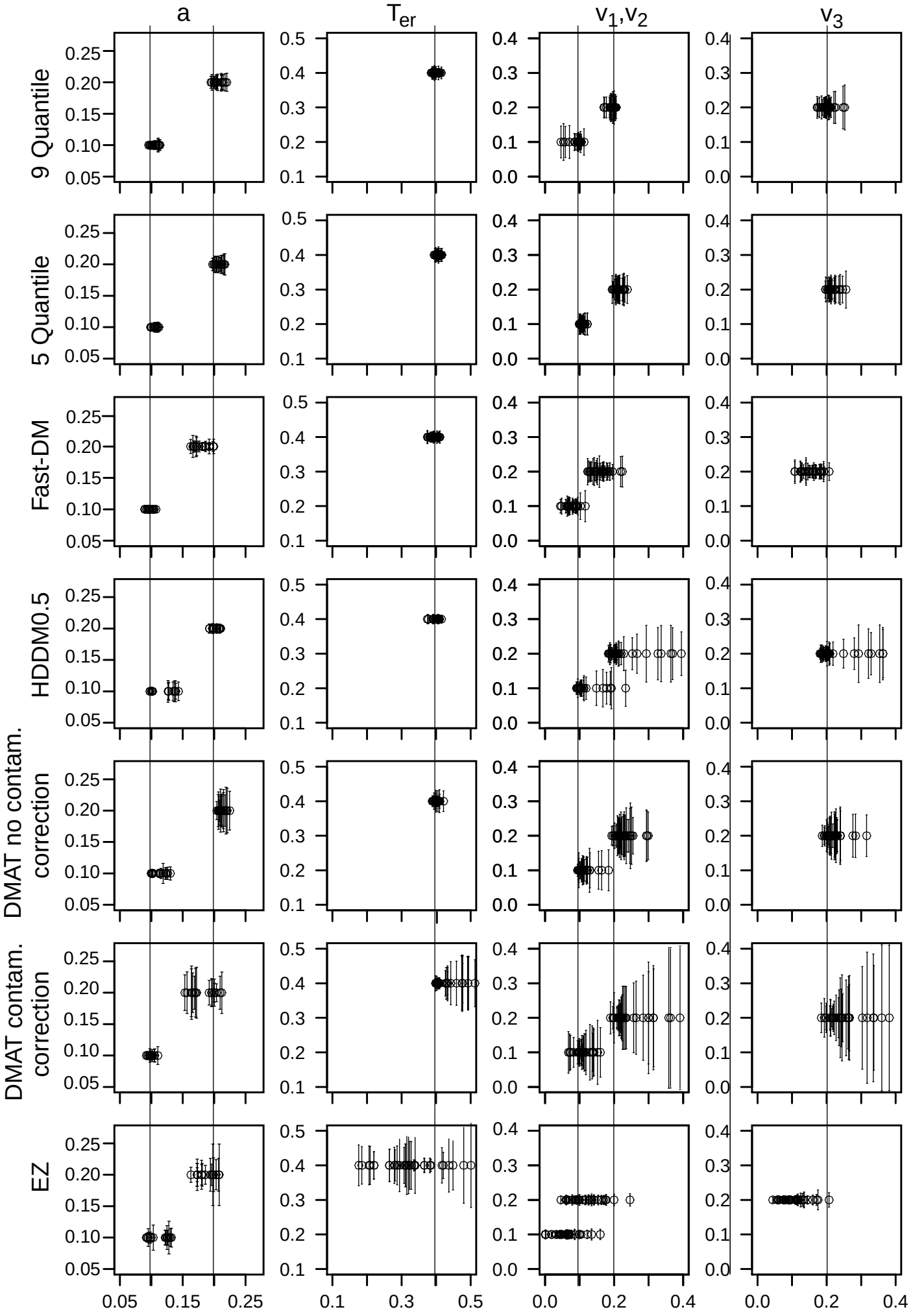

Boundary Separation (a)
0.20
0.15
0.10
0.05
0.00
Nondecision Time (T_{er})
0.55
0.45
0.35
0.25
High Drift Rate (v)
0.6
0.4
0.2
0.0
Low Drift Rate (v)
0.4
0.3
0.2
0.1
0.0
1-80
81-160
161-240
1-160
161-320
1-320
321-640
1-1200
Trial Blocks

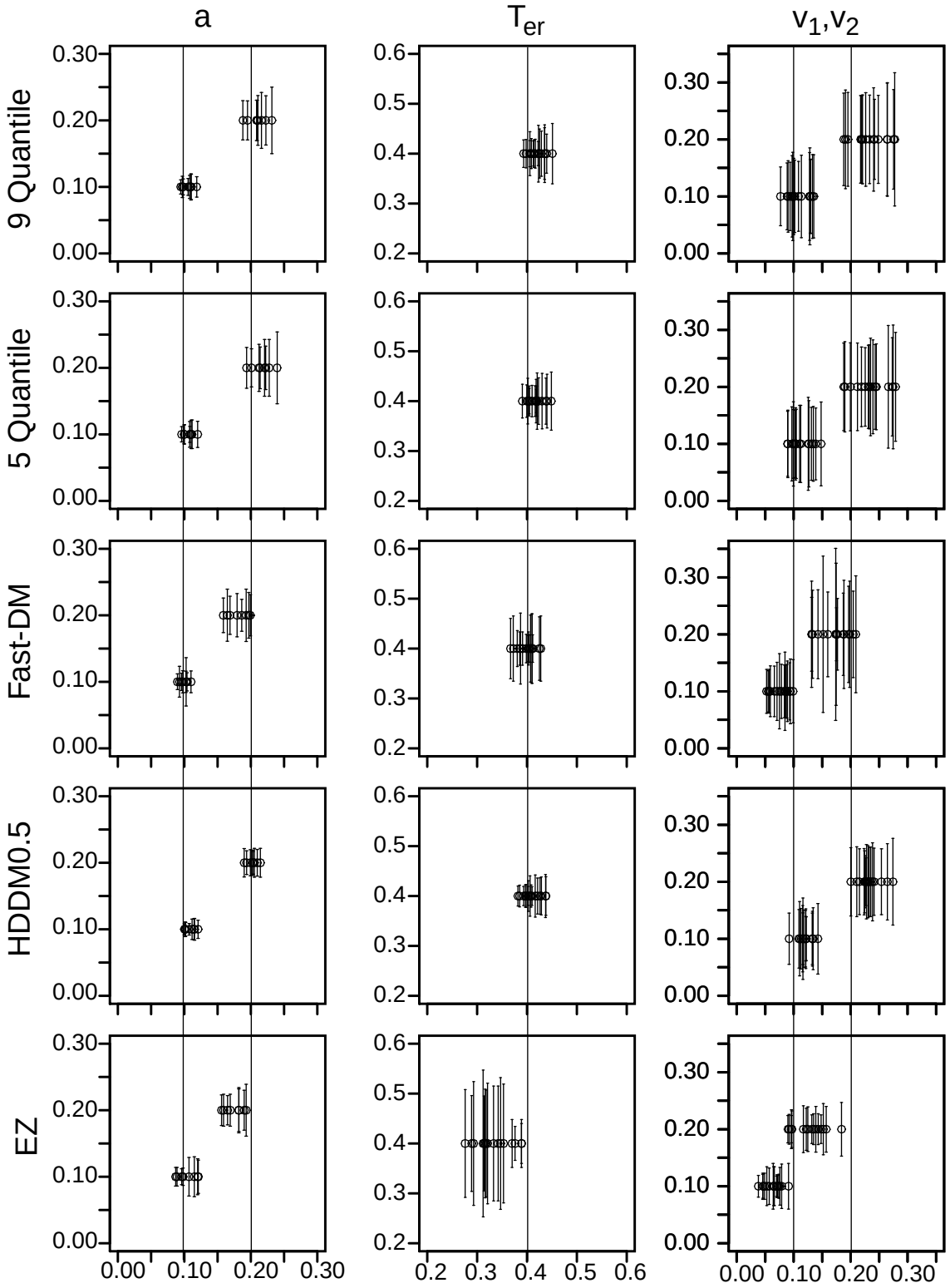

Boundary Separation (a)
0.0
0.1
0.2
0.3
Nondecision Time (T_{er})
0.3
0.4
0.5
0.6
High Freq. Drift Rate (v)
0.0
0.2
0.4
0.6
0.8
1.0
Low Freq Drift Rate (v)
0.0
0.2
0.4
0.6
Very Low Freq. Drift Rate (v)
0.0
0.1
0.2
0.3
0.4
0.5
Nonword Drift Rate (v)
0.0
0.1
0.2
0.3
0.4
0.5
1-120
121-240
241-360
1-240
241-480
1-480
481-960
1-2100
Trial Blocks

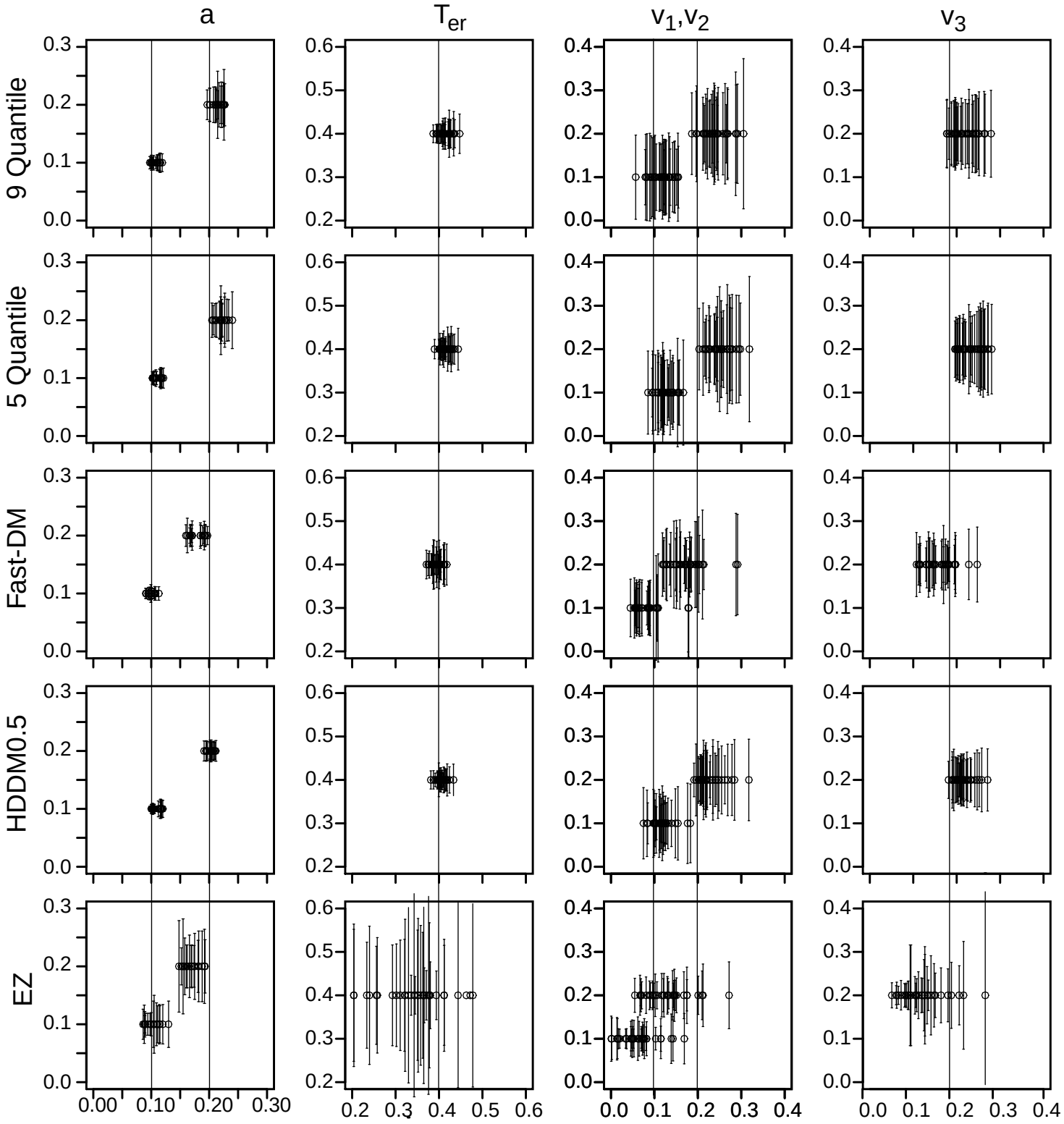

Diff N=1000 p0=0
EZ N=1000 p0=0
Diff N=1000 p0=0.04
EZ N=1000 p0=0.04
a
Ter
v
r=0.98
r=0.98
r=0.99
r=0.80
r=0.96
r=0.93
r=0.96
r=0.76
r=0.91
r=0.93
r=0.91
r=0.77
Parameter values recovered from fitting the simulated data
Parameter values used to generate the simulated data

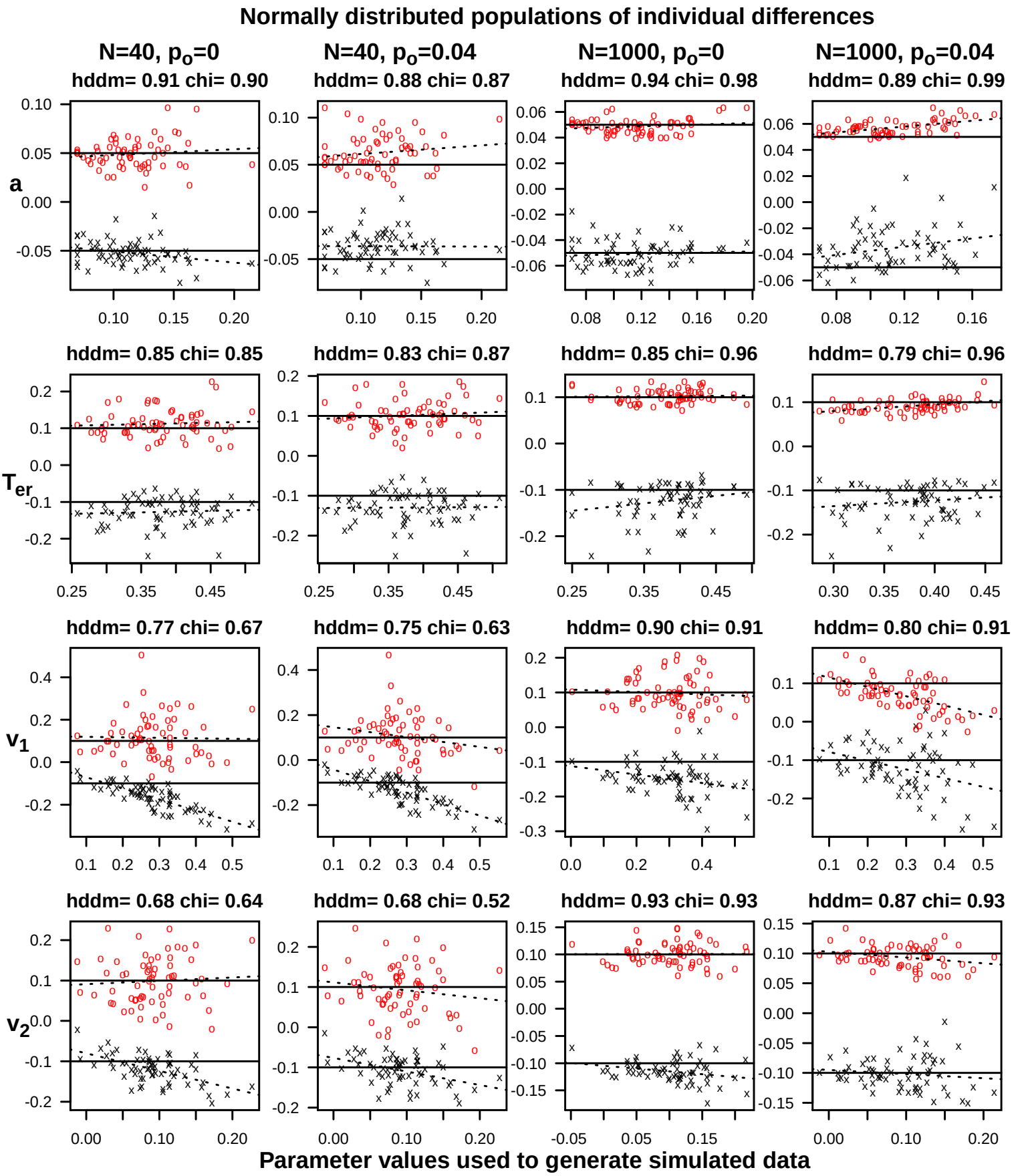

Diff N=40 p_o=0
EZ N=40 p_o=0
Diff N=40 p_o=0.04
EZ N=40 p_o=0.04
a
T_{er}
v
r=0.90
r=0.82
r=0.87
r=0.64
r=0.85
r=0.75
r=0.87
r=0.58
r=0.67
r=0.74
r=0.63
r=0.61
Parameter values recovered from fitting the simulated data
Parameter values used to generate the simulated data

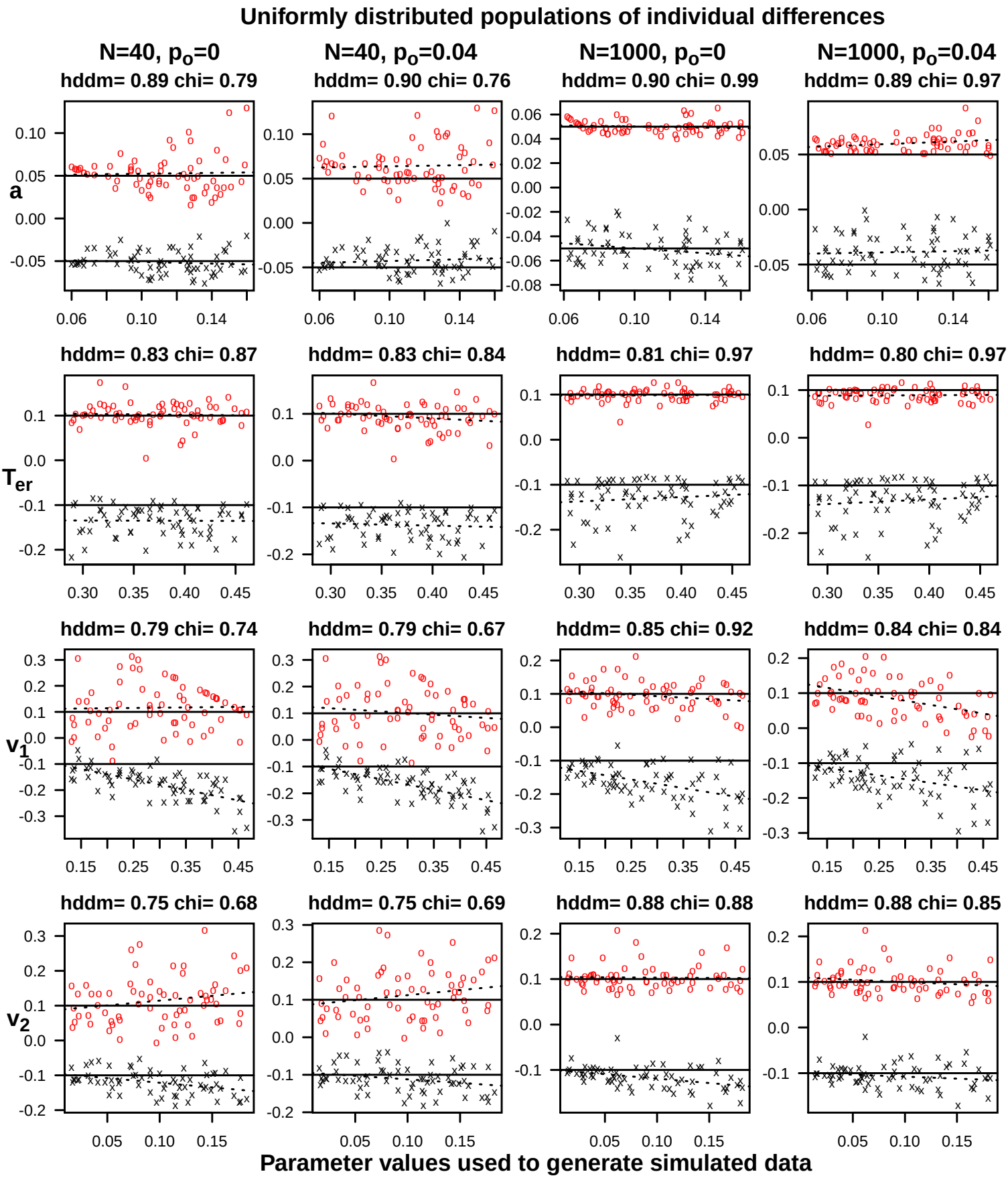

a
T_{er}
v_1,v_2
9 Quantile
5 Quantile
Fast-DM
HDDM0.5
DMAT no contam. correction
DMAT contam. correction
EZ
0.25
0.20
0.15
0.10
0.05
0.50
0.40
0.30
0.20
0.40
0.30
0.20
0.10
0.05
0.15
0.25
0.35
0.50

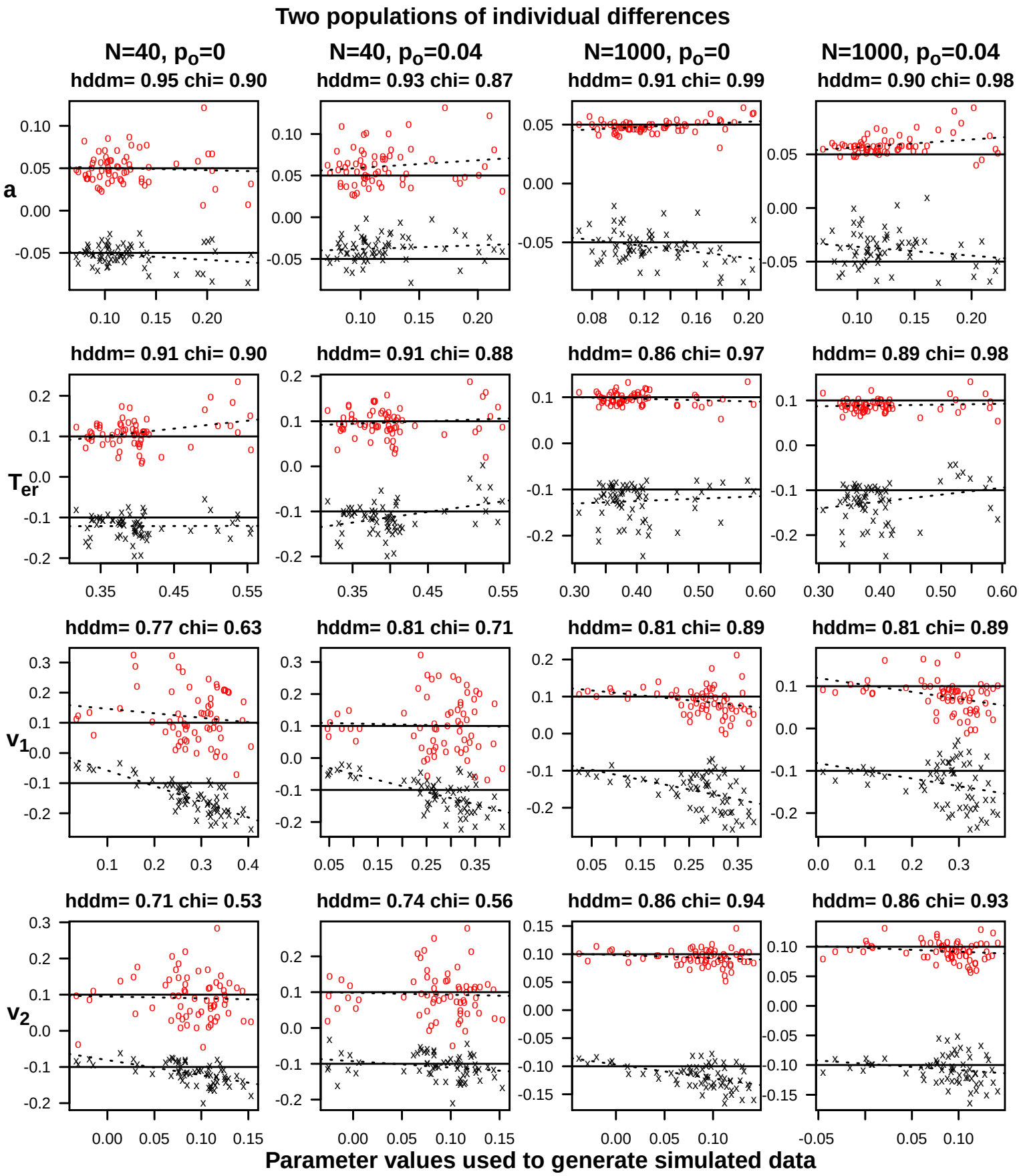

a
T_{er}
v_1,v_2
v_3
9 Quantile
5 Quantile
Fast-DM
HDDM0.5
DMAT no contam. correction
DMAT contam. correction
EZ

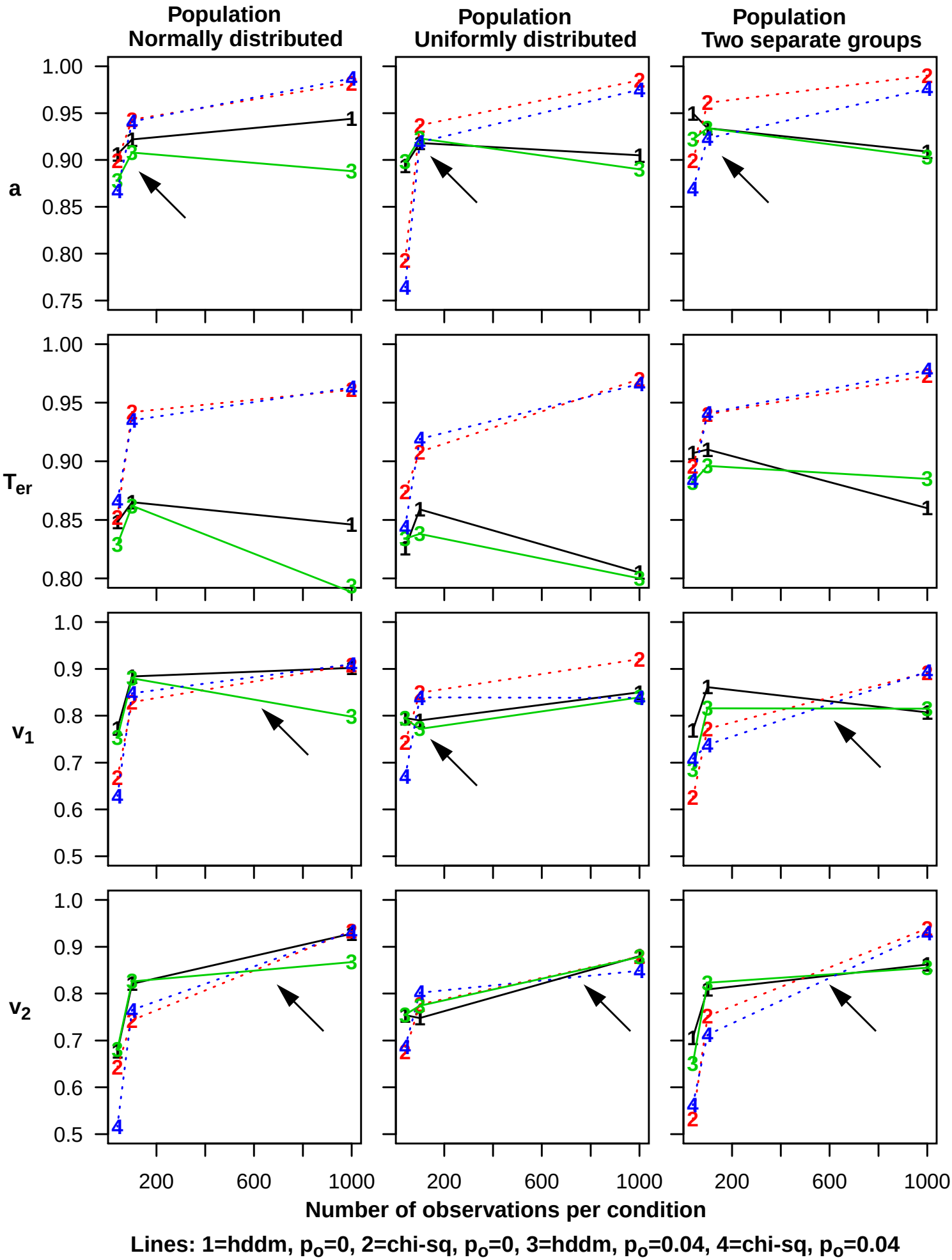

a
T_{er}
v_1,v_2
9 Quantile
5 Quantile
Fast-DM
HDDM0.5
EZ
0.00
0.10
0.20
0.30
0.2
0.3
0.4
0.5
0.6

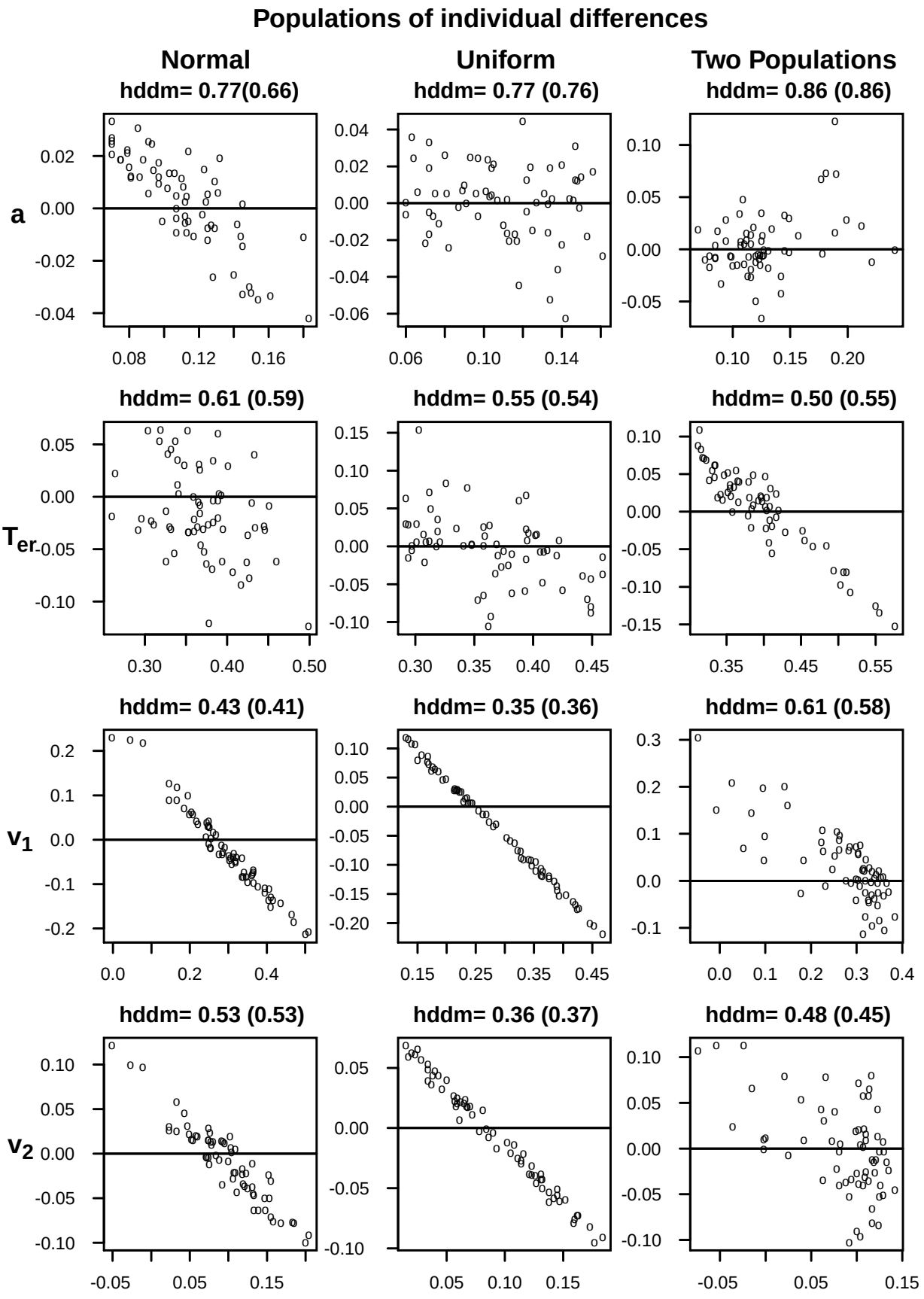

a
T_{er}
v_1,v_2
v_3
9 Quantile
5 Quantile
Fast-DM
HDDM0.5
EZ

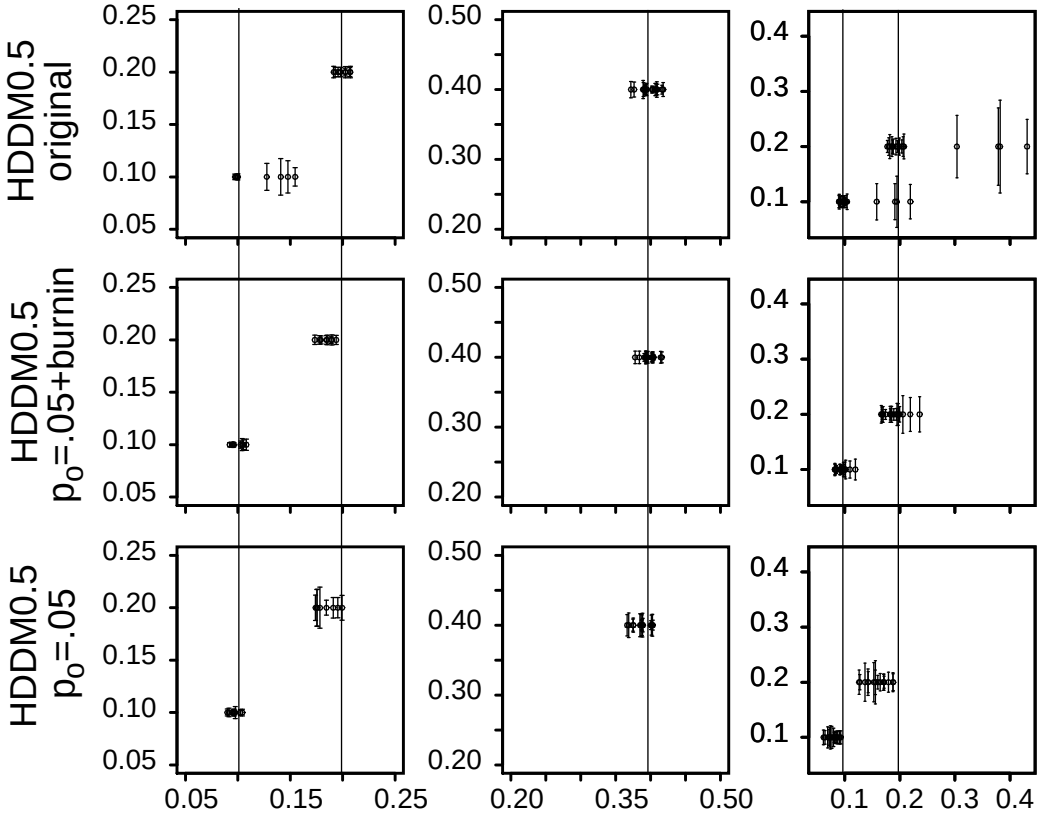

Normally distributed populations of individual differences
N=40, p0=0
N=40, p0=0.04
N=1000, p0=0
N=1000, p0=0.04
hddm= 0.91 chi= 0.90
hddm= 0.88 chi= 0.87
hddm= 0.94 chi= 0.98
hddm= 0.89 chi= 0.99
a
hddm= 0.85 chi= 0.85
hddm= 0.83 chi= 0.87
hddm= 0.85 chi= 0.96
hddm= 0.79 chi= 0.96
Ter
hddm= 0.77 chi= 0.67
hddm= 0.75 chi= 0.63
hddm= 0.90 chi= 0.91
hddm= 0.80 chi= 0.91
v1
hddm= 0.68 chi= 0.64
hddm= 0.68 chi= 0.52
hddm= 0.93 chi= 0.93
hddm= 0.87 chi= 0.93
v2
Parameter values used to generate simulated data


Uniformly distributed populations of individual differences
N=40, p0=0
N=40, p0=0.04
N=1000, p0=0
N=1000, p0=0.04
hddm= 0.89 chi= 0.79
hddm= 0.90 chi= 0.76
hddm= 0.90 chi= 0.99
hddm= 0.89 chi= 0.97
hddm= 0.83 chi= 0.87
hddm= 0.83 chi= 0.84
hddm= 0.81 chi= 0.97
hddm= 0.80 chi= 0.97
hddm= 0.79 chi= 0.74
hddm= 0.79 chi= 0.67
hddm= 0.85 chi= 0.92
hddm= 0.84 chi= 0.84
hddm= 0.75 chi= 0.68
hddm= 0.75 chi= 0.69
hddm= 0.88 chi= 0.88
hddm= 0.88 chi= 0.85
a
Ter
v1
v2
Parameter values used to generate simulated data


Two populations of individual differences
N=40, p0=0
N=40, p0=0.04
N=1000, p0=0
N=1000, p0=0.04
hddm= 0.95 chi= 0.90
hddm= 0.93 chi= 0.87
hddm= 0.91 chi= 0.99
hddm= 0.90 chi= 0.98
a
hddm= 0.91 chi= 0.90
hddm= 0.91 chi= 0.88
hddm= 0.86 chi= 0.97
hddm= 0.89 chi= 0.98
Ter
hddm= 0.77 chi= 0.63
hddm= 0.81 chi= 0.71
hddm= 0.81 chi= 0.89
hddm= 0.81 chi= 0.89
v1
hddm= 0.71 chi= 0.53
hddm= 0.74 chi= 0.56
hddm= 0.86 chi= 0.94
hddm= 0.86 chi= 0.93
v2
Parameter values used to generate simulated data


Population
Normally distributed
Population
Uniformly distributed
Population
Two separate groups
a
T_{er}
v_1
v_2
1.00
0.95
0.90
0.85
0.80
0.75
1.0
0.9
0.8
0.7
0.6
0.5
200
600
1000
Number of observations per condition
Lines: 1=hddm, p_o=0, 2=chi-sq, p_o=0, 3=hddm, p_o=0.04, 4=chi-sq, p_o=0.04


Populations of individual differences
Normal
Uniform
Two Populations
hddm= 0.77(0.66)
hddm= 0.77 (0.76)
hddm= 0.86 (0.86)
a
hddm= 0.61 (0.59)
hddm= 0.55 (0.54)
hddm= 0.50 (0.55)
Ter
hddm= 0.43 (0.41)
hddm= 0.35 (0.36)
hddm= 0.61 (0.58)
v1
hddm= 0.53 (0.53)
hddm= 0.36 (0.37)
hddm= 0.48 (0.45)
v2


HDDM0.5
original
HDDM0.5
p0=.05+burnin
HDDM0.5
p0=.05
0.25
0.20
0.15
0.10
0.05
0.50
0.40
0.30
0.20
0.4
0.3
0.2
0.1
0.05
0.15
0.25
0.20
0.35
0.50
0.1
0.2
0.3
0.4