ML-PEST CTDS

Simulation-based validation report

Frozen website version 42 | 22 October 2026

Conclusion: The implementation performs well when the simulated response model matches its assumptions. The expanded study also identifies large errors that can pass the current review-readiness rule. These results support further development and independent validation; they do not establish clinical usability, MTAT equivalence or treatment-dose safety.

Study at a glance

Evaluation	Completed
Independent simulated assessment histories	127,000 across 127 scenarios; 1,000 replicates per scenario
Primary robustness grid	108,000 histories: 6 true MTs x 3 response spreads x 3 lapse mixtures x 2 curve families
Paired assessment plans	20 and 30 pulses; 254,000 endpoint records; 3,741,759 actual simulated pulses
Frozen production estimator	Default 37% MSO start; fixed 7% model spread; exact production policies
Verification	9/9 simulation checks passed; 37/37 retained isolated scalar reference cases matched

Key 30-pulse results

Cohort	Mean absolute error (MSO points)	Within +/-5 MSO points	Ready with error >10 MSO points
Matched Gaussian model (6,000 histories)	0.79	99.92%	0 / 6,000
Equal-weight robustness grid (108,000 histories)	1.28	96.28%	144 / 108,000

Decision: Retain evaluation-only status. Prefer the 30-pulse plan for further testing, but do not treat a green result as confirmation of accuracy. The largest review-ready error in this grid was an estimate of 49.67% MSO for a simulated true MT of 90% MSO: an underestimation of 40.33 MSO points.

The +/-5 MSO-point and +/-5% of true-MT margins are descriptive study endpoints, not clinically approved acceptance thresholds.

02 / Design and frozen implementation

Algorithm tested

The exact production JavaScript estimator was copied without modification. Its response model is P(positive | x, t) = Phi((x - t)/(0.07t)). Here x is delivered intensity, t is candidate MT, and Phi is the standard normal cumulative distribution. Both intensities are expressed as percent maximum stimulator output (MSO).

The likelihood includes computational 0% No / 100% Yes anchors. They were not delivered or counted as pulses. The search uses a 1-point coarse grid, a 1e-7 plateau rule and 0.01-point local refinement for a unique coarse minimum. The next suggestion is rounded to a whole MSO percentage. No result was clamped to a ceiling.

Independent synthetic subjects

Subject probabilities were generated independently with SciPy ndtr for Gaussian curves and expit for logistic curves. Logistic curves were scaled by pi/sqrt(3) to match Gaussian latent-noise variance; they have different slopes at MT. The 50% response intensity was known exactly. Each binary response was drawn from an independently keyed SHA-256 uniform variate.

For a random-response mixture fraction l, p = l/2 + (1 - l)p_base. Thus a 5% mixture creates a 2.5% low-intensity response floor, not a 5% overall misclassification rate. The fixed estimator did not know or fit l.

Factor	Values
Core true MT	20, 35, 50, 65, 80, 90% MSO
True response spread	3%, 7%, 15% of true MT; estimator fixed at 7%
Random-response mixture	0%, 2%, 5%
Response family	Cumulative Gaussian or variance-matched logistic
Paired plans	20 versus 30 pulses with shared first 20 draws
Stress tests	Abrupt +/-10-point drift after pulse 15; forced wrong first response; 60% ceiling; alternative starts 20/50/70; true MT 5/95

The 30-pulse policy continued regardless of whether its 20-pulse counterfactual was review-ready. Each endpoint was assessed against its own fixed-budget policy. Early ceiling/boundary stops were retained as unresolved runs. Outcomes use raw estimates unless explicitly described as rounded.

The analysis plan was written before this expanded run, informed by the earlier pilot; it was not preregistered. The grid is deliberately equally weighted and does not estimate clinical population prevalence.

03 / Accuracy and the 20-versus-30 comparison

Cohort / plan	Bias	MAE	95th percentile absolute error	Within +/-5 MSO points
Matched model, 20 pulses	-0.06	0.99	2.86	99.60%
Matched model, 30 pulses	-0.03	0.79	2.24	99.92%
Robustness grid, 20 pulses	-0.35	1.58	5.25	94.57%
Robustness grid, 30 pulses	-0.27	1.28	4.28	96.28%

Bias, MAE and percentile errors are MSO percentage points, not percentages of the true MT. For example, an estimate of 55 for true MT 50 has error +5 MSO points but relative error +10%. Small average bias can hide large errors of opposite sign.

Monte Carlo uncertainty

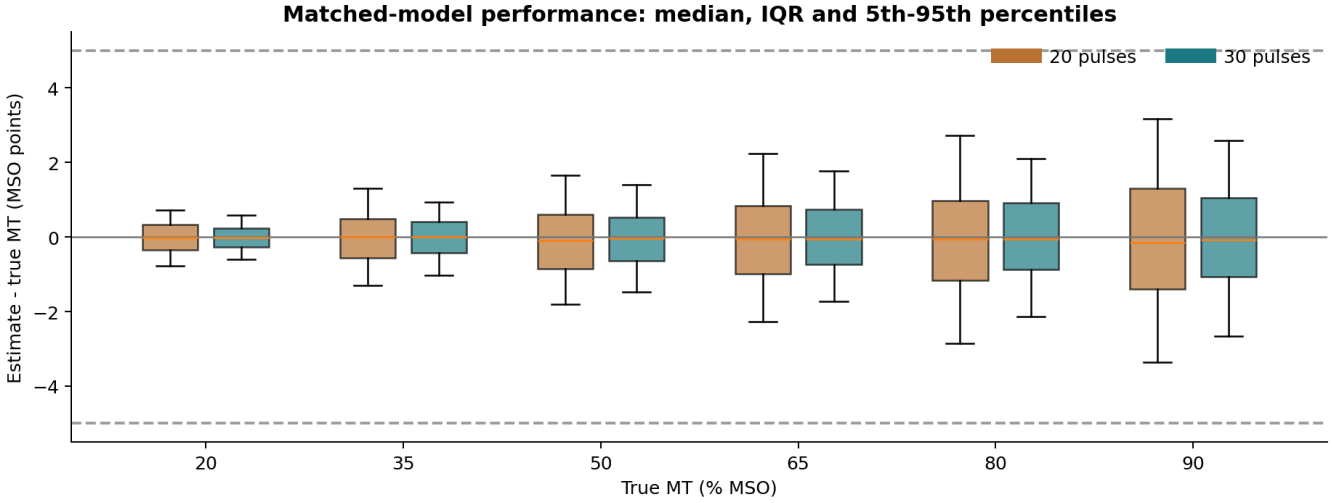
Matched model, 30 pulses: within +/-5 MSO points 99.92% (99.84% to 99.99%); MAE 0.787 (95% Monte Carlo interval 0.770 to 0.804). Within +/-5% of true MT: 99.55%.

Robustness grid, 30 pulses: within +/-5 MSO points 96.28% (96.18% to 96.39%); MAE 1.284 (95% Monte Carlo interval 1.269 to 1.299). Within +/-5% of true MT: 90.28%.

Effect of ten additional pulses

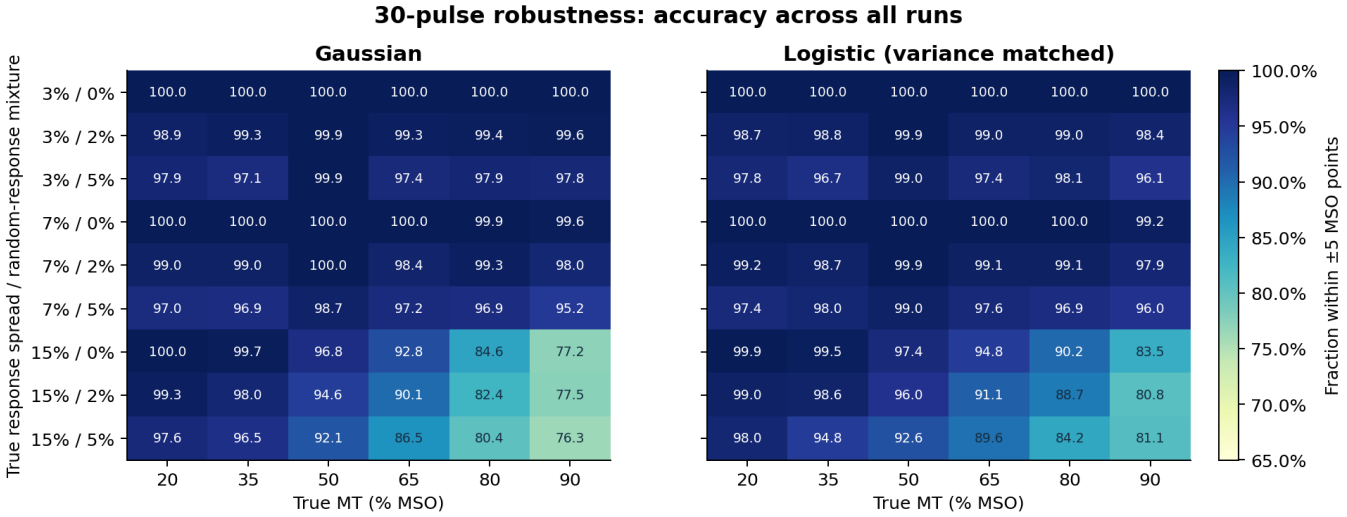
Across the robustness grid, 30 rather than 20 pulses reduced MAE by 0.301 MSO points (paired 95% Monte Carlo interval 0.296 to 0.306), a 19.0% relative reduction. Accuracy within +/-5 points increased by 1.71 percentage points (1.63 to 1.79).

Mean delivered pulses were 19.97 versus 29.95; short runs reflect blocking stops. These are pulse counts, not measured clinic time. A fixed count is not a precision-based stopping criterion.



Boxes show median and interquartile range; whiskers show 5th and 95th percentiles. Outliers are omitted only from the figure, never from analysis. These are error distributions across simulations, not patient-level MT confidence intervals.

04 / Robustness and high-error scenarios



Each cell contains 1,000 independent histories. The plot includes every core run, regardless of readiness. A response spread is the width of the simulated response curve, not uncertainty in the final estimate.

Worst 30-pulse core cells by mean absolute error

True MT / curve	Spread / mixture	MAE	Within +/-5	Ready >10 (of 1,000)
90 / Gaussian	15% / 5%	4.33	76.3%	28
90 / Logistic	15% / 5%	3.81	81.1%	16
90 / Gaussian	15% / 2%	3.72	77.5%	10
80 / Gaussian	15% / 5%	3.58	80.4%	7
80 / Logistic	15% / 5%	3.50	84.2%	5

The hardest cells combined high true MT with broad response curves and lapses. The fixed 7% estimator is then misspecified. The computational 100% positive anchor also has more influence near the upper output limit. These observations identify risk patterns; this study does not isolate a causal contribution from each numerical design choice.

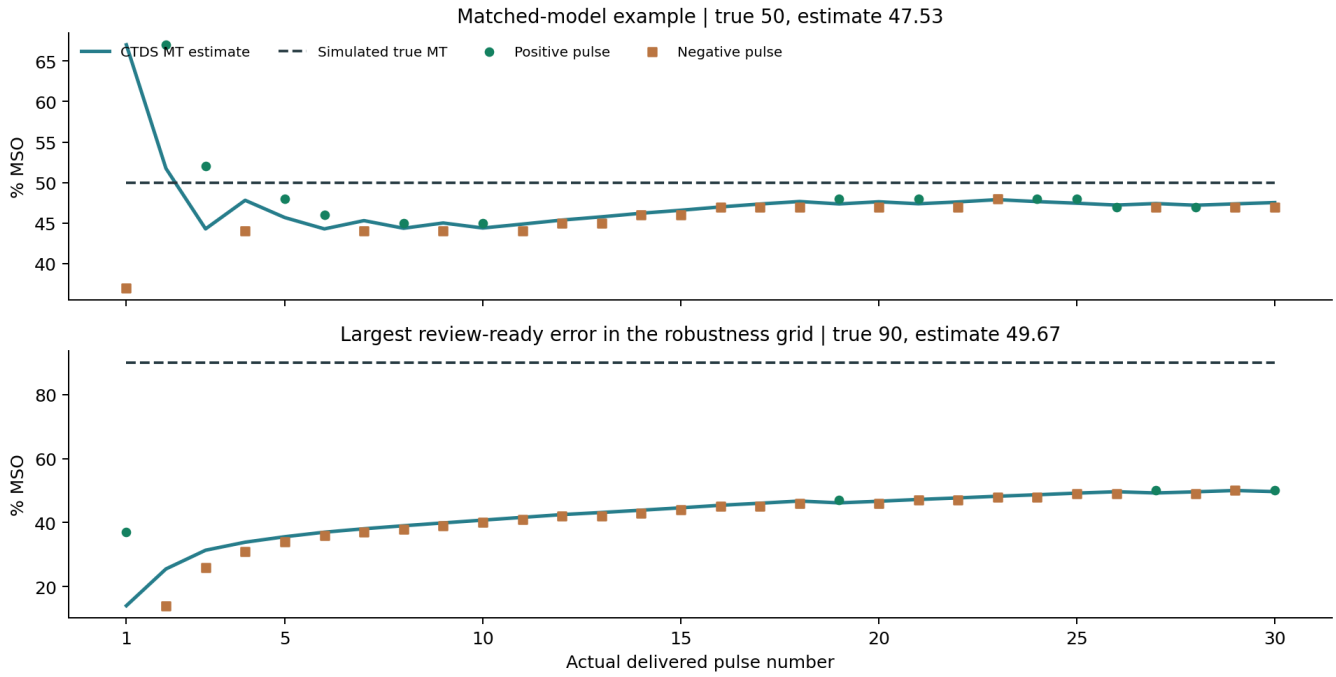
Rounded outputs

Across the core grid, mean absolute error after rounding was 1.546 MSO points at 20 pulses and 1.233 at 30, compared with raw-estimate MAE of 1.584 and 1.284. Rounding does not resolve tail failures.

05 / Review readiness is not accuracy

Robustness-grid outcome	20 pulses	30 pulses
Ready for review	94.38%	95.55%
Ready AND within +/-5 (all-run yield)	91.07%	93.47%
Within +/-5, among ready runs	96.50%	97.82%
Ready with error >5 (all-run risk)	3.30%	2.08%
Ready with error >10 (all-run risk)	0.27%	0.13%
Error >10, among ready runs	0.29%	0.14%

At 30 pulses, 2,247 of 108,000 core runs were ready with errors over 5 points; 144 exceeded 10 points. The latter all-run rate was 0.133% (0.112% to 0.155%). Among ready runs the corresponding rate was 0.140% (0.117% to 0.162%).



The failure example is the largest observed ready-run error, selected automatically rather than as a representative case. True MT 90; Gaussian spread 15%; 5% random-response mixture; replicate 411. An early positive at 37 redirected the search downward. There were two positives in the last ten pulses, so the current 0/1/9/10 imbalance rule did not block readiness. Final error: -40.33 MSO points. Full probabilities, uniforms and pulse logs are supplied.

06 / Stress tests: do not pool with stationary performance

Scenario (1,000 each)	MAE 20 / 30	Ready 20 / 30	Ready >5 30 pulses
Drift: 50 to 40 after pulse 15	8.36 / 6.36	79.2% / 0.4%	0.1%
Drift: 50 to 60 after pulse 15	8.35 / 6.03	82.5% / 10.9%	7.0%
Drift: 80 to 70 after pulse 15	7.84 / 5.34	83.7% / 30.9%	12.9%
Drift: 80 to 90 after pulse 15	7.89 / 5.31	85.8% / 50.5%	27.3%
Forced first negative; MT 20	10.46 / 8.25	0.0% / 0.0%	0.0%
Forced first positive; MT 70	21.91 / 15.22	0.0% / 0.0%	0.0%
Ceiling 60; MT 50; start 37	17.00 / 17.00	0.0% / 0.0%	0.0%
Ceiling 60; MT 70; start 37	3.00 / 3.00	0.0% / 0.0%	0.0%
Start 20; MT 20	0.36 / 0.29	96.8% / 97.9%	0.0%
Start 50; MT 20	0.38 / 0.29	97.4% / 97.8%	0.0%
Start 70; MT 20	0.39 / 0.29	97.6% / 98.3%	0.0%
Start 20; MT 50	0.89 / 0.69	96.8% / 97.2%	0.0%
Start 50; MT 50	0.88 / 0.70	97.5% / 98.2%	0.0%
Start 70; MT 50	0.90 / 0.69	97.2% / 97.4%	0.0%
Start 20; MT 90	1.72 / 1.38	96.1% / 95.9%	0.0%
Start 50; MT 90	1.79 / 1.47	95.4% / 95.5%	0.0%
Start 70; MT 90	1.52 / 1.23	97.2% / 97.7%	0.1%
Boundary case: MT 5	0.11 / 0.08	97.2% / 98.2%	0.0%
Boundary case: MT 95	2.20 / 1.91	84.0% / 84.7%	0.1%

Drift: a cumulative stationary model can lag a changing threshold. Opposite drifts nearly cancel mean bias, while leaving large absolute errors. Post-shift MT is the reference at these endpoints.

Wrong first responses: these deliberately extreme entries produced large persistent errors; the readiness gate rejected all 2,000 runs at both budgets. This demonstrates a useful safeguard, not recovery of accuracy.

Ceilings: all 2,000 capped runs stopped after one pulse when the next model suggestion exceeded 60. No pulse above the ceiling was simulated. This included MT 50, which is below the ceiling: safe stopping can leave a technically reachable threshold unresolved.

07 / Validation interpretation and recommended next steps

What this study supports

A reproducible simulation-based assessment of the frozen algorithm under the stated response models. The default 30-pulse plan improved average and tail performance relative to 20 pulses. Numerical regression matched 37 retained scalar reference fixtures, and all nine simulator checks passed. Complete endpoint data, seeds, source hashes and selected pulse-level replays are available.

What it does not support

Clinical validation, treatment dosing authorization, visual-movement equivalence, active-MT validation, MTAT full-application equivalence, equivalence to Rossini-Rothwell, or a guarantee that the reported estimate is within a particular tolerance. No patient or stimulator was involved. No real EMG amplitude, hotspot error, muscle activation, coil movement, physiological serial dependence or response-classification artifact was modeled, beyond the specified lapse mixture and forced first response.

Recommended development decisions

1. Keep the current evaluation-only restriction and independently obtained clinical-MT pathway. Do not equate green review status with a validated result.

2. Use 30 pulses as the default candidate for subsequent testing. The simulation does not justify a general 20-pulse shortcut under broad curves or lapses.

3. Improve quality assessment before promoting the estimator: investigate inconsistency and model fit, early-response sensitivity, evidence from independent confirmation pulses, and explicit drift detection. Freeze any revised version and evaluate it on new seeds and held-out scenarios; do not tune and claim success on this same dataset.

4. Perform end-to-end comparisons with the original MTAT application using actual intensities, responses, initialization, rounding, stopping and readiness records. Scalar fixture agreement alone does not establish application parity.

5. Plan a prospective clinician-supervised, paired EMG study against an independent reference assessment. Prespecify acceptable limits of agreement, repeatability, failure handling, sample size, device/muscle scope and clinical acceptance criteria with a qualified study team.

Statistical interpretation

Monte Carlo intervals quantify repeat-simulation uncertainty for this fixed scenario design, not patient uncertainty or generalizability. Pooled mean/rate intervals use independent scenario-stratified variances; conditional rates use a ratio-estimator delta method. For zero-event proportions, a Wilson interval supplies a nonzero upper limit. Per-cell rates use Wilson intervals; rates and counts are provided separately. No prespecified clinical pass/fail threshold or equivalence margin was imposed.

08 / Reproducibility, provenance and references

Exact source identity

Website version: 42
Commit: 539405967be8691ceb59cc1bd6721754f342a487
Estimator SHA-256:
46b65b52c3b09b312f558947c57e323b2184b7b5263e2f81b469bde2c6261d7b

Reproduce the study

Install Node.js and Python with the versions/dependencies recorded in requirements.txt and the run manifest. From the extracted study folder, run:

```
python design.py
node --test tests.mjs
node simulate.mjs 8
python analyze.py
node replay.mjs
python integrity.py
python make_report.py
```

The subject tables are generated with independent SciPy probability functions. Seeds depend on scenario, replicate and pulse, so worker count and completion order do not change outcomes. The exact production estimator and numerical fixture provenance are retained under source/. A checksum manifest is supplied. The numeric reference fixtures were previously obtained from isolated MTAT scalar routines with environment stubs; no complete MTAT executable, UI run or clinical result is being claimed here.

Supplied outputs

results/endpoints.csv contains all 254,000 endpoint rows. scenario-summary.csv contains every scenario and budget. summary.json includes Monte Carlo intervals and paired comparisons. all-review-ready-errors-above-10.csv retains every such endpoint. selected-traces.json and selected-pulse-traces.csv include reproducible examples. tests.txt and integrity-checks.json document verification.

Method references

1. Awiszus F. TMS and threshold hunting. Supplements to Clinical Neurophysiology. 2003;56:13-23. doi:10.1016/S1567-424X(09)70205-3. Foundational method reference; the PubMed record has no abstract and is not a complete implementation specification.

2. Mishory A, Molnar C, Koola J, et al. Maximum-likelihood strategy for determining TMS motor threshold using sequential testing [title abbreviated]. Journal of ECT. 2004;20(3):160-165. doi:10.1097/00124509-200409000-00007. Illustrative comparison in one participant; it does not validate CTDS.

3. Ah Sen CB, Fassett HJ, El-Sayes J, et al. Active and resting motor threshold are efficiently obtained with adaptive threshold hunting. PLoS One. 2017;12:e0186007. doi:10.1371/journal.pone.0186007. Empirical ML-PEST/RR comparison informing the 20-pulse scenario; different software and experiment.

4. Koponen LM, Peterchev AV. Preventing misestimation of transcranial magnetic stimulation motor threshold with MTAT 2.0. Brain Stimulation. 2022;15(5):1073-1076. doi:10.1016/j.brs.2022.07.057. PMID:35940558. Simulation and response-imbalance rationale; results for MTAT 2.0 are not automatically transferable to CTDS.