2026 ■ Vol. 22 ■ no. 2

Planning for Robust Inference from Accuracy Rates and Mean Response Times

Russell J. Boag [ab]

[a]University of Western Australia

[b]University of Newcastle, Australia

Abstract ■ Lerche and Voss (2020) showed that accuracy rates and mean response times alone are insufficient to identify which diffusion decision model parameters vary across experimental conditions. Different parameter combinations can produce indistinguishable behavioural summaries, creating a risk of false process-level conclusions. I extend those simulations to examine how parameter mimicry depends on data quality, model specification, and design complexity, and to evaluate how simulation can inform experimental planning before data collection. Across a set of simulation studies, mimicry was reduced by increasing trials per condition, increasing the availability of error responses, fixing nondecision time across conditions unless theory requires otherwise, using designs with more than two conditions, and incorporating response time variability alongside accuracy and mean response time. The strongest constraint came from adding response time variability as an outcome measure. These results position simulation as a tool for model-informed experimental design and provide practical guidance for planning studies that support more robust cognitive inference.

Keywords ■ diffusion decision model, evidence accumulation, parameter mimicry, experimental design, model-informed design. **Tools** ■ R, sh.

russell.boag@uwa.edu.au

10.20982/tqmp.22.2.p025

Acting Editor ■ Cheng-Ta Yang (Department of Psychology, National Cheng Kung University)

Reviewers ■ Two anonymous reviewers.

Introduction

A central goal of cognitive psychology is to infer latent cognitive processes from observed behaviour. In two-choice tasks, the diffusion decision model (DDM) introduced by Ratcliff (1978) is a leading framework for doing so because it links response accuracy and response time (RT) to psychologically interpretable components of the decision process. In the DDM, noisy evidence accumulates over time until it reaches one of two response boundaries. The boundary that is reached first determines the response, and the time taken to reach that boundary forms the decision component of RT.

Observed RT also includes processes outside the decision itself, such as perceptual encoding and motor execution. The model therefore treats RT as the sum of decision time and nondecision time. Its core parameters describe different psychological influences on this process. Drift rate reflects the average quality or speed of evidence accumulation, boundary separation reflects response caution, starting point reflects bias toward one response, and nondecision time captures non-decision processes. Because these parameters jointly determine accuracy and the RT distribution, a change in behaviour is not automatically diagnostic of a unique latent process.

This ambiguity is the problem of parameter mimicry. In the present context, mimicry occurs when different DDM parameter configurations produce behavioural summaries that fall within the same tolerance-defined target, even though they imply different latent processes. Tolerance boundaries define the behavioural windows used to decide whether simulated data are close enough to a target case to count as a match, and parameter patterns describe the qualitatively distinct parameter-change explanations that remain viable among those matches. Mimicry is inferentially problematic because DDM parameters can compensate for one another in their joint effects on accuracy and RT. Consequently, an observed behavioural effect may ap-

Table 1 ■ Summary of target behavioural cases.

Case	Description	Accuracy	Mean RT (ms)
1	Baseline / null effect	94%	550
2	Lower accuracy, slower RT	90%	600
3	Higher accuracy, slower RT	98%	600
4	Equal accuracy, slower RT	94%	600
5	Lower accuracy, equal RT	90%	550

pear to support one process-level explanation even though alternative parameter changes would produce the same behavioural effect within tolerance.

This concern was demonstrated directly by Lerche and Voss (2020). They showed that accuracy and mean RT can be insufficient to identify which DDM parameters differ across conditions, because qualitatively different parameter changes can produce behavioural summaries that are indistinguishable within the relevant tolerances. As a consequence, analyses that rely on those summaries alone can support false conclusions about which latent processes differ across conditions. The same concern is consistent with the logic of the EZ diffusion model described by Wagenmakers et al. (2007), which uses accuracy, mean RT, and RT variability jointly to identify core DDM parameters. If RT variability carries parameter-relevant information, then inferences based only on accuracy and mean RT should be expected to remain underconstrained.

The present study extends those simulations with two aims. First, I examine which features of the data and model specification make mimicry more or less likely. Second, I evaluate practical strategies for reducing mimicry. Across a set of simulation studies, I show that mimicry is reduced by increasing the number of trials per condition, increasing the availability of error responses, fixing nondecision time across conditions unless theory requires otherwise, using designs with more than two conditions, and incorporating RT variability in addition to accuracy and mean RT. More broadly, the paper treats simulation as a tool for planning robust inference. The aim is not only to fit a model after data collection, but to use the model in advance to design experiments that are less vulnerable to ambiguous process-level conclusions, consistent with recent guidance on planning evidence-accumulation modelling studies (Boag et al., 2025).

Method

Diffusion Decision Model

The simulations used the DDM parameters introduced above, namely drift rate (v), boundary separation (a), starting point (z), nondecision time (t_0), and across-trial drift-rate variability (s_v). The simulated parameter ranges are reported below.

General Simulation Approach

The simulations followed the logic of Lerche and Voss (2020) by organising the analysis into two stages. First, for each parameterisation, three model parameters were varied over grids of 100 values, yielding $100^3 = 1{,}000{,}000$ unique parameter combinations. For each combination, a dataset of size N was generated from the DDM, producing one million simulated datasets for each trial count. Each simulated dataset was then compared with one of five target behavioural cases defined by combinations of accuracy and mean RT (Table 1). A dataset was classified as a behavioural match when its summary statistics fell within the prespecified tolerances of a target case. This stage quantified the prevalence of behavioural mimicry, operationalised as the proportion of simulated parameter sets that produced the same aggregate behavioural pattern.

Second, for datasets classified as behavioural matches, I characterised the parameter relations underlying mimicry. Relative to the baseline case (Case 1), each parameter was coded as smaller than ($<$), equal to ($=$), or larger than ($>$) a baseline reference value. Following Lerche and Voss (2020), the reference values were obtained by a sequential narrowing procedure around the medians of the matched baseline distributions.[1] With three parameters and three possible relations for each, 27 unique parameter patterns are possible.

Together, the target cases, tolerance rules, and parameter-pattern coding preserved the target-case logic of Lerche and Voss (2020) while extending the simulations to design choices that researchers can evaluate when planning new studies. The resulting design asked whether common behavioural summaries would remain sufficiently diagnostic under plausible changes in data quantity, task difficulty, model flexibility, and summary information, rather than whether the target cases exhaust all possible DDM behaviours.

[1] The baseline reference values were obtained by successively narrowing the matched Case 1 parameter sets around the median of each parameter in turn, using the parameter tolerances reported in Table 2, and then recalculating the medians within the narrowed subset.

Table 2 ■ Summary of simulated parameter ranges.

Parameter	Model	Lower	Upper	Step size	Tolerance
a	All	0.500	1.985	0.015	0.07425
v	All	1.000	4.960	0.040	0.198
t_0	$a - v - t_0$	0.100	0.595	0.005	0.02475
s_v	$a - v - s_v$	0.000	1.980	0.020	0.099
z	$a - v - z$	0.010	$a - 0.010$	$(a - 0.020)/99$	$0.05 \times (a - 0.020)$

Note. a = boundary separation; v = drift rate; s_v = across-trial drift-rate variability; t_0 = nondecision time; z = starting point bias.

Tolerances

Behavioural Tolerances

Lerche and Voss (2020) used an accuracy tolerance of .005 when determining whether a simulated dataset matched a target case. That tolerance is sensible when very large trial counts are used, but it becomes problematic at lower trial counts because the resolution of observed accuracy is constrained by $1/N$. For example, when $N \leq 200$, an accuracy tolerance of .005 effectively collapses to exact matching. This would imply less mimicry for smaller and noisier datasets, which is not a defensible comparison.

To avoid that artefact, the present simulations scaled the accuracy tolerance to the resolution of each dataset, using a tolerance of $1.01/N$. This rule ensures that the tolerance is slightly larger than the smallest observable change in accuracy at each trial count. For mean RT, I followed Lerche and Voss (2020) in using a fixed tolerance of 5 ms. In analyses that additionally required matches on RT variability, the same 5 ms tolerance was applied to the RT standard deviation.

Parameter Tolerances

For classifying parameter patterns, I followed Lerche and Voss (2020) in using a tolerance equal to 5% of the range of each simulated parameter. Parameter tolerances therefore depended on the parameterisation being considered (Table 2).

Simulation Factors

The simulations examined five planning-relevant factors that researchers can evaluate before data collection. These factors captured the amount of within-participant data, the availability of error responses, the flexibility of the intended DDM parameterisation, the richness of the behavioural summaries, and the number of behavioural cases that must be explained simultaneously. First, each simulation was repeated for $N = 50, 100, 150, 200, 250, 300, 350$, and 400 trials to evaluate the effect of measurement noise. Second, to examine the effect of increased error information, a second set of target cases reduced the accuracy of each original case by 10 percentage points. Third, I compared model parameterisations. In the baseline parameterisation, a, v, and t_0 were free to vary, whereas two alternative parameterisations fixed t_0 and instead allowed either s_v or z to vary. This comparison tested whether eliminating condition-wise variation in nondecision time reduced mimicry. Fourth, in addition to analyses based on accuracy and mean RT, I examined analyses that also required a match on RT standard deviation, with the target RT standard deviation fixed at 145 ms for all behavioural cases. Finally, I examined whether the same local parameter region could reproduce more than one behavioural case, thereby indexing how broadly mimicry generalised across conditions.

Outcome Measures

Four outcomes were considered. The first was the proportion of behaviourally identical datasets, which indexes the overall prevalence of mimicry. The second was the number of unique parameter patterns associated with those matches, which indexes the diversity of viable process-level explanations. The third was the distribution of the most common mimicry patterns, which identifies the particular alternative explanations most likely to mislead an investigator. The fourth was the number of admissible parameter sets that spanned multiple behavioural cases, which indexes how many alternative solutions generalise across conditions rather than within a single contrast alone. Together, these outcomes are informative because they quantify not only how much mimicry is present, but also how varied, structured, and general that mimicry is.

To complement the plots, I also computed a set of omnibus descriptive summaries for each analysis. For the proportion of behaviourally identical datasets, the number of unique parameter patterns, and the cross-case span count outcomes, each measure was first averaged within each trial count over the sampled values of N (50, 100, 150, 200, 250, 300, 350, and 400 trials). I then computed the mean level across trial counts, the proportional change from 50 to 400 trials, and the smallest trial count at which the out-

come dropped to 50% or less of its value at 50 trials (the half-reduction threshold).

To characterise diminishing returns, I calculated the absolute reduction from 50 to 200 trials and from 200 to 400 trials, expressed each reduction as a percentage of the earlier level, and formed an early-versus-late gain ratio by dividing the first reduction by the second. For the pattern-frequency analyses, I additionally computed concentration indices that captured the proportion of all matching datasets accounted for by the single most common mimicry pattern (top-1 share) and by the three most common patterns combined (top-3 share). For contrasts among design choices, I computed proportional reductions in the mean-over-trials summary relative to a matched reference analysis that differed only in the focal manipulation, such as adding error information, changing parameterisation, or requiring a match on RT variability. These summaries were descriptive rather than inferential and were used to quantify the magnitude, concentration, and diminishing returns of the effects shown in the figures.

Target Behavioural Cases

The target cases used the same accuracy and mean RT values as Lerche and Voss (2020), preserving the qualitative logic of the original simulation while making the behavioural contrasts interpretable as common experimental outcomes. Case 1 served as the baseline reference. Cases 2-5 then represented distinct ways in which behaviour can differ across conditions, including lower accuracy with slower RT, higher accuracy with slower RT, equal accuracy with slower RT, and lower accuracy with equal RT. These cases were not intended to exhaust all possible task regimes, but to provide a structured test bed for asking whether accuracy and mean RT contain enough information to distinguish among competing DDM parameter explanations.

Simulated Parameter Ranges

The ranges were chosen to span plausible values for the relevant DDM parameters while keeping the grid resolution comparable across parameterisations. This made the simulations broad enough to reveal mimicry while retaining a common structure for comparing the three-parameter models.

Software and Code Availability

Simulation and analysis code are available at https://osf.io/2dzqe/.

Results

Proportion of Behaviourally Identical Datasets

Figure 1 shows the prevalence of behavioural mimicry, defined as the proportion of simulated datasets that matched the target accuracy and mean RT values within tolerance. Across analyses, increasing the number of trials reduced mimicry, with the steepest decline at lower trial counts and a clear flattening by roughly 300 to 400 trials. Thus, additional data continued to help, but with diminishing returns once measurement noise was substantially reduced.

Increasing the error rate also reduced mimicry. Across parameterisations, the +10% error target cases, in which each target accuracy was reduced by 10 percentage points while the mean RT target was retained, yielded markedly fewer matches than the original cases. This effect was strongest for the high-accuracy case. In the case-level mean-over-trials summaries, Case 3 declined by 77.5% in $a-v-t_0$, 66.4% in $a-v-z$, 37.0% in $a-v-s_v$, and 84.0% in $a-v-t_0$ + RT SD, consistent with the idea that mimicry was greatest when accuracy approached ceiling.

Model parameterisation also mattered. Relative to the $a-v-t_0$ specification, mimicry was lower when t_0 was fixed and either s_v or z was allowed to vary instead. In the original mean-over-trials summaries, the reduction relative to $a-v-t_0$ was 42.1% for $a-v-s_v$ and 67.3% for $a-v-z$, so the stronger reduction occurred in the $a-v-z$ model. This pattern suggests that allowing nondecision time to vary across conditions is a particularly permissive source of behavioural equivalence.

The strongest reduction occurred when RT variability was added to the matching criterion. Requiring datasets to match accuracy, mean RT, and RT standard deviation dramatically reduced the proportion of behaviourally identical datasets, indicating that RT variability provides substantial additional constraint beyond the first moments alone.

The half-reduction thresholds for Figure 1 were correspondingly early. In the original analyses, $a-v-t_0$ and $a-v-t_0$ + RT SD reached half reduction by 100 trials; $a-v-z$ and $a-v-s_v$ reached half reduction by 150 trials. For the +10% error cases, $a-v-t_0$, $a-v-z$, $a-v-s_v$ and $a-v-t_0$ + RT SD reached half reduction by 150 trials. Thus, for the prevalence of mimicry, the threshold results reinforce the visual impression that most of the practical gain occurred within roughly the first 100 to 150 trials per condition.

Number of Unique Parameter Patterns

Figure 2 reports the number of unique parameter patterns associated with behavioural matches. Increasing the num-

ber of trials reduced the diversity of admissible parameter patterns, although this reduction was more gradual than the reduction in the overall proportion of matched datasets. The main implication is that larger datasets do not merely reduce the quantity of mimicry; they also reduce the variety of qualitatively distinct explanations that remain viable.

Adding errors produced a smaller effect on pattern diversity than on the overall prevalence of behavioural matches. Across the two-feature analyses, the relative reduction in mean match prevalence ranged from 46.5% to 65.7%, whereas the relative reduction in the mean number of unique patterns ranged only from 30.6% to 41.6%. Thus, adding errors removed many matching parameter sets, but it produced a smaller proportional reduction in the number of distinct parameter-change patterns that remained viable. Likewise, fixing nondecision time reduced the proportion of matched datasets more than it reduced the number of distinct parameter patterns. Relative to $a - v - t_0$ in the original target set, $a-v-z$ and $a-v-s_v$ reduced mean match prevalence by 67.3% and 42.1%, yet their mean numbers of unique patterns were slightly higher rather than lower (14.13 and 17.08 vs 13.68). By contrast, requiring a match on RT standard deviation produced a pronounced reduction in pattern diversity, reducing the $a-v-t_0$ mean from 13.68 to 7.43 and often leaving only a small number of surviving patterns at the larger trial counts.

The half-reduction thresholds for Figure 2 were later and less uniform than those for Figure 1. The strongest improvement again came from including RT variability. The $a - v - t_0$ + RT SD analysis reached half reduction by 150 trials in the original target set and by 150 trials in the +10% error set. In the original target set, $a-v-t_0$ + RT SD reached half reduction by 150 trials; $a - v - s_v$ reached half reduction by 300 trials; $a - v - z$ reached half reduction by 350 trials; $a-v-t_0$ did not reach half reduction even at 400 trials. In the +10% error target set, $a - v - t_0$ + RT SD reached half reduction by 150 trials; $a - v - s_v$ reached half reduction by 200 trials; $a-v-t_0$ and $a-v-z$ did not reach half reduction even at 400 trials. This pattern indicates that reducing the diversity of viable alternative explanations typically required either richer behavioural constraints or substantially more data than was needed merely to reduce the overall prevalence of mimicry.

Most Common Mimicry Patterns

Figure 3 shows the most prevalent parameter patterns. In the $a - v - t_0$ model, the dominant patterns involved compensatory changes in a, v, and t_0, such that the effects of two parameters on RT and accuracy were offset by the third. This result clarifies why allowing nondecision time to vary is so problematic. It introduces a flexible route by which qualitatively different process explanations can be made behaviourally indistinguishable.

In the $a - v - s_v$ model, the dominant patterns tended to involve coordinated changes in all three parameters, indicating that variation in response caution can offset joint changes in drift rate and drift-rate variability. In the $a - v - z$ model, the most common patterns reflected compensatory trade-offs among drift rate, caution, and response bias. That stability was visible in the original target-set aggregates, where the two most common patterns overall were $><<$ and $<>>$ for $a - v - t_0$, $>><$ and $><>$ for $a-v-z$, and $<<<$ and $>>>$ for $a-v-s_v$. Consistent with the concentration summaries, the single most common pattern accounted on average for 35.8%, 33.3%, and 32.6% of matches in those three parameterisations, whereas the top three patterns accounted for 74.8%, 66.7%, and 65.4%, respectively. Across parameterisations, the key point is that mimicry is not random noise in parameter space but is instead organised around a small set of recurrent compensatory relations. Although the exact top-3 patterns differed across parameterisations, they were highly stable within each model, indicating that each parameterisation carries its own characteristic set of recurring alternative explanations rather than a diffuse cloud of idiosyncratic solutions.

Including RT standard deviation changed the relative frequency of these patterns and reduced the prevalence of several otherwise common configurations. In the original target set, the leading RT-variability patterns shifted to $>>>$ and $>>=$, and the mean number of unique patterns fell from 13.68 to 7.43. Thus, richer behavioural constraints not only reduce the amount of mimicry but also alter which alternative process accounts remain plausible.

Number of Parameter Sets Spanning Multiple Behavioural Cases

Figure 4 shows the number of admissible parameter sets that reproduced more than one behavioural case. These cross-case solutions became less common as the number of trials increased, and the reduction was especially strong when RT variability was included. Expressed as raw counts, the parameterisation contrast was still informative but not perfectly uniform across span criteria. In the original mean-over-trials summaries, the number of solutions spanning at least two cases was highest for $a-v-t_0$ (6,397), intermediate for $a - v - s_v$ (4,267), and lowest for $a - v - z$ (2,401); the same ordering held for solutions spanning at least three cases (4,205, 3,667, and 1,950). For the stricter four- and five-case criteria, however, $a - v - s_v$ often exceeded $a - v - t_0$ across much of the trial range; for example, in the original target set the count of solutions spanning four or more cases was 4,850 versus 3,767 at $N = 100$ and 241 versus 120 at $N = 400$. Increasing error information

further reduced the broader-span counts, for example lowering the $a - v - t_0$ mean from 6,397 to 2,644 for solutions spanning at least two cases and from 4,205 to 2,069 for solutions spanning at least three cases. The main implication is that richer behavioural constraints reduce not only which cross-case alternatives remain viable, but also how many such misleading alternatives survive.

An additional implication is methodological. Parameter regions that can mimic several behavioural cases at once become rarer as the number of experimental conditions increases. Designs with more than two conditions therefore provide stronger protection against false process-level inferences, because any candidate explanation must simultaneously account for a broader set of behavioural constraints.

The half-reduction thresholds for Figure 4 were again relatively early for the count of solutions spanning at least two cases. In the original analyses, $a-v-t_0$ and $a-v-t_0$ + RT SD reached half reduction by 100 trials; $a - v - z$ and $a-v-s_v$ reached half reduction by 150 trials. In the +10% error analyses, $a-v-t_0$, $a-v-z$, $a-v-s_v$ and $a-v-t_0$ + RT SD reached half reduction by 150 trials. Thus, the number of cross-case solutions typically dropped by half within 100 to 150 trials, with the RT-variability analysis reaching that benchmark especially quickly.

Omnibus Summaries

Figure 5 summarises the key omnibus results. The omnibus summaries paralleled the visual patterns across Figures 1 to 4 while making the design contrasts easier to quantify. For the prevalence of mimicry, the mean proportion of matched datasets across trial counts was .00160 for the original $a - v - t_0$ analysis, .00093 for $a - v - s_v$, and .00052 for $a - v - z$. Relative to $a - v - t_0$, this corresponds to reductions of 42.1% for $a - v - s_v$ and 67.3% for $a-v-z$. Increasing error information further reduced the mean proportion of matches by 65.7% in the $a-v-t_0$ analysis, 57.7% in $a-v-z$, and 46.5% in $a-v-s_v$. The strongest contrast was the addition of RT variability. For the original $a-v-t_0$ analysis, requiring a match on RT standard deviation reduced the mean proportion of behavioural matches from .00160 to .000057, a 96.4% reduction.

The same pattern appeared for the diversity of surviving explanations. In the original two-feature analyses, the mean number of unique mimicry patterns across trial counts was 13.68 for $a-v-t_0$, 14.13 for $a-v-z$, and 17.08 for $a-v-s_v$. Adding RT variability reduced the $a-v-t_0$ mean from 13.68 to 7.43, a 45.7% reduction. The concentration summaries further showed that mimicry was not spread evenly across many alternatives. In the original analyses, the single most common pattern accounted on average for 35.8% of all matches in $a - v - t_0$, 33.3% in $a - v - z$, and 32.6% in $a - v - s_v$, whereas the three most common patterns accounted for 74.8%, 66.7%, and 65.4%, respectively. Importantly, increases in these top-1 and top-3 shares under more informative designs do not indicate more mimicry overall; rather, they indicate that the smaller set of surviving matches is concentrated in fewer recurrent patterns after many weaker alternatives have been eliminated.

For cross-case generalisation, RT variability again had the clearest effect. In the original $a - v - t_0$ analysis, the mean number of admissible parameter sets spanning at least two cases was 6,397. When RT variability was also matched, this fell to 168, a 97.4% reduction. The stricter span criteria showed the same pattern. In the original $a-v-t_0$ analysis, the mean counts spanning three or more cases, four or more cases, and all five cases were 4,205, 2,368, and 1,396. With RT variability included, those means fell to 73, 29, and 10, respectively. Because Figure 5 reports mean-over-trials summaries, these values should be read as the average burden of cross-case ambiguity across the sampled trial range rather than as endpoint counts at $N = 400$.

The raw means also make the parameterisation and error-information contrasts easier to see. In the original analyses, parameter sets spanning at least two cases averaged 6,397 for $a - v - t_0$, 4,267 for $a - v - s_v$, and 2,401 for $a - v - z$. In the +10% error analyses, these means fell to 2,644, 2,381, and 1,088, respectively. These changes correspond to reductions of 58.7%, 44.2%, and 54.7%, respectively.

These summaries also reinforce the diminishing-returns pattern seen in the trial-count curves. For example, the mean match proportion in the original $a - v - t_0$ analysis had already fallen below half of its 50-trial value by 100 trials. From 50 to 400 trials, it declined by 88.4% overall, and the reduction from 50 to 200 trials was about 7.0 times the reduction from 200 to 400 trials. For within-subject study planning, the practical implication is that the largest gains are likely to come from moving out of the very low-trial regime, whereas increasing an already moderate design from 200 to 400 trials per condition will often yield diminishing returns.

General Discussion

The present simulations extend Lerche and Voss (2020) by identifying when parameter mimicry is most likely and which design and analysis choices most effectively reduce it. The core conclusion is straightforward. Because compensatory parameter changes can allow the same behavioural summaries to support different latent-process explanations, process-level inference from those summaries is more reliable when the data are more in-

formative and when the model is less permissive. In that sense, the results speak directly to model-informed experimental design. The practical question is not only how to analyse a finished dataset, but how to design a study so that competing parameter explanations are unlikely to survive in the first place.

Three findings were especially consistent. First, increasing the number of trials per condition reduced both the frequency of behavioural mimicry and the diversity of parameter patterns that could support it. In practical terms, the outcome often reached its half-reduction threshold, defined as the trial count at which it first fell to 50% or less of its value at 50 trials, within 100 to 150 trials for match prevalence and within 100 to 150 trials for solutions spanning at least two cases. The analogous threshold for unique mimicry patterns was later and less uniform. Second, increasing the availability of error responses reduced mimicry further, particularly for high-accuracy cases in which near-ceiling performance otherwise leaves little discriminating information. For Case 3, the mean-over-trials reduction was 77.5% in $a-v-t_0$, 66.4% in $a-v-z$, 37.0% in $a-v-s_v$, and 84.0% in $a-v-t_0$ + RT SD. Third, fixing nondecision time across conditions reduced mimicry relative to the $a-v-t_0$ parameterisation, consistent with the idea that freely varying nondecision time can absorb differences that might otherwise be attributed to decision-related parameters. The size of that benefit depended on the outcome being used. For overall match prevalence, the stronger reduction occurred in $a-v-z$ (67.3%) rather than $a-v-s_v$ (42.1%), whereas the unique-pattern summaries did not show parallel reductions. At the same time, the omnibus contrasts show that these improvements do not automatically translate into a smaller diversity of surviving explanations, since a parameterisation can yield fewer behavioural matches overall while the matches that do survive remain at least as diverse, or even more diverse, in terms of unique parameter patterns.

The omnibus summaries sharpen this interpretation in two ways. First, they show that the same design manipulations that reduce mimicry also reduce the overall burden of ambiguity across the full range of trial counts, as indexed by lower mean-over-trials summaries, larger 50–to–400-trial reductions, and larger proportional reductions relative to the matched reference designs. This was especially clear for RT variability, since in the original $a-v-t_0$ analysis, adding RT standard deviation reduced mean match prevalence by 96.4%, mean unique patterns by 45.7%, the mean count of solutions spanning at least two cases by 97.4%, and the mean count spanning at least three cases by 98.3%. Second, they make the diminishing-returns pattern explicit, as in most analyses the absolute and percentage reductions from 50 to 200 trials exceeded those from 200 to 400 trials, meaning that the early-versus-late gain ratio was greater than 1. The concentration summaries likewise showed that ambiguity was often dominated by a limited set of recurrent alternatives, with the single most common pattern often accounting for about one third of all matches and the top three patterns often accounting for about two thirds to three quarters. For planning within-subject data collection, this means that increasing trials per condition from a sparse level to a moderate level is often more valuable than pushing an already moderate design to very large trial counts. Viewed alongside the error-rate contrasts, the same pattern suggests that, when the trial budget is constrained, calibrating the task to yield more incorrect responses can be a more efficient marginal improvement than adding another modest block of trials to an already moderate design. This is useful for study planning because it translates the plotted trends into quantities that can be compared directly when evaluating design trade-offs, as summarised visually in Figure 5.

The half-reduction thresholds illustrate the same point especially clearly. For the prevalence of mimicry, halving occurred within 100 to 150 trials per condition across the single-case analysis and target-set combinations. For the count of cross-case solutions spanning at least two cases, the corresponding range was 100 to 150 trials. By contrast, the number of unique mimicry patterns halved only within 150 to 350 trials in the combinations that reached that threshold, and 3 of the eight combinations did not halve even by 400 trials. In practical terms, this means that a moderate increase in trials is often sufficient to reduce how often mimicry occurs, but a stronger design improvement may be needed to substantially reduce the diversity of qualitatively distinct alternative explanations that remain viable.

The error-rate results should therefore be read as a boundary condition on near-ceiling designs rather than as a general recommendation to make tasks as difficult as possible. The +10% error manipulation was most informative when baseline accuracy was high, because near-ceiling performance leaves little error information with which to distinguish among parameter explanations. For paradigms that naturally operate around 60% to 70% accuracy, further increasing error responses may be less profitable than improving other design features. In such settings, researchers may instead prioritise increasing the number of trials per condition, preserving informative RT variability, adding experimental conditions, or imposing theoretically justified parameter constraints. Pushing such tasks toward still lower accuracy can introduce floor effects, guessing, disengagement, or a qualitative change in the process being measured. At the limit, paradigms approaching 50% or chance accuracy are poor candidates

for this strategy, because the DDM will struggle to identify meaningful evidence-accumulation structure when performance is effectively indistinguishable from guessing.

The strongest countermeasure was to include RT variability in the matching criterion. Accuracy and mean RT constrain only a limited portion of the information contained in a response-time dataset. Once RT standard deviation was also required to match, the proportion of admissible parameter solutions dropped sharply. This result is theoretically important because it shows that the risk identified by Lerche and Voss (2020) is not simply a property of the DDM itself; it is partly a consequence of attempting to infer latent processes from an impoverished summary of the data.

The pattern analyses clarify the structure of the problem. Across parameterisations, mimicry was dominated by compensatory parameter relations rather than by diffuse or idiosyncratic solutions. In practical terms, apparent evidence for a particular process interpretation may often have a small number of recurrent compensatory alternatives, which are systematic parameter-change patterns that can produce the same behavioural summaries within tolerance. The specific alternatives depend on the chosen parameterisation. In the $a - v - t_0$ model, for example, apparent effects of drift rate and caution may be masked or exaggerated by compensatory changes in nondecision time. In the $a - v - z$ and $a - v - s_v$ models, different but equally systematic compensation patterns emerge. Importantly, the recurring top-3 patterns were stable within each model but not identical across models, which suggests that each parameterisation has its own characteristic ambiguity structure. This logic also helps explain the concentration summaries and the cross-case span counts, since adding more diagnostic information, such as more errors or RT variability, can increase top-1 and top-3 shares even as total mimicry falls, because the remaining ambiguity is compressed into a smaller number of robust alternative patterns rather than spread across many weak ones. The count-based span summaries show the complementary point that more informative designs leave fewer parameter sets capable of generalising across multiple behavioural cases.

The cross-case results also qualify any attempt to rank parameterisations with a single best-to-worst ordering. In the original summaries, $a - v - t_0$ produced the most solutions spanning at least two and three cases (6,397 and 4,205), whereas $a - v - z$ produced the fewest (2,401 and 1,950). Yet the stricter four-case criterion showed that $a - v - s_v$ could exceed $a - v - t_0$ over much of the trial range, for example 4,850 versus 3,767 at $N = 100$ and 241 versus 120 at $N = 400$. In practical terms, study planning should account for different forms of ambiguity, since a design choice may reduce the overall prevalence of mimicry while still leaving a diverse or broadly generalisable subset of viable alternative explanations.

These results also have implications beyond full diffusion-model fitting. The logic of the EZ diffusion model (Wagenmakers et al., 2007) is that accuracy, mean RT, and RT variability jointly constrain the core DDM parameters, so that richer summary information can support useful process-level inference even without fitting the full RT distribution. The present simulations support that claim by showing that, once all three data features are considered together, mimicry is greatly reduced relative to analyses based only on accuracy and mean RT. At the same time, the present findings also clarify why the EZ logic should not be treated as a license to rely on overly sparse summaries. Accuracy and mean RT alone were often insufficient to identify the generating parameter set, whereas adding RT variability markedly improved identifiability. Thus, the broader point is that simpler summary-based approaches are useful only when they capture the behavioural features needed to discriminate among competing parameter explanations.

This interpretation is also consistent with Dutilh et al. (2019). In that blinded modelling challenge, teams using detailed formal models did not uniformly outperform simpler approaches. Of particular relevance here, the Evans and Brown team achieved strikingly good performance using a model-informed heuristic approach that effectively eyeballed mean RT, accuracy rates, and RT variability rather than relying on elaborate model estimation. Their success is informative because it shows that, when analysts attend to the right combination of summaries, a great deal of the relevant process information is already visible in the data. Nevertheless, the challenge does not imply that formal modelling is unnecessary. Rather, it underscores the same point as the present simulations. That is, the quality of inference depends critically on whether the analysis captures a sufficiently diagnostic set of behavioural constraints. Summary-based reasoning can therefore be surprisingly effective when it includes accuracy, mean RT, and RT variability, but it becomes much less trustworthy when it is based on only one or two underinformative moments.

Crucially, these results should not be read as an argument against formal modelling. Full model fitting remains the appropriate approach when researchers wish to exploit the full RT distribution, compare theoretically motivated parameterisations, and quantify uncertainty. Rather, the present point is that even before model fitting, researchers can substantially reduce the risk of false conclusions by collecting better-constrained data and by avoiding underinformative summaries.

Several limitations should be acknowledged. First, the

simulations are based on aggregate summaries rather than full model fits, so they illuminate identifiability at the level of behavioural features rather than parameter recovery under specific estimation procedures. Second, the analyses consider only a limited set of parameterisations and fixed parameter ranges. Different tasks, parameter priors, or theoretically motivated constraints may alter the prevalence and structure of mimicry. Third, the present design manipulations vary trials per condition but not the number of participants, individual differences, or hierarchical structure. As a result, the findings are most directly relevant to within-participant data quality rather than to full study-level planning. Nevertheless, the present results support a shift toward model-informed experimental design (Boag et al., 2025), in which cognitive models are used not only to interpret data after collection but also to choose task settings that maximise inferential robustness before data are collected.

Conclusions

Several practical recommendations follow from these simulations. Researchers should collect enough trials per condition to reduce measurement noise, ensure that task performance yields a meaningful number of error responses, avoid freeing nondecision time across conditions without strong theoretical justification, use designs with multiple behavioural conditions where possible, and include RT variability alongside accuracy and mean RT in descriptive and modelling analyses. These steps do not eliminate parameter mimicry entirely, but they substantially reduce it and thereby improve the credibility of process-level inference from behavioural data.

Design Recommendations

For researchers planning experiments, the present results suggest a straightforward model-informed workflow. First, select task settings that are likely to produce a nontrivial proportion of errors rather than near-ceiling performance. When pilot data suggest very high accuracy and the trial budget is small, adjusting task difficulty to produce informative errors may matter more than adding a modest number of trials once the design is already in the roughly 100- to 150-trial-per-condition range, because the +10% error manipulation generally produced larger improvements in the outcome measures than the next trial-count increment. The error-information recommendation is weaker when the task already produces low accuracy. For paradigms operating around 60% to 70% accuracy, it may be more useful to increase trials, retain RT variability, or add constraining conditions than to further increase errors and risk chance-level performance. Second, prioritise collecting enough within-subject trials per condition to stabilise accuracy and RT estimates before increasing model flexibility; the omnibus summaries suggest that moving from very small trial counts toward roughly 100 to 200 trials per condition may often buy more inferential protection than extending an already moderate design from 200 to 400 trials per condition. Third, prefer designs with multiple conditions or graded manipulations when theoretically possible, because they place stronger constraints on competing parameter explanations. Fourth, treat response time variability as a design-relevant outcome. Finally, use targeted simulations before data collection to evaluate whether the planned combination of trials, difficulty, and conditions is likely to support robust inference under the intended model parameterisation.

The omnibus summaries introduced here provide a practical language for that planning exercise. Mean-over-trials summaries quantify the typical burden of mimicry across plausible trial counts, half-reduction thresholds identify when an outcome falls to half of its 50-trial value, early-versus-late gain summaries show whether most improvement occurs before or after 200 trials, concentration indices show whether ambiguity is dominated by a few recurrent alternatives, and cross-case span summaries quantify how many of those alternatives generalise across conditions. In that sense, the simulations can support not only qualitative judgement from plots but also compact numerical comparisons among candidate designs.

As a concrete example, consider a researcher planning a two-choice RT experiment with two candidate task settings from pilot data. One setting yields high accuracy, for example around 95% to 98%, with modest differences in mean RT across conditions. The other setting yields lower but still interpretable accuracy, for example around 80% to 90%, while preserving the same qualitative RT contrast. The present framework would favour simulating both settings before data collection rather than choosing between them informally. For each setting, the researcher would simulate candidate DDM parameter sets or use pilot-informed parameter ranges, compare the resulting behavioural summaries with the expected accuracy, mean RT, and RT standard deviation, and then inspect the omnibus summaries. If the near-ceiling setting produces many tolerance-defined matches, many unique parameter patterns, or solutions that span multiple behavioural cases, it would be a poor basis for process-level inference even if the behavioural effect is reliable. A more error-informative setting may be preferable if it reduces mimicry without making the task so difficult that performance approaches floor or chance.

The same exercise can guide the trial-count decision. If increasing from 50 to 150 or 200 trials per condition sharply reduces mimicry prevalence but further increases

produce diminishing returns, the moderate design may be sufficient. If unique parameter patterns remain diverse, the researcher might instead add RT variability as an explicit constraint, add another condition, or impose theoretically motivated restrictions on which DDM parameters may vary across conditions. For paradigms that already operate around 60% to 70% accuracy, the same logic would usually point away from further increasing errors; increasing trial numbers or adding stronger design constraints would be more defensible than pushing the task toward chance performance. The selected design would therefore be the one that provides enough errors and RT variability to constrain the model while preserving an interpretable behavioural regime and a feasible participant burden.

Authors' note

Correspondence concerning this article should be addressed to Russell J. Boag (russell.boag@uwa.edu.au), School of Psychological Sciences, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. ORCID: 0000-0002-7689-0682. Simulation and analysis code are available at: https://osf.io/2dzqe/. The author declares no conflicts of interest. Russell J. Boag was supported by Australian Research Council Discovery grant DP250101740 and an Australia-US Multi-University Research Initiative grant (DSTG AUSMURIV000003, ONR/-DoD N00014-23-1-2792). This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

References

Boag, R. J., Innes, R. J., Stevenson, N., Bahg, G., Busemeyer, J. R., Cox, G. E., Donkin, C., Frank, M. J., Hawkins, G. E., Heathcote, A., Hedge, C., Lerche, V., Lilburn, S. D., Logan, G. D., Matzke, D., Miletic, S., Osth, A. F., Palmeri, T. J., Sederberg, P. B., Forstmann, B. U., et al. (2025). An expert guide to planning experimental tasks for evidence-accumulation modeling. *Advances in Methods and Practices in Psychological Science*, *8*(2), 1–41. doi: 10.1177/25152459251336127.

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., Maanen, L. v., C., D., et al. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. doi: 10.3758/s13423-017-1417-2.

Lerche, V., & Voss, A. (2020). When accuracy rates and mean response times lead to false conclusions: A simulation study based on the diff usion model. *The Quantitative Methods for Psychology*, *16*(2), 107–119. doi: 10.20982/tqmp.16.2.p107.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. doi: 10.3758/BF03194023.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/2dzqe/

Citation

Boag, R. J. (2026). Planning for robust inference from accuracy rates and mean response times. *The Quantitative Methods for Psychology*, *22*(2), 25–39. doi: 10.20982/tqmp.22.2.p025.

Received: 02/06/2026 ~ Accepted: 07/07/2026

Figures 1 to 5 follow.

Figure 1 ■ Effects of simulation factors on the proportion of behaviourally identical datasets. Panels show, from top left to bottom right, $a - v - t_0$, $a - v - z$, $a - v - s_v$, and $a - v - t_0$ when matching within tolerance on accuracy, mean RT, and RT standard deviation. Within each panel, the left facet shows the original target cases and the right facet shows the +10% error cases. Vertical dashed lines mark the half-reduction threshold, defined as the smallest trial count at which the mean outcome falls to 50% or less of its value at 50 trials. In-panel labels report the early-late gain ratio, defined as the 50–to–200-trial reduction divided by the 200–to–400-trial reduction. The parameter labels are a for boundary separation, v for drift rate, s_v for drift-rate variability, t_0 for nondecision time, and z for starting point bias.

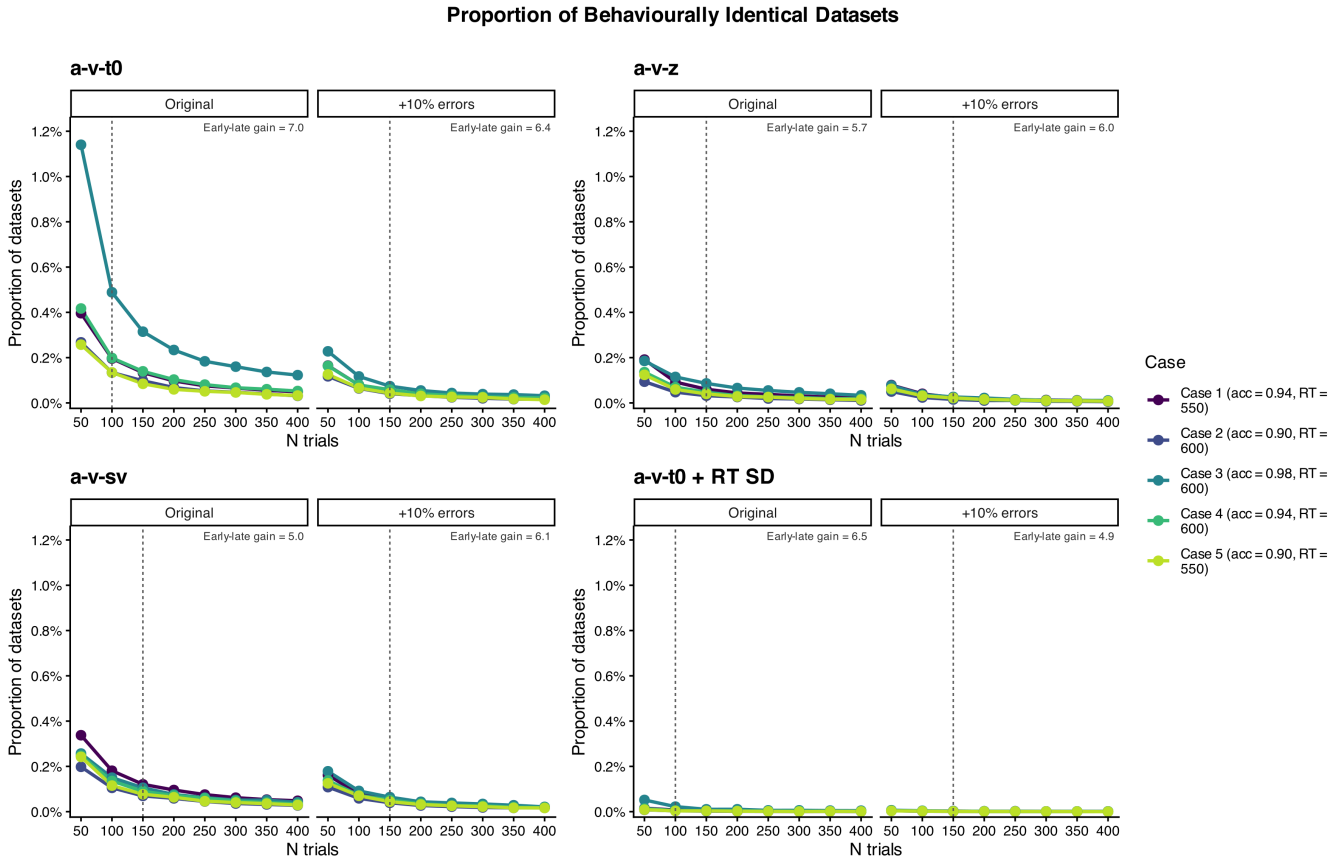

Figure 2 ■ Effects of simulation factors on the number of unique parameter patterns producing behavioural mimicry. Panels show, from top left to bottom right, $a - v - t_0$, $a - v - z$, $a - v - s_v$, and $a - v - t_0$ when matching within tolerance on accuracy, mean RT, and RT standard deviation. Each panel counts unique qualitative patterns for the three parameters named in that panel. Within each panel, the left facet shows the original target cases and the right facet shows the +10% error cases. Vertical dashed lines mark the half-reduction threshold, defined as the smallest trial count at which the mean outcome falls to 50% or less of its value at 50 trials. Panels without a dashed line did not reach half reduction even at $N = 400$. In-panel labels report the early-late gain ratio, defined as the 50–to–200-trial reduction divided by the 200–to–400-trial reduction. The parameter labels are a for boundary separation, v for drift rate, s_v for drift-rate variability, t_0 for nondecision time, and z for starting point bias.

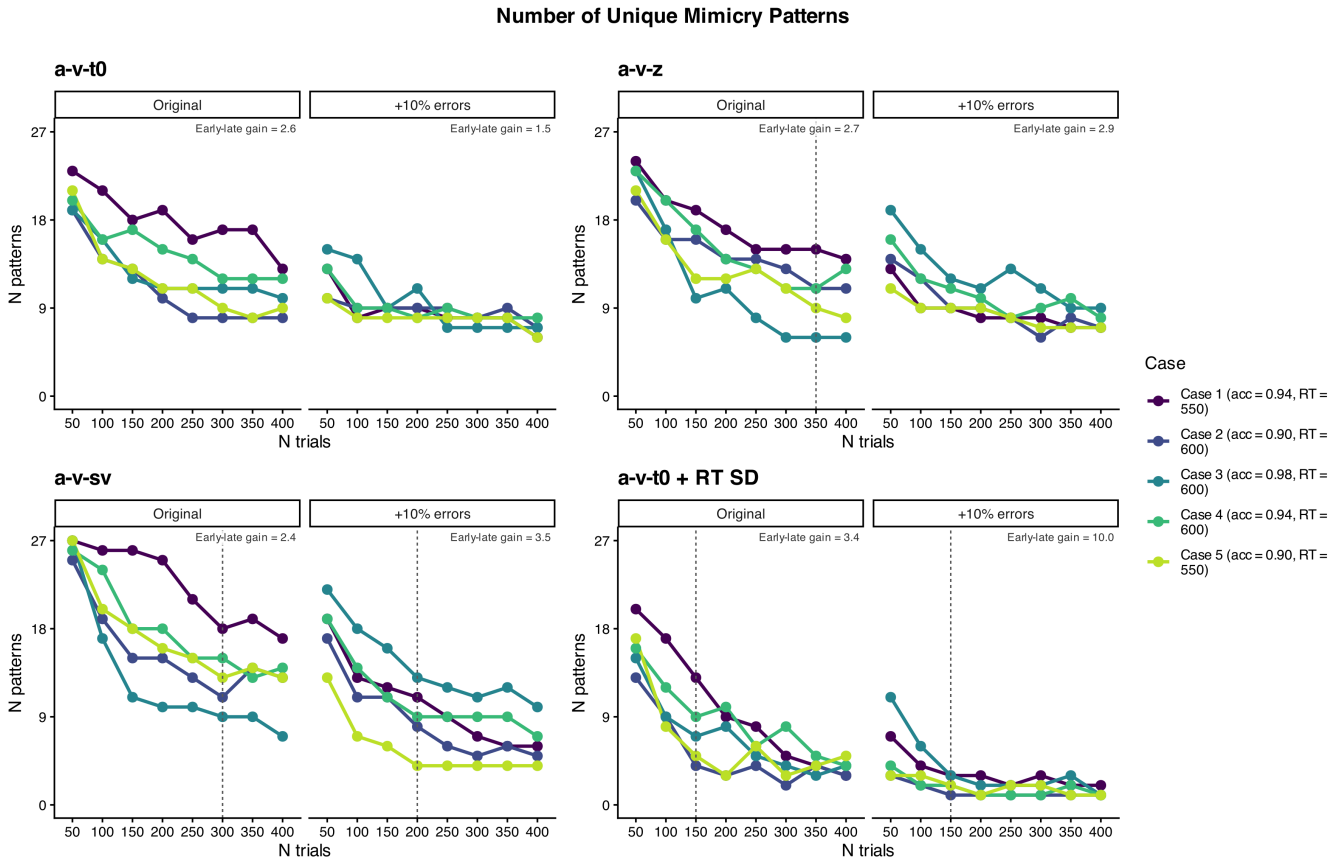

Figure 3 ■ Most common parameter patterns producing behavioural mimicry. Panels show, from top left to bottom right, $a - v - t_0$, $a - v - z$, $a - v - s_v$, and $a - v - t_0$ when matching within tolerance on accuracy, mean RT, and RT standard deviation. Symbols indicate whether each parameter was smaller than ($<$), equal to ($=$), or larger than ($>$) the baseline reference values. Pattern symbols are ordered according to the parameter names shown for each panel; for example, in an $a - v - s_v$ panel, the first symbol refers to a, the second to v, and the third to s_v. The six most prevalent mimicry patterns are shown for each model. The parameter labels are a for boundary separation, v for drift rate, s_v for drift-rate variability, t_0 for nondecision time, and z for starting point bias.

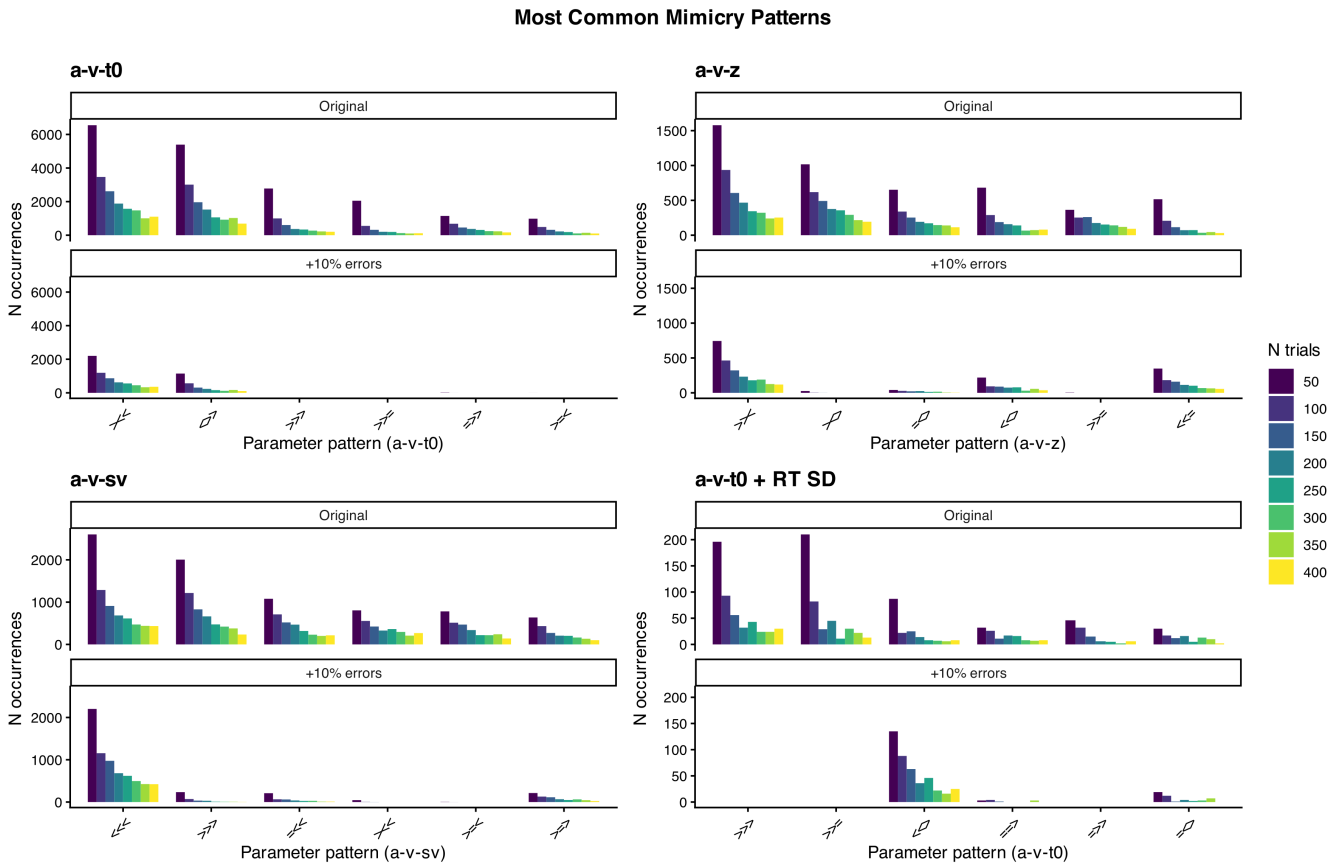

Figure 4 ■ Effects of simulation factors on the number of admissible parameter sets that span multiple behavioural cases. Panels show, from top left to bottom right, $a - v - t_0$, $a - v - z$, $a - v - s_v$, and $a - v - t_0$ when matching within tolerance on accuracy, mean RT, and RT standard deviation. Within each panel, the left facet shows the original target cases and the right facet shows the +10% error cases. Vertical dashed lines mark the half-reduction threshold based on the number of solutions spanning at least two cases. In-panel labels report the early-late gain ratio, defined as the 50–to–200-trial reduction divided by the 200–to–400-trial reduction. For readability, the y-axis scales differ across panels. The parameter labels are a for boundary separation, v for drift rate, s_v for drift-rate variability, t_0 for nondecision time, and z for starting point bias.

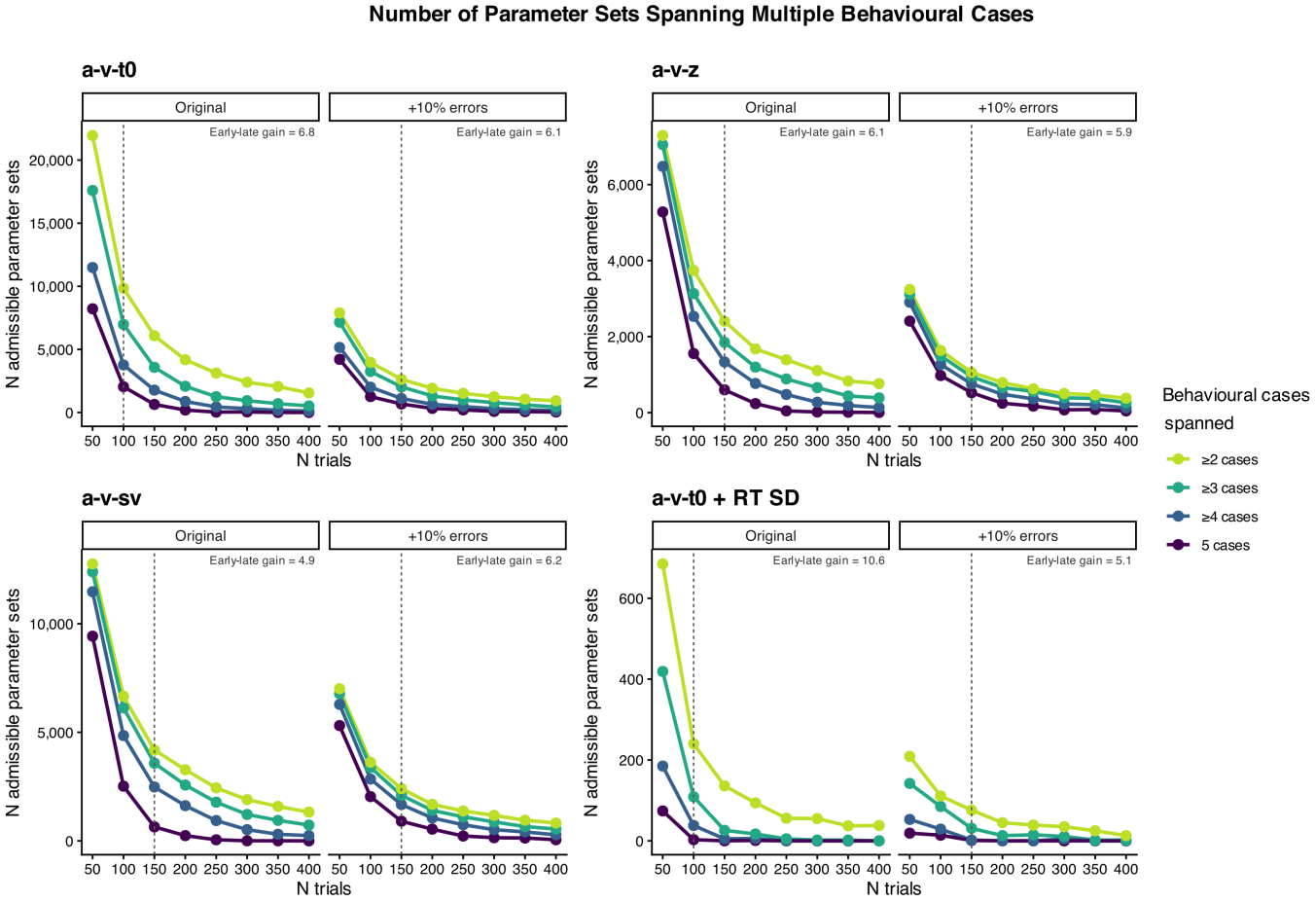

Figure 5 ■ Key omnibus summaries across model parameterisations and target sets. Each panel plots the mean-over-trials summary for one omnibus outcome, averaged across the sampled trial counts from 50 to 400 per condition. For the mimicry prevalence outcome and top-N pattern concentration outcomes, values are shown as percentages; for diversity of unique patterns and cross-case span outcomes, values are raw counts. Open circles show the original target cases and filled circles show the +10% error cases, in which target accuracies were reduced by 10 percentage points while mean RT targets were unchanged. Grey line segments link the two target sets within each parameterisation to show the direction and magnitude of change. Together, the panels summarise how parameterisation, added error information, and inclusion of RT variability affect the amount, diversity, concentration, and cross-case generality of behavioural mimicry. The parameter labels are a for boundary separation, v for drift rate, s_v for drift-rate variability, t_0 for nondecision time, and z for starting point bias.