UC Irvine
UC Irvine Previously Published Works

Title
Robust Bayesian hypothesis testing with the hierarchical EZ-DDM

Permalink
https://escholarship.org/uc/item/4rn2g7td

Journal
Behavior Research Methods, 58(7)

ISSN
1554-351X

Authors
Chávez De la Peña, Adriana F
Shin, Eunice
Vandekerckhove, Joachim

Publication Date
2026

DOI
10.3758/s13428-026-03066-1

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Robust Bayesian hypothesis testing with the hierarchical EZ-DDM

Adriana F. Chávez De la Peña[a,b,*], **Eunice Shin**[a], **and Joachim Vandekerckhove**[a,b,c,*]

This is the author final version of:

Chávez De la Peña, A.F., Shin, E. & Vandekerckhove, J. (2026). Robust Bayesian hypothesis testing with the hierarchical EZ-DDM. *Behavior Research Methods, 58,* 177. https://doi.org/10.3758/s13428-026-03066-1

The EZ-diffusion model (EZ-DDM) uses a method of moments to provide closed-form estimators for the three-parameter drift-diffusion model from summary statistics. In previous work, we showed that using the sampling distributions of these statistics enables the implementation of hierarchical EZ-DDM extensions, supporting scalable Bayesian inference in cognitive psychometrics applications. However, the summary statistics used in EZ-DDM implementations (the mean and variance of the response time distribution) are sensitive to contaminant data points, limiting its utility in real-world applications. To address this, we propose a variation on the EZ-DDM implementation in which the summary statistics are replaced with robust alternatives, substituting mean RT with median RT and RT variance with an estimate derived from the interquartile range. We explore and evaluate the effectiveness of this substitution through simulation studies using a within-subject t test design across varying sample sizes and effect sizes. We show that the robust variant matched the diagnostic accuracy of the EZ-DDM implementation on uncontaminated data while maintaining diagnostic accuracy under contamination, unlike the standard model. This extension preserves efficiency while adding robustness in real-world applications. We recommend the use of the robust EZ-DDM in practical applications.

Robust inference | hierarchical Bayesian | EZ diffusion | cognitive psychometrics | hypothesis testing

Robustness is a necessary quality for trustworthy scientific inference. Broadly defined, it refers to the capacity of a method to preserve its intended function in the presence of perturbations that threaten its integrity. We want to ensure that the methods we use do not break under the presence of small departures from the assumptions underlying their theoretical foundation (Lee et al., 2019). Such robust procedures allow valid inference even in the presence of skewness, contaminant data points, heteroscedasticity, and other irregularities commonly encountered in real-world data sets (Huber & Ronchetti, 2009).

Estimation methods that rely on summary statistics that are sensitive to contaminants and other irregularities in the data are not *robust*. This is the case for the EZ-DDM (Wagenmakers, van der Maas, & Grasman, 2007), a simplification of the drift diffusion model (Ratcliff, 1978) that enables parameter estimation for the drift rate, boundary separation, and non-decision time parameters from three summary statistics: the mean and variance of the RT distribution, and the accuracy rate. In this paper, we explore a robust implementation of the EZ-DDM, in which the mean and variance are replaced by two approximations that are less sensitive to contaminants: the median and variance estimate derived from the interquartile range (IQR). Specifically, we embed the proposed modification into the hierarchical Bayesian framework that we introduced in previous work (Chávez De la Peña & Vandekerckhove, in press) and evaluate its performance in simulation studies.

A hierarchical Bayesian EZ-DDM

In previous work, we introduced a Bayesian hierarchical EZ-DDM implementation that recasts the observed accuracy rate as the binomial rate for correct response counts and uses the sampling distributions of the mean and variance of RT to define a proxy likelihood that can be easily implemented in probabilistic programming languages (Chávez De la Peña & Vandekerckhove, in press). This proxy enables latent variable and covariate regression structures, and is well-suited for scalable inference due to its reliance on summary statistics.

A robust Bayesian EZ-DDM

The robust implementation of the EZ-DDM substitutes the mean and variance of the RT distribution with the sample median and the following variance estimate:

$$\widehat{\text{Var}}_{\text{IQR}} = \left(\frac{Q_3 - Q_1}{1.349}\right)^2 \quad [1]$$

where Q_1 and Q_3 are the first and third quartiles of the RT distribution.

Aside from the summary statistic substitution, nothing in the model changes – the robust implementation of the hierarchical Bayesian EZ-DDM continues to use the sampling distributions of the mean and variance of RTs, thus preserving the same proxy likelihood structure.

Hypothesis testing simulation design

We evaluated and compared the performance of the robust and standard implementations of the hierarchical Bayesian EZ-DDM for hypothesis testing through simulation studies. We generated trial-level data from a Wiener diffusion process with a within-subject

[a]Department of Cognitive Sciences, University of California, Irvine; [b]Department of Statistics, University of California, Irvine; [c]Department of Logic and Philosophy of Science, University of California, Irvine
All authors contributed to the final draft.
[*]Correspondence should be addressed to Adriana F. Chávez De la Peña (achavezd@uci.edu) or Joachim Vandekerckhove (joachim@uci.edu).
JV was supported by National Science Foundation grants #1850849 and #2051186.

AUTHOR VERSION

t test design on the drift rate parameter $\nu_{p,k}$ given by the following hierarchical meta-regression model:

$$\begin{aligned} \nu_{p,k}^{\text{pred}} &= \mu_\nu + \beta X_k \\ \nu_{p,k} &\sim \mathcal{N}\left(\nu_{p,k}^{\text{pred}},\ \sigma_\nu^2\right) \\ \alpha_p &\sim \mathcal{N}\left(\mu_\alpha,\ \sigma_\alpha^2\right) \\ \tau_p &\sim \mathcal{N}\left(\mu_\tau,\ \sigma_\tau^2\right) \end{aligned} \quad [2]$$

such that the predicted drift rate for participant p in condition k ($\nu_{p,k}^{\text{pred}}$) is defined as the linear combination of a population-level drift rate intercept μ_ν and the population-level regression coefficient capturing the condition effect β (with $X_k \in \{0, 1\}$ as the condition indicator). In this data-generating model, the dispersion of the person-by-condition drift rates $\nu_{p,k}$ around these predicted values is captured by the variance parameter σ_ν^2, while hierarchical parameters μ_α, μ_τ, and σ_α^2, σ_τ^2 capture the mean and variance of the population-level distributions of the boundary separation α_p and nondecision time τ_p parameters, respectively.

In our simulation studies, we generated 1,000 independent data sets for each design cell defined by (1) number of participants P, (2) number of trials per condition T, and (3) true fixed effect size β. Across all data sets generated, we fixed the value of all variance parameters $\sigma_\nu^2 = 0.75, \sigma_\alpha^2 = 0.5, \sigma_\tau^2 = 0.1$. The rest of the true parameters were generated by drawing population-level values for μ_ν, μ_α, and μ_τ from uniform distributions (see below), and then sampling individual-level boundary separation (α_p) and nondecision time (τ_p) parameters from the corresponding group-level normal distributions. Similarly, we drew participant-specific drift rates $\nu_{p,k}$ from the regression structure described in Equation 2.

Comparison conditions. Each simulated dataset consisted of clean trial-level data from a Wiener diffusion process. For each dataset, we constructed a contaminated counterpart by replacing 5% of each participant's trials with contaminant data (distributed equally over conditions), while leaving the remaining 95% of trials identical to the clean data. From both versions (clean and contaminated), we then computed standard and robust summary statistics, yielding a 2×2 factorial comparison design: **data integrity** (clean vs. contaminated) $\times$ **summary method** (standard vs. robust).

Contamination procedure. In each participant-by-condition run of trials, we contaminated 5% of trials using a two-stage procedure. Following Vandekerckhove, Tuerlinckx, and Lee (2008), we generated a binary indicator for each of these trials to determine the type of contamination. With 50% probability, the trial was a *delayed startup*, for which we added uniform noise between 2s and 3s to the observed RT. For the other contaminant trials, we generated a *guess* trial by setting the drift rate to 0 and redrawing the observation.

Simulation study

We conducted an extensive simulation study in which we generated 1,000 datasets for each cell defined by a $4 \times 5 \times 4$ factorial design. This design included 4 levels for the number of participants $P \in \{20, 40, 80, 160\}$, 5 levels for the number of trials per condition $T \in \{20, 40, 80, 160, 320\}$, and 4 levels for the true fixed effect size $\beta \in \{0.0, 0.1, 0.2, 0.4\}$. Across all datasets, the true values for all variance parameters were fixed at $\sigma_\nu^2 = 0.75, \sigma_\alpha^2 = 0.5, \sigma_\tau^2 = 0.1$. Each dataset was generated from its own set of true values for μ_ν, μ_α, and μ_τ sampled from the following uniform distributions:

$$\mu_\nu \sim \mathcal{U}(-3, 3) \qquad \mu_\alpha \sim \mathcal{U}(2, 4) \qquad \mu_\tau \sim \mathcal{U}(0.2, 0.4)$$

These hierarchical parameters were then used to sample individual-level boundary separation (α_p) and nondecision time (τ_p) parameters and participant-by-condition specific drift rates $\nu_{p,k}$ according to the regression structure described in Equation 2.

The results of our simulation study focus on the effect size parameter β. First, we evaluate the diagnostic accuracy performance of the robust and standard EZ-DDM implementations under clean and contaminated data, by testing the null hypothesis $\beta = 0$ in a Bayesian framework. Then, we evaluate the parameter recovery of the effect size parameter β across all simulation conditions.

Diagnostic accuracy results. The hypothesis testing performance of both hierarchical Bayesian implementations was evaluated using their Receiver Operating Characteristic (ROC) curves. To draw an ROC curve, we plot the detection rate (proportion of $\beta \neq 0$ simulations in which we correctly detect an effect) on the vertical axis and the false alarm rate (the proportion of $\beta = 0$ simulations in which we incorrectly claim there is an effect) on the horizontal axis. The curve is then drawn by varying the decision threshold from maximally conservative (where both are 0) to maximally liberal (where both are 1). The Area Under the Curve (AUC) metric then quantifies diagnostic accuracy, with values closer to 1 indicating stronger discrimination between true effects ($\beta \neq 0$) and null effects ($\beta = 0$).

Figure 1 summarizes the results. Within each panel, the horizontal axis displays the true effect sizes (β) and the vertical axis shows the AUC computed for each simulation condition. By definition, the AUC is 0.5 where $\beta = 0$. Different lines contrast the standard and robust EZ-DDM implementations under clean and contaminated data. Panels are arranged in a grid showing the different trial sizes per condition (T) across columns and the sample sizes (P) across rows.

The result of interest is captured by the fact that the red dashed line is dominated by all other lines, which overlap. That is, the robust and standard EZ-DDM produced nearly identical AUC results under clean data, but the standard EZ-DDM exhibits noticeable reductions in AUC under contaminated data.

Estimation performance results. We also explore the parameter recovery performance of the robust and standard implementations of the EZ-DDM within a hierarchical Bayesian framework by computing the root mean square error (RMSE) of the estimated effect size parameter β for each simulation condition. Our focus on β is motivated by the fact that it is the central parameter of interest in hypothesis testing and the main determinant of the AUC diagnostic accuracy reported in Figure 1. The RMSE summarizes overall estimation error by combining both bias and variability, with values closer to 0 indicating more accurate recovery of the true parameter.

Figure 2 summarizes the results. Within each panel, the RMSE of the estimated effect size parameter β is plotted as a function of the true β value, with separate lines for each of the four comparison conditions. The result of interest is that the RMSE obtained for the robust implementation across both data conditions (clean and contaminated) is comparable to that of the standard implementation under clean data, with similar RMSE values reported across all true effect sizes β. However, the RMSE of the standard implementation under contaminated data exhibits noticeable changes as a

FINAL PREPRINT

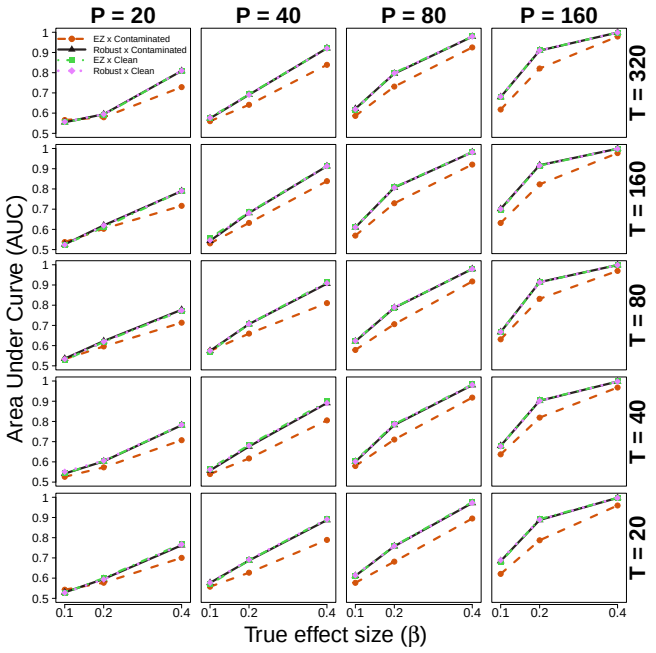


Fig. 1. Simulation study results (diagnostic accuracy). Each panel shows the AUC (vertical axis) for different true effect sizes (horizontal axis) and simulation comparison conditions (different lines), across different trial and participant sizes (rows and columns).

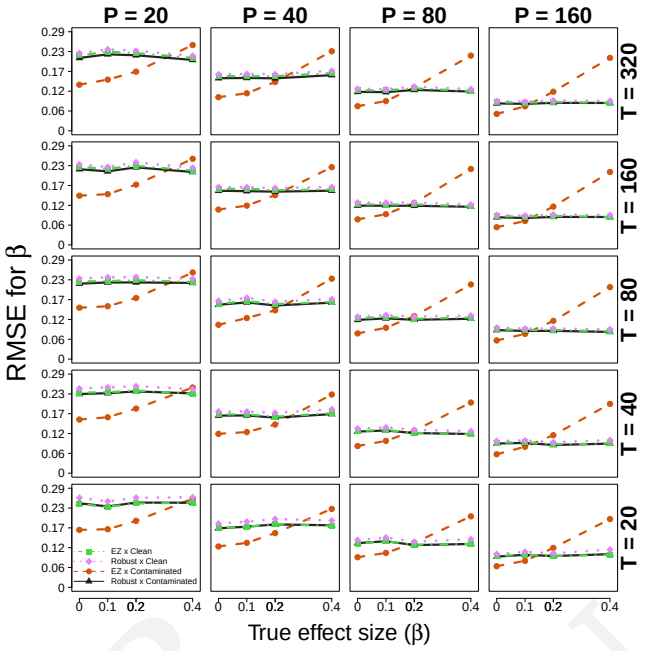


Fig. 2. Simulation study results (parameter recovery). Each panel shows the RMSE (vertical axis) of the estimated effect size parameter β (true value on the horizontal axis) for each comparison condition (different lines), across different trial and participant sizes (rows and columns).

function of the true effect size β. In particular, the RMSE is lower for smaller true effect sizes β and increases as the true effect size β increases.

To get a better understanding of the systematic differences in RMSE observed across simulation conditions, we inspect the estimation bias and variance components separately. Figure 3 shows the estimation bias for the effect size parameter β as a function of its true value, across all simulation conditions. The estimation bias refers to the difference between the true and the estimated value—in this case, the mean posterior value—as such values closer to 0 indicate a more accurate (i.e., unbiased) parameter recovery. According to Figure 3, the estimation bias for the effect size parameter β is generally close to 0 across most simulation conditions. However, under the standard implementation of the hierarchical EZ-DDM under contaminated data, the effect size parameter seems to be systematically underestimated for $\beta \neq 0$, a result that increases notably as the true effect size β increases.

To further investigate the change in performance seen in the standard implementation under contaminated data, we conducted additional simulation studies to explore the effects of specific parameter configurations (i.e., combinations of high vs. low drift rate and boundary separation parameters) on both the diagnostic accuracy and parameter recovery.

Follow-up simulations

We conducted a follow-up simulation study in which we generated 1,000 datasets with 160 participants ($P = 160$), 40 trials per condition ($T = 40$), using the same four effect size levels as in our extended simulation study ($\beta \in \{0.0, 0.1, 0.2, 0.4\}$). These simulations introduce four distinct true parameter configurations defined by combining two levels of the population intercept of the drift rate (high μ_ν vs low μ_ν) and two levels of the population mean boundary separation (high μ_α vs low μ_α). These parameter configurations were created by sampling true parameter values from different uniform distributions, as indicated below:

$$\mu_\nu^{\text{low}} \sim \mathcal{U}(0, 1) \qquad \mu_\nu^{\text{high}} \sim \mathcal{U}(2, 3)$$
$$\mu_\alpha^{\text{low}} \sim \mathcal{U}(2, 2.5) \qquad \mu_\alpha^{\text{high}} \sim \mathcal{U}(3.5, 4)$$

The population mean nondecision time (μ_τ) and all individual-level parameters were generated in the same way as in our first simulation study.

Diagnostic accuracy by parameter configuration. We evaluated the diagnostic performance of the standard and robust implementations of the EZ-DDM under clean and contaminated data across parameter configurations. Results are summarized in Figure 4, where the two levels (high vs. low) of the population intercept for the drift rate parameter μ_ν and the two levels (high vs. low) of the population mean boundary separation parameter μ_α are organized across rows and columns, respectively. Just like in Figure 1, each panel in Figure 4 shows the AUC (vertical axis) as a function of the true effect size (horizontal axis) across all simulation comparison conditions (different lines). The main result highlighted by Figure 4 is that both implementations of the EZ-DDM perform similarly across data conditions when the true population intercept for the drift rate is low (i.e., $0 \leq \mu_\nu \leq 1$, as it would be in a hard task condition). However, when the population intercept is high (i.e., $2 \leq \mu_\nu \leq 3$, as it would be in an easy task condition), the standard implementation shows a notable decrease in AUC performance under contaminated data. This result holds across low and high levels of the population boundary separation parameter ($2 \leq \mu_\alpha \leq 2.5$ and $3.5 \leq \mu_\alpha \leq 4$).

We further inspect these results by examining changes in the AUC as a function of the true population intercept for the drift rate parameter μ_ν, across high and low levels of the population boundary separation parameter μ_α, for a fixed effect size $\beta = 0.2$ (which was the value for which we observed the largest vertical distance in Figure 4). Results are summarized in Figure 5, where

AUTHOR VERSION

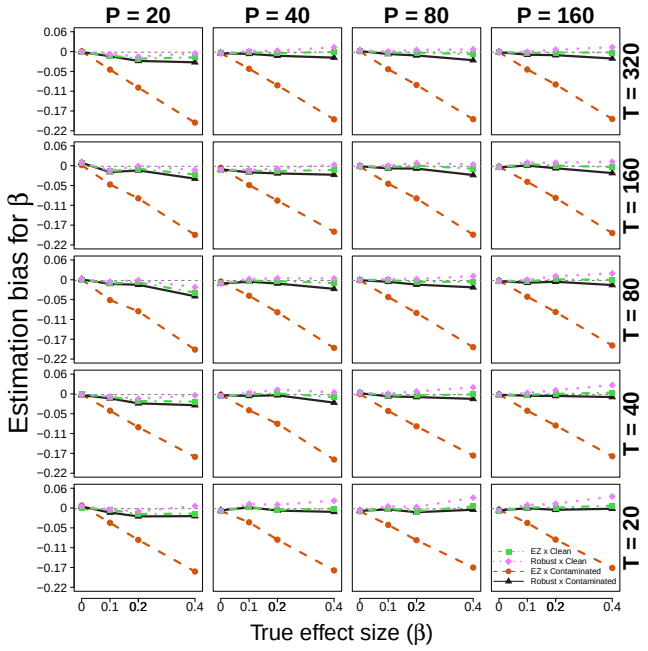


Fig. 3. Simulation study results (parameter recovery). Each panel shows the RMSE (vertical axis) of the estimated effect size parameter β (true value on the horizontal axis) for each comparison condition (different lines), across different trial and participant sizes (rows and columns).

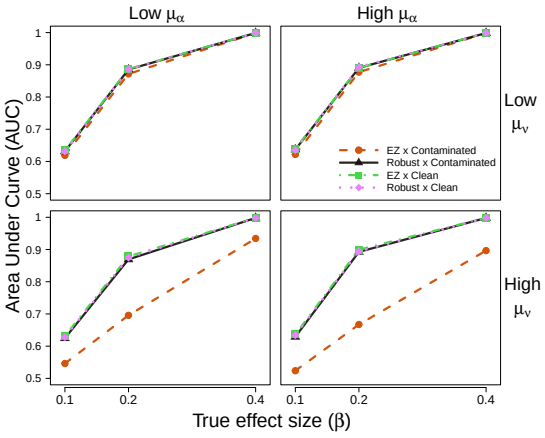


Fig. 4. Diagnostic performance of the robust and standard implementations of the EZ-DDM under clean and contaminated data across different parameter configurations. Each panel shows the AUC (vertical axis) for different true effect sizes (horizontal axis) and simulation comparison conditions (different lines). The panels on the left show the results for low levels of the population boundary separation parameter ($2 \leq \mu_\alpha \leq 2.5$) and the panels on the right show the results for high levels of the population boundary separation parameter ($3.5 \leq \mu_\alpha \leq 4$). The top panels correspond to low levels of the population intercept drift rate parameter ($0 \leq \mu_\nu \leq 1$) and the bottom panels correspond to high levels of the population intercept drift rate parameter ($2 \leq \mu_\nu \leq 3$).

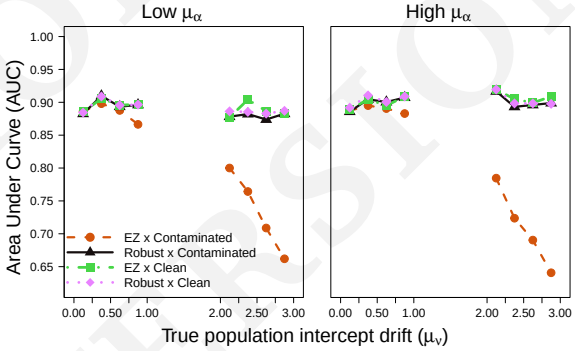


Fig. 5. AUC plotted as a function of the true population intercept for the drift rate parameter (μ_ν) in simulations with fixed effect size $\beta = 0.2$ and a within-subject t-test design on the drift rate parameter with 40 trials per condition across 160 participants. Population intercept drift rate values are aggregated into bins and shown on the horizontal axis, with the vertical axis showing the AUC. Panels distinguish between true population boundary separation (μ_α) conditions, with low boundary separation values shown on the left and high boundary separation values shown on the right. We use different lines to distinguish between the simulation comparison conditions.

the AUC is plotted as a function of the true population intercept for the drift rate parameter μ_ν (aggregated across 10 bins: 5 "low" and 5 "high" bins), with panels corresponding to low (left column) and high (right column) values of μ_α. This Figure confirms that the AUC for the standard implementation of the EZ-DDM under contaminated data decreases as the true population intercept for the drift rate parameter μ_ν increases, across both levels of the population boundary separation parameter μ_α.

Estimation performance by parameter configuration. To further understand the differences in AUC observed for the standard EZ-DDM under contaminated data, we examined how different true values of the population intercept for the drift rate parameter μ_ν affect the estimation of both the effect size parameter β and μ_ν itself. Results are summarized in Figure 6, where the first two columns show the estimation bias for μ_ν and the last two columns show the estimation bias for β. The first and third columns correspond to low values of the boundary separation parameter μ_α whereas the second and fourth columns correspond to high values of μ_α. Each row corresponds to a different true effect size β level. Within each panel, the vertical axis shows the estimation bias (computed as the mean difference between the true and the estimated values based on posterior means), and the horizontal axis aggregates values of the true population intercept for the drift rate parameter μ_ν into 10 bins.

Figure 6 shows that under contaminated data, the population intercept for the drift rate parameter μ_ν is increasingly underestimated as its true value increases across both implementations of the EZ-DDM, with the robust implementation showing less severe underestimation. As expected, the underestimation appears to be independent of the effect size parameter β. However, the underestimation is more pronounced when the population boundary separation parameter μ_α is high. On the other hand, Figure 6 also shows that, under contaminated data, the standard implementation of the EZ-DDM underestimates the effect size parameter β towards 0, with the magnitude of underestimation increasing as its true value increases and near-zero bias for $\beta = 0$.

Taken together, these results provide an intuitive account of the AUC differences observed in Figure 4. The AUC reflects how well the model can discriminate between null and true effects, which, given our simulation study design, corresponds to detecting differences in drift rate across conditions within participants. When the population intercept for the drift rate parameter μ_ν is is high, contamination induces a systematic underestimation of both μ_ν and β in the standard implementation of the EZ-DDM, attenuating the estimated within-subject difference in drift rate between conditions. As a result, the estimated effect sizes $\beta > 0$ are pulled closer to zero, reducing their separation from the null condition ($\beta = 0$). This attenuation makes it more difficult to distinguish datasets generated under the null from those generated under true effects, increasing the likelihood of Type II errors (i.e., failing to detect true effects when they are present) and bringing AUC val-

AUTHOR VERSION

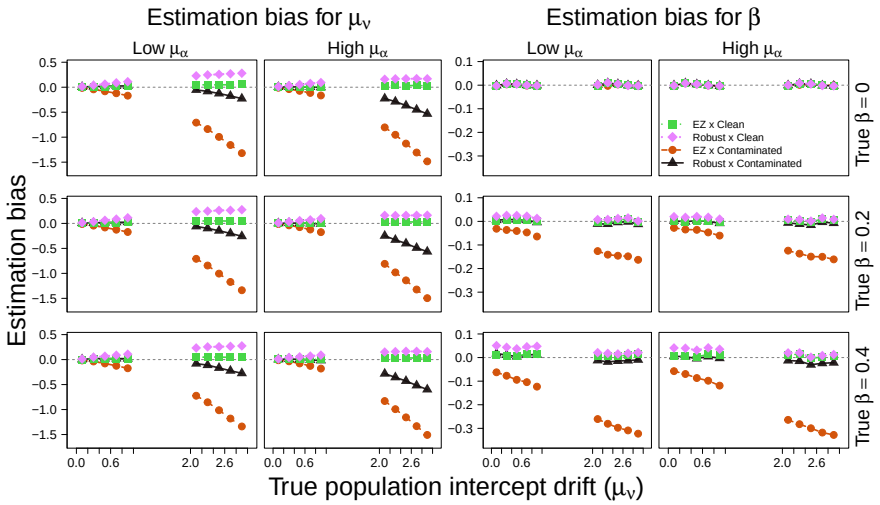


Fig. 6. Estimation bias for the population intercept drift rate μ_ν (first two columns) and the effect size parameter β (last two columns) as a function of the true population intercept drift rate μ_ν (horizontal axis) across high and low levels of the population boundary separation parameter μ_α (first and third columns vs second and fourth columns, respectively), and different levels of the true effect size β (different rows). All simulation comparison conditions are shown in different lines.

ues closer to 0.5. In contrast, the robust implementation mitigates this attenuation by reducing the bias in both parameters, which preserves the within-subject differences in drift rate induced by β, allowing the robust implementation to maintain its diagnostic accuracy under contamination.

Discussion and conclusion

The motivation for this study was to address concerns about the sensitivity of the EZ-DDM framework to contaminant data, which limits its applicability to real-world data sets. To this end, we proposed a robust implementation of the EZ-DDM in which the mean and variance of the observed RTs are replaced by the median and a robust variance estimator. The performance of this robust variant was evaluated in an elaborate simulation study using a factorial design that crossed data integrity (cleaned vs. contaminated) and summary method (mean/variance vs. quantile-based), across varying levels of sample size, trial count, and true effect size.

The results from the simulation study show that the robust version performed indistinguishably from the standard EZ-DDM when applied to clean data. Crucially, the robust implementation remained consistent under contamination conditions that otherwise substantially hindered the performance of the EZ-DDM. This general pattern was observed both in terms of the diagnostic accuracy (AUC) and parameter recovery performance (RMSE and estimation bias of the estimated effect size parameter β). Overall, these results suggest that the robust replacement for the standard EZ-DDM maintains its efficiency even under data contamination.

We further investigated the differences in performance found between both implementations across data integrity conditions through a series of follow-up simulations in which we controlled for the values of the population drift rate and boundary separation parameters. This inspection revealed that the diagnostic accuracy of the standard implementation under contaminated data decayed as a function of the true population intercept drift rate across both high and low levels of the population boundary separation parameter. This pattern could also be found in the parameter recovery performance of both the effect size parameter and the population intercept drift rate parameter, which were increasingly underestimated as the true population intercept drift rate increased (with the effect size parameter showing larger biases for larger true effect size values). Altogether, these results suggest that the standard implementation of the EZ-DDM is most affected by the presence of contaminant data when the drift rate is high, such as when the task difficulty is low, and where we expect to find lower RTs and thus, the presence of outliers has a more pronounced impact on the standard mean and variance summary statistics.

To conclude, the robust implementation of EZ-DDM provides a simple and efficient alternative to the standard EZ-DDM that combines computational tractability with resilience to contamination. Under all conditions, the robust version performed as well or better than the original implementation of EZ-DDM. We recommend researchers use the robust EZ-DDM in all cases.

References

Chávez De la Peña, A. F., & Vandekerckhove, J. (in press). An EZ Bayesian hierarchical drift diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: Wiley-Blackwell.

Lee, M. D., Criss, A., Devezer, B., Donkin, C., Etz, A., Leite, F., ... Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*, 141–153.

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

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision-making. In (pp. 1429–1434).

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