VDA: A VARIANCE DECOMPOSITION ANALYSIS FOR PARAMETER SCREENING BY SEPARATING SIGNAL AND STOCHASTICITY IN SIMULATIONS

Kaidong Hu[a], Mubbasir Kapadia[a], Vladimir Pavlovic[a], and Sejong Yoon[b]

[a]Rutgers University, New Jersey, US
[b]The College of New Jersey, New Jersey, US

ABSTRACT

Agent-based evacuation simulations depend on many parameters spanning physical and cognitive factors, yet downstream tasks such as calibration, validation, and design optimization need to know which parameters materially affect key outcomes like collision risk. Identifying key parameters is challenging because simulation outputs are inherently noisy: two runs with identical parameters can yield different outcomes due to stochasticity. We propose Variance Decomposition Analysis (VDA), a replication-based approach that estimates how much variance a parameter contributes relative to stochastic noise. Applying VDA to 10 agent parameters across approximately 9 million runs, we find that stochasticity dominates—even the most influential parameter explains less than 25% of observed variance. VDA recovers a stable parameter hierarchy while resolving ambiguities that plague traditional pairwise t-tests. We further show that population heterogeneity amplifies cognitive parameter effects and identify a "dominant perception" mechanism explaining the anomalous vision angle behavior.

Keywords: variance decomposition, crowd evacuation, parameter screening, stochasticity, agent-based modeling.

1 INTRODUCTION

One of the popular applications of agent-based simulations is to study pedestrian evacuation and other safety-critical crowding scenarios. These models expose many parameters, including but not limited to physical properties (e.g., body size) and cognition or perception (e.g., reaction time, field of view). On the other hand, downstream tasks such as calibration [1], validation [2], and design optimization [3] depend on knowing which parameters significantly affect key performance outcomes such as collision risk.

However, finding the critical set of parameters in such applications is challenging because simulation outputs are often inherently noisy, i.e., two runs with identical simulator parameters can produce significantly different outcomes due to stochasticity and complex multi-agent interactions occur in the middle of simulations. This makes naïve significance testing and correlation-based parameter selection brittle and, in practice, leads to uncertainty about which parameters should be tuned versus fixed.

A classic pipeline for range-based sensitivity analysis discretizes a parameter into bins, runs many simulations per bin, then performs pairwise t-tests between bins and counts the number of significant differences. This approach suffers from several failure modes.

First, it introduces ambiguity: when a parameter shows some but not many significant bin-pair differences (e.g., 4–6 out of 10 comparisons for 5 bins), there is no principled rule for deciding if the parameter should be considered important or not. Second, it exhibits sample-size sensitivity: with large sample sizes, tiny effects may become statistically significant in the statistical testing, even if practically irrelevant. Third, it

may face computational overhead: increasing the number of bins from 5 to 100 increases pairwise tests per parameter from 10 to $\binom{100}{2} = 4{,}950$. Finally, arbitrarily defined, different bin boundaries can yield different conclusions.

We propose **Variance Decomposition Analysis (VDA)**, which replaces the classic discretization-and-many-tests approach with a replication-based design that directly estimates how much variance a parameter contributes relative to stochastic noise. VDA yields an interpretable effect size, is stable to sample size in the way p-values are not, and requires only a single bootstrap-based test per parameter. With a set of agent-based evacuation simulation experiments, we show the proposed VDA's benefits and utility over the traditional statistical testing method. Code for this paper can be found in https://github.com/kd-research/VDAtraj.

2 RELATED WORK

2.1 Crowd Simulation Models

Agent-based models have long been used to study pedestrian crowds and evacuation. The Social Force Model [4] became a foundational microscopic approach, reproducing phenomena such as lane formation and, in emergency contexts, panic-driven stampedes and clogging at exits [5]. Fire safety research developed evacuation simulators with diverse methodologies [6]. Over three decades, the field matured into a rich interdisciplinary area with comprehensive surveys cataloguing cellular automata, force-based, and continuum approaches [7]. Recent systematic reviews note persistent challenges including limited empirical data for calibration, behavioral realism requirements, and high computational costs [1].

2.2 Stochasticity and Replications

A key insight from simulation methodology is that stochastic noise significantly affects outcome measures, necessitating careful experimental design [8]. Unlike deterministic analyses, a single simulation run is merely one sample from a distribution of possible outcomes. Generally, it is advised using multiple independent replications to estimate metrics with confidence [9, 10]. In agent-based models, outputs exhibit considerable run-to-run variability, so determining adequate run counts is essential for variance stability [11]. Lorscheid et al. advocate systematic design-of-experiments approaches combining factorial sampling with repeated runs to improve reproducibility [12]. The literature strongly suggests that replication-aware experiment design is crucial for stochastic evacuation models to separate true signal from noise.

2.3 Sensitivity Analysis

Given the high-dimensional parameter space of evacuation models, researchers commonly apply sensitivity analysis to determine which inputs most influence outcomes. The Morris screening method [13] performs one-at-a-time random perturbations to compute "elementary effects" for each factor, efficiently identifying which parameters have negligible versus large nonlinear or interaction effects. For quantitative variance apportioning, Sobol's method partitions total output variance into contributions from each input and their interactions [14], yielding first-order and total-order sensitivity indices. These methods have been applied in crowd simulation to understand which behavioral parameters most affect evacuation time or congestion [15, 16, 17, 18, 19].

However, variance-based sensitivity analysis is most straightforward when the model evaluation is effectively deterministic, or when the response at each input is estimated with enough replications to treat random error as negligible. For stochastic simulators, there are two standard options to remedy this. First, we can run sufficient independent replications per design point, so sensitivity is computed on the estimated mean response. Second, we can build an explicit noise-aware metamodel, e.g., stochastic kriging [20], and

analyze the surrogate instead [9, 8, 21, 22]. In comparison, mean-response (replicate-averaged) sensitivity treats noise as a nuisance and does not report it as a first-class quantity, while surrogate-based workflows shift the burden to fitting and validating a metamodel. Our VDA method addresses both limitations by using replication-based variance decomposition: we use paired replications at matched parameter settings to estimate within-setting stochastic noise, then convert the corresponding between-setting component into a noise-normalized effect size that is directly comparable to variance-based global sensitivity measures, e.g., Sobol indices [14]. VDA estimates both signal and noise directly from the same replications and reports a per-parameter impact score. Carmona et al. [23] recently proposed decomposing global sensitivity analysis (GSA) results for stochastic agent-based model into deterministic and stochastic sensitivity indices, reporting each input's contribution as a fraction of total variance. Our VDA takes a different approach: rather than estimating Sobol indices via a full sample matrix, we use paired replications at matched parameter settings to directly measure noise, then normalize the between-setting variance by this noise baseline. The resulting effect size quantifies parameter influence relative to the noise baseline, rather than as a fraction of total variance, connecting parameter screening to classic signal-to-noise decomposition.

3 METHOD: REPLICATION-BASED VARIANCE DECOMPOSITION

In this section, we describe the proposed Variance Decomposition Analysis (VDA), which leverages the law of total variance to quantify parameter influence through controlled replication experiments.

3.1 Problem Formulation

Consider a stochastic simulation system where a model parameter X may influence an outcome metric Y (e.g., collision count or evacuation time). Repeated simulations with identical parameters may produce different outcomes due to random seeds, agent initialization, or other stochastic elements. We seek to quantify whether variation in X meaningfully changes the distribution of Y beyond this inherent simulation noise by designing experiments where the noise component is directly measurable. By comparing runs that share the same parameter value against runs with different parameter values, we can isolate the parameter effect from the noise baseline.

The theoretical foundation for the proposed VDA rests on the law of total variance. Given input X and output Y, we can decompose the total variance of an outcome into within-group and between-group components as: $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$. The first term, $\mathbb{E}[\mathrm{Var}(Y|X)]$, represents the expected variance of outcomes when the parameter is held fixed and this is the irreducible noise component arising from simulation stochasticity. The second term, $\mathrm{Var}(\mathbb{E}[Y|X])$, captures how the expected outcome varies as the parameter changes and this is the component we wish to analyze.

3.2 Triple-Replication Design

We define a scenario as a combination of particular simulation configuration. For each scenario i with fixed environment and initial conditions, we generate three simulation runs following a structured protocol depicted in Figure 1. The *base run* uses parameter value $X = x$ and produces output H_i. The *truthy run* uses the same parameter value $X = x$ but employs a different random number generator seed, producing output H_i'. The *random run* uses a different parameter value $X = x'$ drawn independently from the same distribution, producing output H_i''.

From the three output values, we compute two difference quantities:

$$D_{\text{same},i} = H_i - H_i', \tag{1}$$

$$D_{\text{random},i} = H_i - H_i''. \tag{2}$$

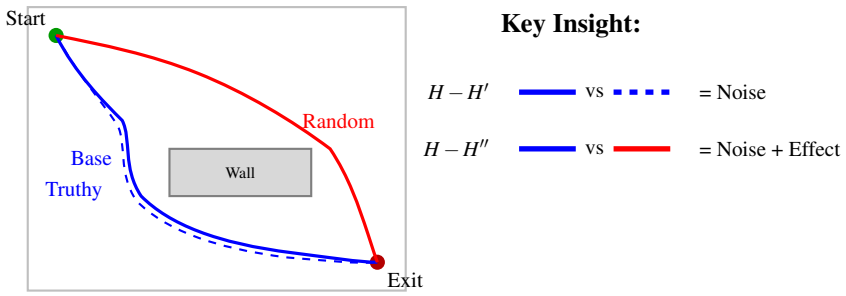


Figure 1: Triple-replication design. Base and Truthy share parameter value $X = x$ but differ in random seed (blue); Random uses $X = x'$ (red). $H - H'$ measures pure noise; $H - H''$ measures noise plus parameter effect.

$D_{\text{same},i}$ captures noise only, since both runs share the same parameter value. $D_{\text{random},i}$ captures noise plus any parameter effect. Below, we use D_{same} to denote all $D_{\text{same},i}$ values across scenarios. As a concrete example, suppose we want to test an agent radius parameter with range $[0.1, 0.5]$ meters. For each scenario i, we sample a base parameter value x_i uniformly from this range, then execute the triple-replication: one base run with x_i, one truthy run with the same x_i but a different random seed, and one random run with an independently sampled x_i' from the same range. This yields one $D_{\text{same},i}$ and one $D_{\text{random},i}$ per scenario. Across N scenarios, we collect all differences into the sets $D_{\text{same}} = \{D_{\text{same},1}, \ldots, D_{\text{same},N}\}$ and $D_{\text{random}} = \{D_{\text{random},1}, \ldots, D_{\text{random},N}\}$.

3.3 Impact and Effect Size

We compute sample variances $\text{Var}_{\text{same}} = \text{Var}(D_{\text{same}})$ and $\text{Var}_{\text{random}} = \text{Var}(D_{\text{random}})$ from the observed differences, and define the *impact* as

$$\text{Impact} = \text{Var}_{\text{random}} - \text{Var}_{\text{same}}\,. \tag{3}$$

Proposition 1. *Let H and H' be independent draws from $Y|X = x$, and let H'' be drawn from $Y|X = x'$ where x' is sampled independently from the same distribution as x. Assume that stochastic noise is i.i.d. and independent of X. Define $D_{same} = H - H'$ and $D_{random} = H - H''$. Then*

$$\textit{Impact} = \text{Var}(D_{random}) - \text{Var}(D_{same}) = 2\,\text{Var}(\mathbb{E}[Y|X]).$$

Proof. Given $X = x$, we have $\mathbb{E}[D_{\text{same}}|X] = 0$ and $\text{Var}(D_{\text{same}}|X) = 2\,\text{Var}(Y|X)$ by independence of H and H'. By the law of total variance, $\text{Var}(D_{\text{same}}) = 2\mathbb{E}[\text{Var}(Y|X)]$. Since X and X' are independent, H and H'' are unconditionally independent, so $\text{Var}(D_{\text{random}}) = 2\,\text{Var}(Y) = 2\,\text{Var}(\mathbb{E}[Y|X]) + 2\mathbb{E}[\text{Var}(Y|X)]$ by the law of total variance. Subtracting yields $\text{Impact} = 2\,\text{Var}(\mathbb{E}[Y|X])$. □

We define the *effect size* by normalizing the impact by the noise baseline:

$$ES = \frac{\text{Impact}}{\text{Var}_{\text{same}}}. \tag{4}$$

An effect size of $ES = 0.30$ means that varying the parameter adds variance equal to 30% of the noise baseline. We assess the statistical significance via a paired bootstrap, testing $H_0 : \text{Impact} \leq 0$ against $H_1 : \text{Impact} > 0$.

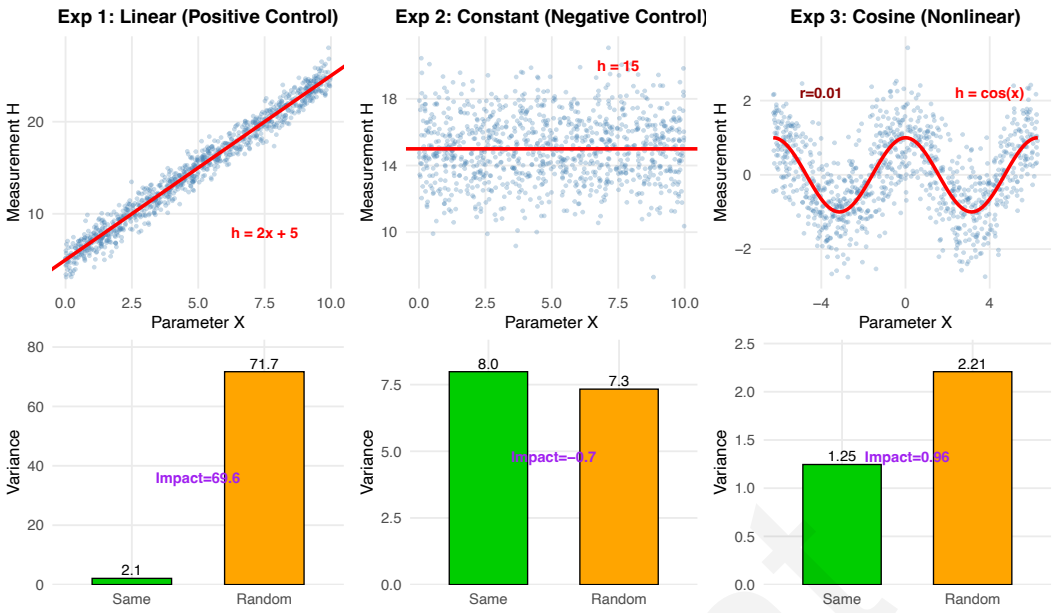


Figure 2: Synthetic validation. Top: ground-truth relationships. Bottom: variance comparisons (green = same, orange = random). Exp 1: linear detection. Exp 2: no false positives. Exp 3: VDA detects nonlinear effects missed by correlation ($r \approx 0$).

3.4 Validation using Synthetic Data

Before applying VDA to the agent-based simulation data, we validate the method through controlled synthetic experiments where ground truth is known. These experiments demonstrate sensitivity (detecting genuine effects), specificity (avoiding false positives), and superiority over traditional correlation analysis. We conducted three experiments and the results are depicted in Figure 2.

Experiment 1: Positive Control. We simulate $h(x) = 2x + 5 + \mathcal{N}(0,1)$ with $x \sim \text{Uniform}(0,10)$. The theoretical impact variance is $2 \cdot 2^2 \cdot \text{Var}(X) \approx 66.7$. Using $N = 1000$ triple-replication samples, VDA recovers $\text{Var}_{\text{same}} \approx 2.07$ and Impact ≈ 67.86, matching theory within 0.2% relative error ($p < 0.001$).

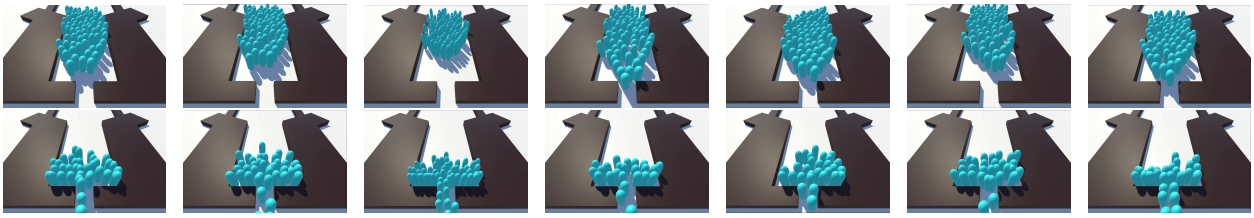
Experiment 2: Negative Control. We simulate $h(x) = 15 + \mathcal{N}(0,4)$ where output is independent of parameters. VDA correctly identifies no effect: Impact ≈ 0.38 relative to noise baseline $\text{Var}_{\text{same}} \approx 7.70$, with $p = 0.324$ (correctly failing to reject H_0).

Experiment 3: Nonlinear Detection. We simulate $h(x) = \cos(x) + \mathcal{N}(0, 0.64)$ with $x \in [-2\pi, 2\pi]$. Correlation analysis yields $r \approx 0.009$, incorrectly suggesting no influence (cosine is symmetric). VDA correctly detects the relationship: Impact ≈ 1.04, $ES \approx 0.81$, $p < 0.001$. This demonstrates VDA's superiority for non-monotonic parameter effects common in complex simulations.

4 EXPERIMENTAL SETUP

4.1 Simulator and Scenario

We use a pedestrian evacuation corridor testbed with a 20 meters $\times$ 7 meters geometry containing 75 agents navigating through a 4 meter-wide corridor. This configuration creates sufficient density for meaningful agent-agent interactions while remaining computationally tractable for large-scale replication studies.



(a) (b) (c) (d) (e) (f) (g)

Figure 3: Evacuation simulations with varying parameters. Top: entering bottleneck. Bottom: exiting. (a) Default (b)–(c) Agent Radius (P2) max/min (d)–(e) Max Speed (P4) max/min (f)–(g) Reaction Time (P8) max/min.

Table 1: Agent parameters evaluated in this study with their sampling ranges.

PID	Parameter Name	Range	Description
2	Agent Radius	0.15–0.30 m	Physical body radius
3	Body Mass	60–100 kg	Body mass
4	Maximum Speed	1.0–2.0 m/s	Maximum walking speed
5	Maximum Force	1.0–2.0 N	Maximum physical force
6	Query Radius	0.5–1.5 m	Detection radius
7	Vision Angle	0.78–2.35 rad	Visual field width
8	Reaction Time	0.1–1.0 s	Perception delay
9	Vision Resolution	10–100 rays	Sampling directions
10	Vision Range	1.0–20.0 m	Perception distance
11	Repulsion Coef.	2000–8000	Contact repulsion

We borrowed a setup similar to the one used in a data-driven parameter exploration [3, 24]. They built a simulation testbed based on the SteerSuite [25] simulator, which provides the core infrastructure for scenario configuration, agent initialization, and trajectory recording. We also used Unity for visualization purposes. Figure 3 shows some examples of evacuation simulations.

For pedestrian steering behavior, we implemented an egocentric, vision-based cognitive model following the heuristics proposed by Moussaïd et al. [26]. This model represents a departure from traditional force-based approaches (such as Social Forces or ORCA) and instead models pedestrian navigation through two simple cognitive procedures guided by visual information. Each agent perceives its environment by sampling the distance to obstructions within a set of candidate sight lines determined by its vision field angle and vision range. Based on this visual input, agents adapt their walking direction to minimize visual field occlusion while maintaining progress toward their goal.

4.2 Parameters Under Study

We screen 10 agent parameters spanning physical and cognitive factors. Table 1 summarizes each parameter with its tested range. These parameters fall into two categories: **physical parameters** governing body properties and contact forces (P2, P3, P5, P11), and **cognitive/perception parameters** governing movement decisions and sensory capabilities (P4, P6, P7, P8, P9, P10). For the parameter index, we borrowed the definition from Hu et al. [24]. Since P1 is the doorway width which is not an agent parameter, we omit it in this study.

Table 2: Parameter significance ($p < 0.05$) across all eight population type $\times$ metric combinations. ✓ = statistically significant.

Pop. Type	Outcome	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Hetero.	Time	✓	–	✓	–	–	–	✓	–	–	–
Hetero.	Collision	✓	–	✓	✓	–	–	✓	–	–	–
Hetero.	Distance	✓	–	✓	✓	–	–	✓	–	–	–
Hetero.	PLE	✓	–	✓	–	–	–	✓	–	–	–
Homo.	Time	✓	–	✓	–	–	✓	✓	–	–	–
Homo.	Collision	✓	–	✓	✓	–	–	✓	–	–	–
Homo.	Distance	✓	–	–	✓	–	✓	–	–	–	–
Homo.	PLE	✓	–	✓	✓	–	✓	✓	–	–	–

4.3 Outcomes and Population Structures

We evaluate parameter influence on four trajectory-based outcome metrics originally defined in [3, 24]: **collision times**, the average number of collisions per agent (count); **agent time enabled**, the average duration each agent is active in the simulation (seconds); **distance traveled**, the average path length traversed per agent (meters); and **PLE energy**, the average estimated locomotion effort per agent following the principle of least effort.

We also study two distinctive population types: In **heterogeneous** simulations, each agent has a different value for the active parameter while other parameters remain constant. In **homogeneous** simulations, all agents share identical parameter values. In the results, we show that this distinction is critical as some parameters show influence only when uniformly applied or only when varied across agents.

4.4 Dataset Scale

For each parameter and population type combination, we generate 2,000 scenarios with three runs per scenario (base, truthy, and random) and 75 agents per scenario, yielding 450,000 runs per parameter per population. With 10 parameters and two population types, the complete dataset comprises approximately 9 million simulation runs. All bootstrap analyses used 1,000 iterations at a 95% confidence level.

5 RESULTS AND DISCUSSION

Table 2 provides an overview of parameter significance across all simulation conditions and measurements. Three parameters (P2, P4, P8) show consistent significance across most conditions, while five parameters (P3, P6, P9, P10, P11) show no significance in any condition. Notably, P7 shows significance only in homogeneous conditions. We will discuss this observation in detail, in Section 5.5.

5.1 Parameter Impact Ranking for Physical Safety Metric

Table 3 shows that VDA yields a clear separation for the primary safety metric (collision times) under heterogeneous conditions. Three parameters exceed the practical significance threshold of $ES \geq 0.10$: P8 reaction time (0.30), P4 maximum speed (0.20), and P2 agent radius (0.18) as depicted in Figure 4. P5 (maximum force) presents an interesting case. It is statistically significant ($p = 0.001$), but with $ES = 0.015$, we can say that its practical impact on the tested measures in this simulation scenario is negligible.

Table 3: VDA results for collision times under the heterogeneous population condition. Parameters are sorted by Impact value. YES* indicates statistical significance ($p < 0.05$) but practical negligibility ($ES < 0.10$).

Parameter	Impact	Effect Size (ES)	p-value	Significant?
P8 (Reaction Time)	262.97	0.2995	< 0.001	YES
P4 (Maximum Speed)	225.91	0.2015	< 0.001	YES
P2 (Agent Radius)	206.81	0.1814	< 0.001	YES
P5 (Maximum Force)	17.87	0.0145	0.001	YES*
P10 (Vision Range)	3.00	0.0025	0.293	NO
P7 (Vision Angle)	1.15	0.0009	0.405	NO
P3 (Body Mass)	−0.54	−0.0004	0.513	NO
P6 (Query Radius)	−1.63	−0.0014	0.623	NO
P9 (Vision Resolution)	−6.10	−0.0050	0.865	NO
P11 (Repulsion Coef.)	−7.31	−0.0062	0.919	NO

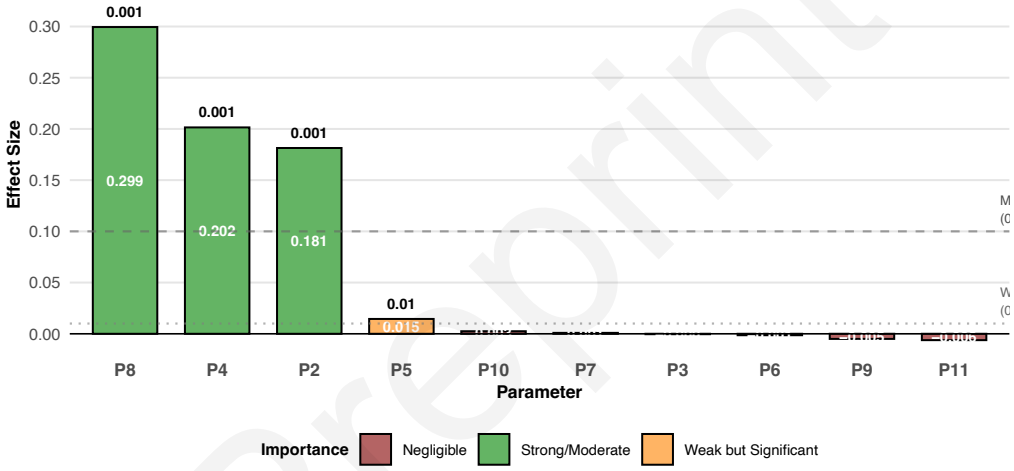


Figure 4: Effect size ranking for collision times under the heterogeneous population condition. Three parameters (P8, P4, P2) exceed the practical significance threshold ($ES \geq 0.10$), while P5 shows statistical but not practical significance.

5.2 Noise Versus Signal Decomposition

To illustrate the variance decomposition capability of VDA, we consider P8 (reaction time), the parameter with the largest impact. The same-parameter variance $\text{Var}_{\text{same}} = 878$ establishes the noise baseline, while $\text{Var}_{\text{random}} = 1,141$ captures noise plus signal. The difference yields Impact $= 263$, giving $ES = 263/878 \approx 0.30$. This means the outcome decomposes into 77% stochastic noise and 23% signal. In other words, stochasticity is large in this evacuation setting. Even the strongest parameter explains less than 25% of total variance, so replication is required to distinguish signal from stochasticity.

5.3 Comparison to Pairwise t-tests

VDA can address two common failure modes of traditional screening: the gray zone and false positive. Suppose we run pairwise t-tests of the 10 parameters we used in this study. We partition each parameter's

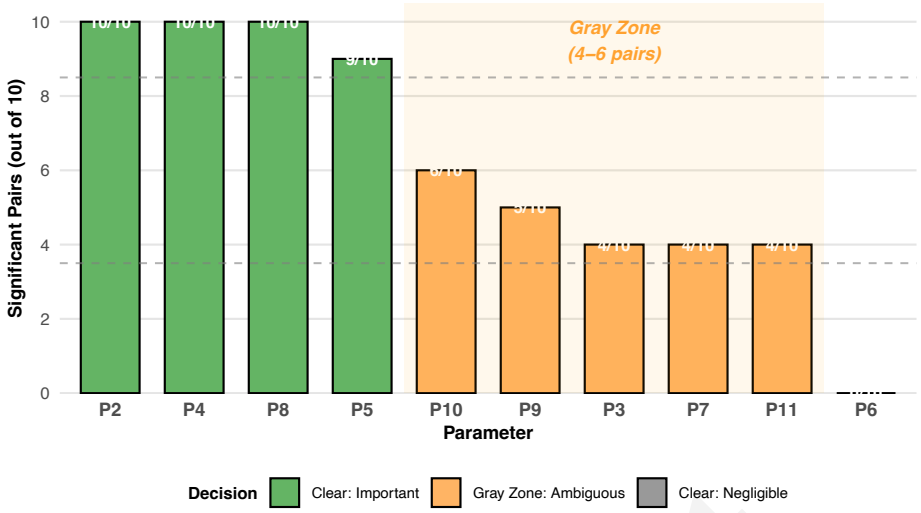


Figure 5: Pairwise t-test results for collision times under the heterogeneous population condition. Five parameters (P3, P7, P9, P10, P11) fall in the gray zone with $4 \sim 6$ significant pairs out of 10, providing no clear guidance on parameter importance.

possible value range into five bins, and choose two bins at a time to see if parameters selected from different bins within a pair can be distinguishable by conducting t-test on the outcome. If we can see a significant difference between the two, then we may say the parameter has impact on the outcome.

Figure 5 shows the result of this analysis. We ranked the parameters based on the number of significant pairs in the test. As it can be seen, several parameters show $4 \sim 6$ significant bin-pair differences out of 10 comparisons for the five bins, leaving practitioners uncertain whether the parameter matters or not. VDA resolves this ambiguity by showing all such parameters have $ES < 0.02$, identifying them as clearly negligible, as shown in Figure 6.

VDA can also identify false positive in the pairwise t-test. The t-test results suggest P5 (Maximum Force) importance with 9/10 significant comparisons, but VDA finds $ES = 0.015$, practically negligible despite the statistical significance.

5.4 Population Structure Effects

Another interesting finding is that the cognitive parameters amplify their impact under heterogeneity. Table 4 shows the comparison. P8 (Reaction Time) and P4 (Maximum Speed) show much larger effects when agents have diverse parameter values, suggesting interaction effects between differently-configured agents. On the other hand, physical constraints remain stable. P2 (Agent Radius) shows nearly identical effect sizes regardless of population structure. P5 (Maximum Force) doubles its effect size in homogeneous setting, but still the smallest among the four largest-effect parameters.

5.5 The P7 (Vision Angle) Anomaly: Dominant Perception

As shown in Table 5, P7 (Vision Angle) shows a unique pattern compared to other cognitive parameters. It is significant only in homogeneous populations. P7 shows significance in 0/4 heterogeneous conditions but 3/4 homogeneous conditions.

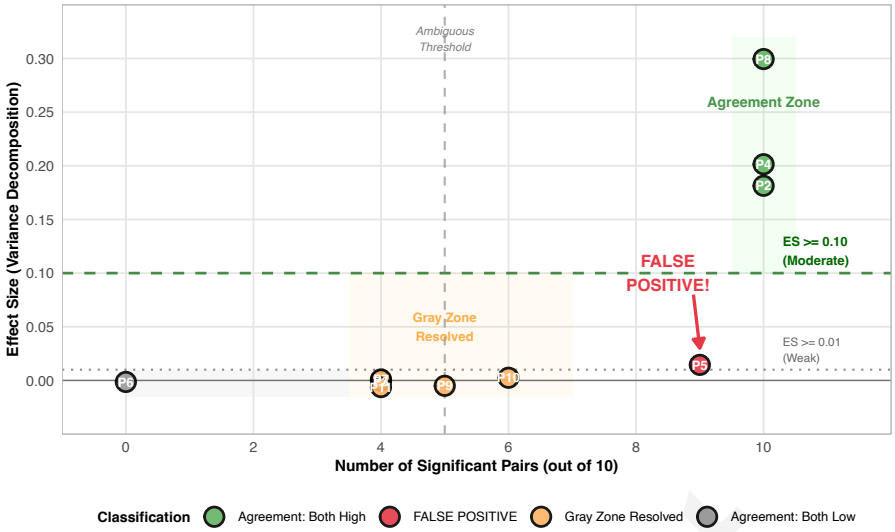


Figure 6: T-test significant pairs vs. VDA effect size. P5: high significance (9/10 pairs) but negligible $ES = 0.015$. Gray-zone parameters (4–6 pairs) are resolved as negligible by VDA.

Table 4: Effect size comparison for collision times: heterogeneous vs homogeneous population types.

Parameter	ES in Heterogeneous Population	ES in Homogeneous Population	Ratio
P8 (Reaction Time)	0.300	0.049	6.1
P4 (Maximum Speed)	0.202	0.094	2.1
P2 (Agent Radius)	0.181	0.205	0.9
P5 (Maximum Force)	0.015	0.033	0.5

To explain this observation, we hypothesize a *dominant perception mechanism*. In heterogeneous crowds with varying vision field angles, the bi-directional nature of pedestrian interactions means the agent with the wider field of view will always detect the interaction first and initiate avoidance behavior. This asymmetry effectively neutralizes the parameter's influence at the population level. In contrast, when all agents share the same vision angle (homogeneous setting), there is no compensatory agent, and the effect of vision field limitation collectively affect the outcome measures. Figure 7 describes the dominant perception mechanism.

Table 5: P7 (Vision Angle) effect sizes and p-values across conditions.

Metric	ES in Hetero.	p-value in Hetero.	ES in Homo.	p-value in Homo.
Time Enabled	-0.011	0.867	0.027	0.001
Collision Times	0.001	0.405	-0.004	0.799
Distance Traveled	-0.017	0.846	0.066	< 0.001
PLE Energy	-0.012	0.854	0.036	< 0.001

6 CONCLUSION

This paper presents an empirical parameter-importance study demonstrating that stochasticity can dominate evacuation simulation outcomes and that replication-based evaluation is essential for reliable conclusions regarding parameter sensitivity. Using a large-scale triple-replication dataset comprising approximately 9

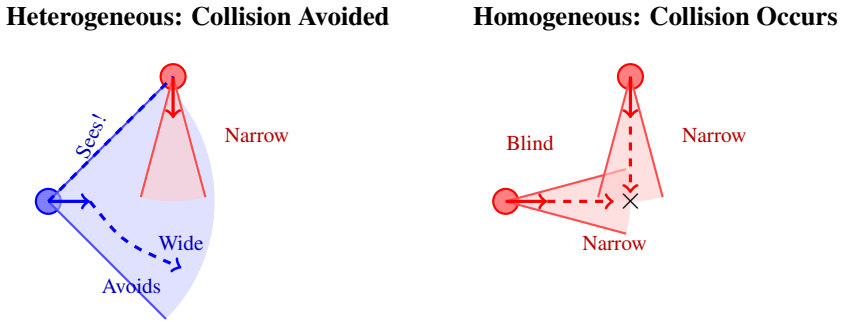


Figure 7: Dominant perception mechanism. Left: in heterogeneous populations, the wider-vision agent (blue) detects and avoids the other. Right: in homogeneous populations with narrow vision, neither agent detects the other, causing collision.

million simulation runs, we applied Variance Decomposition Analysis (VDA) to separate stochastic noise from parameter-driven signal.

Our central finding is that stochasticity is large in this setup. Even the most influential parameter (P8, Reaction Time) explains less than 25% of observed variance in collision counts, with 77% attributable to simulation noise. Despite substantial noise, VDA recovers a stable parameter hierarchy: three parameters, namely P8 (Reaction Time), P4 (Maximum Speed), and P2 (Agent Radius), exceed the practical significance threshold and materially affect collision outcomes, while the remaining seven tested parameters are negligible for this metric.

We further show that population heterogeneity changes parameter importance, with cognitive parameters amplifying under heterogeneous conditions and P7 (Vision Angle) exhibiting a "dominant perception" anomaly where it matters only in homogeneous populations.

Methodologically, VDA offers advantages over the traditional binning-and-testing approaches. It produces sample-size-invariant effect sizes, resolves gray-zone ambiguity, and avoids false positives. For practitioners working with stochastic evacuation models, VDA can provide a principled basis for deciding which parameters warrant calibration effort and which can be safely fixed at default values.

7 LIMITATIONS AND FUTURE WORKS

The proposed method and the evaluation we present in this paper has room for improvements. First, our experiments are based on a single scenario geometry. Findings are specific to 20m x 7m corridor with 75 agents. Second, our VDA application only considered one parameter at a time. It is possible parameters may have correlations or dependencies, or interactions with each other. Third, while we made an i.i.d. assumption in bootstrap, some evacuation simulation scenarios may have latent structure to be considered. In addition to ways to address these limitations, we aim to investigate following aspects in the future: (1) A systematic noise-vs-scale study, (2) diverse environment geometries, (3) alternative behavioral simulation models (social force, ORCA, RVO).

ACKNOWLEDGMENTS

The manuscript was drafted with an assistance of Claude Opus 4.5. All sections were human-verified and corrected.

REFERENCES

[1] G. P. Senanayake, M. Kieu, Y. Zou, and K. Dirks, "Agent-based simulation for pedestrian evacuation: A systematic literature review," *International Journal of Disaster Risk Reduction*, vol. 111, p. 104705, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.ijdrr.2024.104705

[2] R. G. Sargent, "Verification and validation of simulation models," *J. Simul.*, vol. 7, no. 1, pp. 12–24, Feb. 2013.

[3] K. Hu, S. Yoon, V. Pavlovic, P. Faloutsos, and M. Kapadia, "Predicting crowd egress and environment relationships to support building design optimization," *Computers & Graphics*, vol. 88, p. 83–96, May 2020. [Online]. Available: http://dx.doi.org/10.1016/j.cag.2020.03.005

[4] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, p. 4282–4286, May 1995. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.51.4282

[5] D. Helbing, I. Farkas, and T. Vicsek, "Simulating dynamical features of escape panic," *Nature*, vol. 407, no. 6803, p. 487–490, Sep. 2000. [Online]. Available: http://dx.doi.org/10.1038/35035023

[6] S. Gwynne, E. Galea, M. Owen, P. Lawrence, and L. Filippidis, "A review of the methodologies used in the computer simulation of evacuation from the built environment," *Building and Environment*, vol. 34, no. 6, p. 741–749, Nov. 1999. [Online]. Available: http://dx.doi.org/10.1016/S0360-1323(98)00057-2

[7] D. C. Duives, W. Daamen, and S. P. Hoogendoorn, "State-of-the-art crowd motion simulation models," *Transportation Research Part C: Emerging Technologies*, vol. 37, p. 193–209, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.trc.2013.02.005

[8] J. P. Kleijnen, *Design and Analysis of Simulation Experiments*, 2nd ed., ser. International Series in Operations Research & Management Science. Springer International Publishing, 2015. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18087-8

[9] A. M. Law, *Simulation Modeling and Analysis*, 6th ed. McGraw-Hill, 2024.

[10] K. Hoad, S. Robinson, and R. Davies, "Automating des output analysis: How many replications to run," in *2007 Winter Simulation Conference*. IEEE, Dec. 2007, pp. 505–512. [Online]. Available: http://dx.doi.org/10.1109/wsc.2007.4419641

[11] J.-S. Lee, T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, G. Polhill, Z. Sun, and D. C. Parker, "The complexities of agent-based modeling output analysis," *Journal of Artificial Societies and Social Simulation*, vol. 18, no. 4, 2015. [Online]. Available: http://dx.doi.org/10.18564/jasss.2897

[12] I. Lorscheid, B.-O. Heine, and M. Meyer, "Opening the 'black box' of simulations: increased transparency and effective communication through the systematic design of experiments," *Computational and Mathematical Organization Theory*, vol. 18, no. 1, p. 22–62, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1007/s10588-011-9097-3

[13] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, p. 161–174, May 1991. [Online]. Available: http://dx.doi.org/10.1080/00401706.1991.10484804

[14] I. M. Sobol', "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1–3, p. 271–280, Feb. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0378-4754(00)00270-6

[15] M. Gödel, R. Fischer, and G. Köster, "Sensitivity analysis for microscopic crowd simulation," *Algorithms*, vol. 13, no. 7, p. 162, Jul. 2020. [Online]. Available: http://dx.doi.org/10.3390/a13070162

[16] G. Wurzer, M. Ausserer, H. Hinneberg, C. Illera, and A. Rosic, *Sensitivity Visualization of Circulation under Congestion and Blockage*. Springer US, 2011, pp. 899–902. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-9725-8_96

[17] D. Li, R. Huang, and Y. Wu, "Sensitivity analysis of pedestrian simulation on train station platforms," in *Proceedings of the 26th Conference on Computer Aided Architectural Design*

Research in Asia (CAADRIA) [Volume 2], ser. CAADRIA 2021, vol. 2. CAADRIA, 2021, pp. 529–538. [Online]. Available: http://dx.doi.org/10.52842/conf.caadria.2021.2.529

[18] E. Ronchi, P. A. Reneke, and R. D. Peacock, "A method for the analysis of behavioural uncertainty in evacuation modelling," *Fire Technology*, vol. 50, no. 6, pp. 1545–1571, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1007/s10694-013-0352-7

[19] E. Kugu, J. Li, F. D. McKenzie, and O. K. Sahingoz, "Fuzzy logic approach and sensitivity analysis for agent-based crowd injury modeling," *SIMULATION*, vol. 90, no. 3, pp. 320–336, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1177/0037549713518598

[20] B. Ankenman, B. L. Nelson, and J. Staum, "Stochastic kriging for simulation metamodeling," *Operations Research*, vol. 58, no. 2, pp. 371–382, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1287/opre.1090.0754

[21] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis. The Primer*. Wiley, 2007. [Online]. Available: http://dx.doi.org/10.1002/9780470725184

[22] J. L. Hart, A. Alexanderian, and P. A. Gremaud, "Efficient computation of sobol' indices for stochastic models," *SIAM Journal on Scientific Computing*, vol. 39, no. 4, pp. A1514–A1530, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1137/16M106193X

[23] Á. Carmona-Cabrero, R. Muñoz-Carpena, W. S. Oh, and R. Muneepeerakul, "Decomposing variance decomposition for stochastic models: Application to a proof-of-concept human migration agent-based model," *Journal of Artificial Societies and Social Simulation*, vol. 27, no. 1, 2024. [Online]. Available: http://dx.doi.org/10.18564/jasss.5174

[24] K. Hu, S. Yoon, V. Pavlovic, and M. Kapadia, "Toward realistic human crowd simulations with data-driven parameter space exploration," in *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, Jan. 2024, p. 221–225. [Online]. Available: http://dx.doi.org/10.1109/aixvr59861.2024.00035

[25] S. Singh, M. Kapadia, P. Faloutsos, and G. Reinman, "SteerBench: A benchmark suite for evaluating steering behaviors," *Computer Animation and Virtual Worlds*, vol. 20, no. 5-6, pp. 533–548, 2009.

[26] M. Moussaïd, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, p. 6884–6888, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1073/pnas.1016507108