American Journal of Innovation in Science and Engineering (AJISE)

ISSN: 2158–7205 (ONLINE)

VOLUME 5 ISSUE 3 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Innovation in Science and Engineering (AJISE)

Volume 5 Issue 3, Year 2026
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajise.v5i3.8952
https://journals.e-palli.com/home/index.php/ajise

Hierarchical Shrinkage Baselines with Independent Absolute Thresholds for Context-Adaptive Evaluation of Commercial Driver Behavior

Bair Nurbaev[1*]

Article Information

Received: August 08, 2026

Accepted: October 25, 2026

Published: November 04, 2026

Keywords

Absolute Safety Threshold, Credibility Estimation, Driver Behavior Evaluation, Feedback Cadence, Hierarchical Shrinkage

ABSTRACT

A driver's history serves as a reference for flagging unsafe behavior until it absorbs the behavior. Under a shrinkage profile with pseudo-count κ_0 = 30 segments, the expected gap between the profile and a stable braking-composite mean of 2.82 (population mean 2.35) falls from 0.29 at 18 accumulated segments to 0.043 at 300, and the deviation test loses the signal. To prevent this degradation, the proposed architecture couples an empirical Bayes shrinkage baseline with an invariant absolute threshold, evaluating behavioral features against both individual history and safety envelopes simultaneously. A deviation triggers an alert whenever either baseline condition is met. This dual-layer hierarchy directly informs the composite safety and efficiency indices, governs directional conflict tests, and calibrates the automated feedback cadence based on observed driver responsiveness. In one simulated twelve-week fleet of 42 drivers, 36 vehicles and 36 context classes, the adaptive cadence ended 5.2% below a static cadence in mean Safety Behavior Index under a shared adaptive reference (23.8 against 25.1), and mean time headway met the conflict definition for 11 of 42 drivers. Every value describes the behavior of the model under the stated assumptions in one simulated realization and measures no effect in a physical fleet.

INTRODUCTION

A baseline that follows the driver also follows the driver's habits. Under a shrinkage profile with pseudo-count κ_0 = 30 segments, a driver whose stable braking-composite mean of 2.82 exceeds the population mean of 2.35 sees the expected distance between the profile and that mean shrink from 0.29 at 18 accumulated segments to 0.043 at 300, an 85% loss of the quantity that a standardized-deviation test reads, and the driver's absolute risk stays where it was. The loss matters because flagged events decide who receives feedback, and feedback carries measurable weight: recorder-based coaching reduced risky-behavior event rates by 37% and 52% at two carriers (Hickman & Hanowski, 2011), and across 315 vehicles coaching added to in-cab warning lights reached an adjusted odds ratio of 0.52 for risky driving against a control group (Bell *et al.*, 2017). A driver whose unsafe pattern the reference has absorbed stops receiving that intervention.

Event recorders judge each segment against thresholds on accelerometer readings that were set from vendor guidance (Bell *et al.*, 2017), so a driver's habitual level and a deterioration of that driver produce the same flag. Credibility and empirical Bayes estimators blend a unit mean with the population mean at weight $Z = n/(n + \varkappa)$, where $\varkappa$ is the ratio of expected within-unit variance to the variance of unit means (Bühlmann & Straub, 1970); road-safety practice applies the same blend to site crash counts to remove regression to the mean (Hauer *et al.*, 2002). The blend targets the unit mean under squared-error loss and offers no rule for judging a new observation once the weight has moved toward the unit. Insurance telematics condenses a policy period into a few covariates per policyholder, for example K-means classes and principal components of speed-acceleration heatmaps (Gao *et al.*, 2019), and delivers a risk score after the period, at a granularity too coarse to flag one segment. By evaluating each driving segment against a context-conditioned shrinkage profile alongside an independent absolute boundary, the framework decouples longitudinal habituation from critical risk detection. Triggering warnings via either threshold ensures that the marginal detection probability for severe infractions remains invariant across accumulated mileage, whereas an unanchored shrinkage profile asymptotically attenuates detection sensitivity inversely with sample size. Safety and Efficiency Behavior Indices, a directional conflict test and a response-driven feedback cadence sit on the same hierarchy. The evidence is analytical for one worked case and simulated for one twelve-week fleet of 42 drivers, and every number reports the behavior of the model under the stated assumptions.

LITERATURE REVIEW

Event-recorder programs and kinematic risk markers share one assumption: a fixed event definition, written as thresholds on accelerometer or GPS signals, applies to every driver in every context, and the effect is judged on the program or the driver group. Risky-behavior event rates fell by 37% and 52% at two long-haul carriers after recorder-based coaching (Hickman & Hanowski, 2011). In 315 vehicles, adding supervisor coaching to in-

[1] Specialist in Commercial Trucking Safety and Efficiency, United States
[*] Corresponding author's e-mail: sayansayanov0709@gmail.com

cab warning lights lowered the odds of risky behavior to 0.61 of the level under warning lights alone and to 0.52 of the control level (Bell *et al.*, 2017). The rate of elevated gravitational-force events among newly licensed teenage drivers tracked their crash and near-crash rates (Simons-Morton *et al.*, 2012), which supports kinematic exceedance as a risk marker, and Stipancic *et al.* (2018) extract harsh acceleration and braking from smartphone GPS traces as surrogate measures of collision risk. Each design fixes the reference before data arrive: a driver with 3,000 observations in a context meets the same criterion as a driver with three.

Insurance telematics studies reduce a policyholder's driving record to covariates aggregated over the policy period and judge them by their power to predict claims. K-means classes, principal components and bottleneck-network activations extracted from speed-acceleration heatmaps led Gao *et al.* (2019) to recommend telematics covariates for car insurance pricing, and Wüthrich (2017) develops a covariate selection procedure for telematics car driving data. Counts of near-miss events are predicted from telematics covariates by Guillen *et al.* (2020), and Ayuso *et al.* (2019) add mileage and driver-behavior measures to claim-frequency ratemaking. Aggregation over a period averages across contexts: a driver who brakes harder on highway legs and softer on local roads receives the same mean as a driver with the reverse pattern, and a deviation that begins in week 9 of a policy year reaches the score after week 52.

The shrinkage weight comes from a theory that optimizes a different loss. The credibility factor $Z = n/(n + \kappa)$, with $\varkappa$ the ratio of expected within-unit variance to the variance of unit means (Bühlmann & Straub, 1970), yields the weight of the posterior mean in a normal-normal model, and Efron and Morris (1973) read the same estimator as an empirical Bayes rule that lowers total squared error for four or more units when the grand mean comes from the same data (factor $k - 3$ in place of $k - 2$). Small-area income estimates in Fay and Herriot (1979) apply the logic with weights set by the ratio of model variance to sampling variance, and Hauer *et al.* (2002) apply it to road sites, where the weighted blend of observed and predicted counts removes regression to the mean. In all four settings the loss averages over units, so a unit in the tail of the population pays for the aggregate gain, and the estimand is the unit mean; the departure of a new observation from that mean lies outside the loss. The posterior of the between-unit standard deviation depends on its prior when the number of units is small (Gelman, 2006), which bears on how $\varkappa$ can be set for a context class that few drivers share.

Models fitted to commercial-driver naturalistic data represent heterogeneity and time structure explicitly. Zhang *et al.* (2022) use Bayesian nonhomogeneous Poisson process models to describe jointly the intensities of several safety-critical event types and the lengths of on-duty breaks within a driving shift, which places time since shift start among the covariates of event risk. Such models return rates and fit statistics, and their gain in predictive accuracy has a counterpart in causal interpretability, a trade-off that Mannering *et al.* (2020) set out for highway-safety analysis. A responsiveness estimate that decides whether feedback continues needs the causal reading, and the control behind it has to reproduce the selection that triggered the message, the same selection that produces regression to the mean in before-after safety evaluation (Hauer *et al.*, 2002).

MATERIALS AND METHODS

A driving segment closes on an ignition-off event or on a stop beyond a configured duration, and receives a context class from four dimensions: road type (limited-access highway, arterial road, local road), traffic density (low or high, from the ratio of segment mean speed to the posted limit), cargo load state (loaded or empty) and time of day (daytime, dusk or peak period, night). The product of levels gives 36 classes. Twelve behavioral features describe a segment: acceleration aggressiveness, braking smoothness, cornering speed margin, steering smoothness, three headway features (mean time headway, headway variability, short-headway exposure ratio), idle time ratio, cruise control utilization ratio, engine speed efficient-band adherence, and two fuel features (driver-relative efficiency deviation and fleet-relative fuel efficiency gap). Headway is split into three features because a driver who follows at a variable distance and a driver who follows consistently close call for different corrections. Each feature loads on the Safety Behavior Index, the Efficiency Behavior Index or both, with default weights that sum to one within each index.

For each driver d, context class c and feature f, the engine stores the segment count $n(d,c,f)$, the driver's empirical mean $\bar{x}(d,c,f)$ and variance $sd^2(d,c,f)$, and the population prior mean $\mu_0(c,f)$ and variance $\sigma0^2(c,f)$ estimated from all drivers' observations in class c. The profile mean is

$$\hat{\mu}(d,c,f) = \frac{\kappa_0\,\mu_0(c,f) + n(d,c,f)\,\bar{x}(d,c,f)}{\kappa_0 + n(d,c,f)} \qquad (1)$$

Equation (1) equals the posterior mean of a normal-normal model when $\kappa_0 = \sigma w^2/\tau^2$, with σw^2 the within-driver variance and τ^2 the variance of driver means in the class, so κ_0 counts the driver segments that carry as much information as the population prior, and the weight $n/(\kappa_0 + n)$ reaches 0.5 at $n = \kappa_0$. A driver entering a class has $n = 0$ and is judged against the population; the count restarts at zero in every new class, because behavior in one class does not predict behavior in a materially different one. The profile variance is

$$\hat{\sigma}^2(d,c,f) = \frac{\kappa_0[\sigma_0^2(c,f) + (\mu_0(c,f) - \hat{\mu}(d,c,f))^2] + n(d,c,f)[s_d^2(d,c,f) + (\bar{x}(d,c,f) - \hat{\mu}(d,c,f))^2]}{\kappa_0 + n(d,c,f)} \qquad (2)$$

Equation (2) is the variance of a pooled sample of κ_0 prior pseudo-observations and n driver observations: it adds the within-group variances to the separation of

the group means, and the separation term equals $\kappa_0 n(\bar{x} - \mu 0)^2/(\kappa_0 + n)^2$. Profile statistics are recalculated only after evaluating the current segment against pre-existing priors. This sequence is essential; updating the baseline prior to score assignment attenuates the observed deviation by a shrinkage factor proportional to the pseudo-count ratio, thereby dampening transient behavioral anomalies. Consequently, the baseline configuration assigns a uniform pseudo-count across all tracked features. Selection by variance ratio, $\kappa_0(c,f) = \sigma w^2/\tau^2$, needs an estimate of τ^2 that is unstable when few drivers share a class, so a weakly informative prior on τ is required (Gelman, 2006), and a grand mean estimated from the same data requires at least four drivers in the class (Efron & Morris, 1973).

A new feature value x is compared with the profile that existed before the segment:

$$z(d,c,f) = \frac{x - \hat{\mu}_{t-1}(d,c,f)}{\hat{\sigma}_{t-1}(d,c,f)} \quad (3)$$

The second reference is an absolute threshold $T_{hard}(c,f)$, fixed without reference to any driver's history. A feature value is flagged when

Flag(d,c,f) = TRUE when $|z(d,c,f)| \geq \tau_z$,or when x crosses $T_{hard}(c,f)$ in the direction of increased risk (4)

The default thresholds are 3.0 on the deceleration-and-jerk composite of braking smoothness in highway contexts, 0.15 of the estimated rollover threshold for cornering speed margin in any context, and 0.20 of car-following distance below the minimum headway for short-headway exposure on highways at high traffic density; fleets configure further values. T_{hard} is defined on the raw scale of the feature, so accumulating history leaves the probability that a given raw value crosses it unchanged. For a driver whose feature in a cell has stationary mean μ_d, the expected offset of a new value from the pre-update profile is

$$E[x - \hat{\mu}_{t-1}] = \frac{\kappa_0 (\mu_d - \mu_0)}{\kappa_0 + n} \quad (5)$$

and the probability of an absolute flag equals $P(x \geq T_{hard})$, independent of n. A flag raised by the absolute layer alone, without a significant z, goes directly to the fleet safety manager.

A target range $[L_f(c), U_f(c)]$ per class and feature defines the excess

$$r_f(x,c) = \frac{\max\{0, L_f(c) - x, x - U_f(c)\}}{s_f(c)} \quad (6)$$

with the direction of the excess, above or below the range, recorded in a separate flag. The Safety Behavior Index over reporting period T is where $D_{d,c,T}$ is the distance

$$SBI(d,T) = 100 \cdot \frac{\sum_c \sum_{f \in F_S} D_{d,c,T}\, w_f^S \min(|z(d,c,f)|, z_{max})/z_{max}}{\sum_c D_{d,c,T} \sum_{f \in F_S} w_f^S} \quad (7)$$

driven by driver d in class c during T, F_S the safety-loaded features, w_f^S the default safety weight and $z_{max} = 4$ the clipping constant. The Efficiency Behavior Index follows from the same expression with efficiency-loaded features and weights. Both indices run from 0 (no deviation in any loaded feature) to 100 (deviation at or beyond z_{max} in every loaded feature, weighted by distance). A feature missing from one segment drops out of that segment's numerator and denominator; a feature missing for a whole period has its weight redistributed proportionally among the remaining features of the index, and the index carries a degraded-computation indicator. A feature loaded on both indices satisfies the definition of a conflict feature in a class when a corrective adjustment a that moves its target-range excess toward zero changes the two indices in opposite directions: The derivatives are estimated within

$$\text{sign}\left(\frac{\partial SBI}{\partial a}\right) \cdot \text{sign}\left(\frac{\partial EBI}{\partial a}\right) = -1 \quad (8)$$

the class from the fleet's accumulated relation between a change in the feature and the changes in the other loaded features that feed the opposite index.

A responsiveness record is kept per driver and message type. After delivery of a message, the standardized deviation of the associated feature is measured over a follow-up window of 14 days, restricted to segments in the context class that triggered the message, and a delivery yields a value when the driver covers a minimum distance in that class during the window. Responsiveness is the change in $|z|$ between the pre- and post-delivery windows, evaluated against a control estimate from the same driver's segments in the same class during periods without a recent message of that type. A rolling mean of these values per driver and message type governs delivery. A rolling mean at or below -0.30 (the minimum effect size) keeps the type in automated delivery, and a rolling mean above -0.30 after $N_{min} = 3$ deliveries suppresses the type for that driver and escalates the deviation to a fleet safety manager. Responses at or below -0.60 are classed as strong, responses between -0.60 and -0.30 as moderate, and responses above -0.30 as negligible.

Simulated fleet. All values come from one realization of a simulated fleet that exercises Equations (1) to (8). The fleet has 36 commercial vehicles and 42 drivers observed for 12 weeks, which produce 18,720 driving segments and 638,400 km of travel (34.1 km per segment and 15,200 km per driver on average) across 36 context classes. The classes combine 3 road types, 2 traffic-density levels, 2 load states and 3 time-of-day levels ($3 \times 2 \times 2 \times 3 = 36$). Sensor data are incomplete in 3.7% of segments, and 2.9% of segments go to the default class defined by load state and time of day alone. The reference configuration (RC) sets $\kappa_0 = 30$ segments for every feature, $\tau_z = 2.0$, $z_{max} = 4$ and the default absolute thresholds listed above.

Input data. Segment records are synthetic. Feature values per driver, class and feature are generated inside the simulation, and the population prior of each class

is estimated from all drivers' observations in that class. Simulated drivers react to feedback through a response function that is an input of the simulation, so group differences below express detection sensitivity multiplied by that input and carry no information on the size of a real driver response.

Case W. Driver D-017 in the class highway, low traffic density, loaded, daytime, feature braking smoothness: $\mu_0 = 2.35$, $\bar{x} = 2.82$ over n = 18 segments, $\kappa_0 = 30$, adaptive standard deviation 0.31, observed segment value 3.18. Equation (1) gives a profile mean of 2.53 (2.526 unrounded) and the logged z is (3.18 - 2.53)/0.31 = 2.10 (2.11 from unrounded values). The adaptive variance 0.0961 splits into a separation term of 0.0518 and a within-component variance of 0.0443; the latter corresponds to $\sigma 0 = s_d \approx 0.21$ when the two component standard deviations coincide, an assumption used wherever a table below varies n or κ_0.

Groups. The 42 drivers form three groups of 14. Group A runs the full method: adaptive profile and adaptive cadence. Group B runs the adaptive profile with a static cadence. Group C replaces the adaptive profile by fixed class-level fleet values $\mu_0(c,f)$ and $\sigma_0(c,f)$ without individual shrinkage and runs a static cadence. SBI and EBI are computed against each group's own detection reference.

Starting values. The pool-level starting values of SBI and EBI are 31.2 and 29.4 and are logged for Group A. Declines for Groups B and C are computed against these pool-level values, which assumes equal starting levels across groups; the note to Table 3 gives the alternative implied by the 9.7% and 8.8% declines reported for Group C.

Runs. The record contains one realization; run-to-run variance is unavailable, so group differences are point values.

Metrics. SBI and EBI follow Equation (7), with values from 0 to 100. Significant deviation frequency is the number of feature-segment observations with $|z| \geq \tau_z$ per 1,000 km. Responsiveness is the change in $|z|$ defined in the feedback cadence description, taken at 14 days. Conflict prevalence is the number of drivers with at least one conflict feature among 42. Automated and escalated shares count the messages of Group A that stayed in automated delivery or went to a fleet safety manager.

Validity bounds. The z test treats feature values within a cell as approximately normal, which gives a two-sided tail share of 4.55% per feature at $\tau_z = 2.0$. The profile treats each driver's mean as stationary over the 12 weeks. Under uniform occupancy the 1,512 driver-class cells hold 12.4 segments each, so the own-history weight at the end of the period averages 0.29, and an average cell reaches $n = \kappa_0 = 30$ after about 29 weeks, beyond the simulated horizon; cells in heavily used classes, such as case W with 18 segments, sit above the average.

RESULTS AND DISCUSSION

Under case W with the driver's stable mean held at 2.82, the expected offset between the pre-update profile and the driver's level, in braking-composite units, equals $\kappa_0(\bar{x} - \mu_0)/(\kappa_0 + n)$ because Equation (1) moves the profile toward $\bar{x}$ with weight $n/(\kappa_0 + n)$; it falls from 0.29 at n = 18 to 0.235 at n = 30, 0.157 at n = 60, 0.108 at n = 100 and 0.043 at n = 300 (Table 1). The segment of 3.18 that carries z = 2.10 at n = 18 loses its adaptive flag at n = 23, where Equation (2) with equal component variances gives z = 1.99 (at n = 25 when $\hat{\sigma}$ is held at 0.31), and z settles at 1.71 as n grows without bound; the absolute layer flags the same segment at every n because 3.18 exceeds $T_{hard} = 3.0$. Against the same segment and the same $\hat{\sigma} = 0.31$, a profile at the population mean gives z = 2.68, the shrunk profile at n = 18 gives 2.10, and a profile at the driver's own mean gives 1.16: the three references differ in the profile center alone, and the adaptive layer fires for the population profile and the shrunk profile and misses the own-mean profile.

The result holds for exceedances above T_{hard}. The driver's stable mean of 2.82 lies 0.18 below T_{hard}, and a segment

Table 1: Persistence of the adaptive and absolute flags for the case W segment (x = 3.18) as the segment count n grows

n	Own-history weight	Profile mean $\hat{\mu}$	Offset $\bar{x} - \hat{\mu}$	$\hat{\sigma}$ (Eq. 2)	z at x = 3.18	Adaptive flag	Absolute flag
18	0.375	2.526	0.294	0.310	2.11	Yes	Yes
23	0.434	2.554	0.266	0.314	1.99	No	Yes
30	0.500	2.585	0.235	0.316	1.89	No	Yes
60	0.667	2.663	0.157	0.306	1.69	No	Yes
100	0.769	2.712	0.108	0.289	1.62	No	Yes
300	0.909	2.777	0.043	0.250	1.61	No	Yes
∞	1.000	2.820	0.000	0.211	1.71	No	Yes

Note: Case W, $\kappa_0 = 30$, driver mean held at 2.82, $\mu_0 = 2.35$. $\hat{\sigma}$ follows Equation (2) with equal component variances (within-component variance 0.0443). Values come from unrounded intermediate quantities; the logged z of 2.10 at n = 18 uses the two-decimal profile 2.53. Absolute flag: x = 3.18 against $T_{hard} = 3.0$.

at that mean carries z = 0.95 at n = 18 and 0.17 at n = 300 under case W: neither layer flags the driver's typical segment, which sits 0.47 above the population mean.

Under case W at n = 18, the z statistic of the 3.18 segment rises from 1.70 at κ_0 = 10 to 2.45 at κ_0 = 100 when $\hat{\sigma}$ is held at 0.31, and from 1.71 to 2.81 when $\hat{\sigma}$ follows Equation (2) with equal component variances (Table 2), because a larger $\varkappa 0$ cuts the own-history weight $n/(\kappa_0 + n)$ from 0.643 to 0.153 and leaves the profile nearer the population mean. The adaptive flag exists for $\varkappa 0$ above 22.3 at fixed $\hat{\sigma}$ and above 23.8 under Equation (2). The trade-off sits in this one parameter: a small κ_0 individualizes a profile after few segments, with the cost that an early unrepresentative stretch of driving moves it, and at κ_0 = 10 the profile gives 64% of its weight to the driver's own 18 segments and loses the flag; a large κ_0 keeps tail drivers visible and holds the profile near the population for longer. The sweep fixes $\bar{x}$ = 2.82 and the 3.18 segment, so the flip point moves with other drivers and segments, and as κ_0 grows without bound the profile reduces to the population reference (z = 2.68 at $\hat{\sigma}$ = 0.31).

Under RC over 12 weeks with 14 drivers per group, Group A ends with a mean SBI of 23.8 against 25.1 for Group B, which shares its adaptive profile and differs by a static cadence: 5.2% lower SBI, 10.0% lower EBI (21.6 against 24.0) and 12.9% fewer significant deviations (5.4 against 6.2 per 1,000 km; Table 3). Both groups evaluate against the same adaptive profile, so the contrast isolates the cadence module, in which message types with a measured response keep automated delivery and types without one are escalated. Group B against Group C (SBI 25.1 against 28.7, EBI 24.0 against 26.9, deviations 6.2 against 7.8 per 1,000 km) gives 12.5%, 10.8% and 20.5% in favor of the adaptive profile with both groups on static cadence. These two groups differ in the reference itself, so the contrast mixes behavior change with a reference effect: an adaptive profile moves toward each driver's cell mean, which lowers |z| for a driver whose mean differs from the population mean even when behavior stays unchanged. At the logged cell occupancy the average profile has covered 29% of the distance to the cell mean by the end of the period, which caps the average reduction of the individual offset at that share. Of the 4.9 index-point SBI gap between Groups A and C, 1.3 points (27%) arise between groups that share a reference and 3.6 points (73%) between groups that differ in reference. Within Group A, significant deviations fell from 8.3 to 5.4 per 1,000 km (34.9%), and SBI and EBI fell from the pool-level starts of 31.2 and 29.4 to 23.8 and 21.6 (23.7% and 26.5%); that decline includes the same movement of the profile. The record holds one realization and lacks a reference-neutral outcome such as raw feature means or exceedances of T_{hard}, so the contrasts stand as point values under RC. Under RC in Group A, the change in |z| at 14 days ranges from -0.71 for harsh braking (46 deliveries) and -0.63 for excess

Table 2: Adaptive flag for the case W segment as a function of the pseudo-count κ_0 at n = 18

κ_0	Own-history weight	Profile mean $\hat{\mu}$	z ($\hat{\sigma}$ = 0.31)	$\hat{\sigma}$ (Eq. 2)	z (Eq. 2)	Adaptive flag (Eq. 2)
10	0.643	2.652	1.70	0.308	1.71	No
15	0.545	2.606	1.85	0.315	1.82	No
30	0.375	2.526	2.11	0.310	2.11	Yes
60	0.231	2.458	2.33	0.289	2.50	Yes
100	0.153	2.422	2.45	0.270	2.81	Yes

Note: Case W, n = 18, segment value 3.18. z ($\hat{\sigma}$ = 0.31) holds the adaptive standard deviation at its logged value; z (Eq. 2) uses equal component variances. The flag appears at $\varkappa_0$ = 22.3 for fixed $\hat{\sigma}$ and at 23.8 under Equation (2).

Table 3: Simulated driver groups at the close of the twelve-week period under the reference configuration

Group	Drivers	Reference	Cadence	Final SBI	Final EBI	Deviations per 1,000 km	SBI decline	EBI decline
A	14	Adaptive profile	Adaptive	23.8	21.6	5.4	23.7%	26.5%
B	14	Adaptive profile	Static	25.1	24.0	6.2	19.6%	18.4%
C	14	Fixed class-level fleet values	Static	28.7	26.9	7.8	8.0%	8.5%

Note: Declines are computed against the pool-level starting values 31.2 (SBI) and 29.4 (EBI), logged for Group A. If Group C started at 31.78 (SBI) and 29.50 (EBI), its declines equal 9.7% and 8.8%. Each group evaluates against its own detection reference.

cornering speed (24) to -0.54 for excess idle (39), -0.42 for cruise control underuse (31) and -0.09 for headway variability (27; Table 4). Responses at or below -0.60 are strong, responses between -0.60 and -0.30 are moderate, and a type whose rolling mean stays above -0.30 after three deliveries is suppressed and escalated to a fleet safety manager, which applies to headway variability. The mean reduction following messages classified as effective is 0.58, which matches the delivery-weighted mean of the four types with a response beyond the minimum effect size. Of the 167 messages, 116 (69.5%) stayed in automated delivery and 51 (30.5%) went to a fleet safety manager; the record reports the allocation of escalations by type for headway variability alone. Each value is a change between pre- and post-delivery windows relative to a control estimate from message-free periods in the same class, and a delivery followed by too little distance in the matched class yields no value. The estimate carries a selection structure: a message fires on a segment with $|z| \geq 2.0$, the expected $|z|$ of a normally distributed feature is 0.80, and a post-delivery window drifts toward that value whatever the driver does; the control estimate comes from message-free periods selected independently of any flag. Under RC, 37 of 42 drivers (88.1%) recorded at least one significant deviation, and 11 of 42 (26.2%) were associated with a conflict feature. The conflict

Table 4: Simulated coaching responsiveness by message type, Group A

Message type	Deliveries	Δ|z| at 14 days	Classification	Cadence applied
Harsh braking	46	-0.71	Strong response	Automated delivery continues
Excess idle	39	-0.54	Moderate response	Once per 7 days
Cruise control underuse	31	-0.42	Moderate response	Once per 10 days
Headway variability	27	-0.09	Negligible response	Escalated after 3 deliveries
Excess cornering speed	24	-0.63	Strong response	On each new significant deviation

Note. Group A, 14-day follow-up window, context-matched segments. Strong response: ⊿|z| ≤ -0.60; moderate: between -0.60 and -0.30; negligible: above -0.30. Minimum effect size 0.30; $N_{min} = 3$.

feature is mean time headway in the highway, low-density class: an increase in mean time headway reduced the safety-relevant component of the feature's deviation by 50.0% on average and raised the efficiency-relevant component by 63.4%, because wider headways coincided with throttle modulation that rose from 4.2 to 5.8 events per 10 km (38.1%) and raised fuel consumption per ton-kilometer. Equation (8) held for mean time headway; headway variability and short-headway exposure ratio did not show the pattern in the same data, which is the reason the three headway features are tracked separately. The designation was observed in this class, and the record reports no conflict designation for other classes. Throttle modulation is not among the twelve features, so the coupling that produces the opposite signs enters the test through derivatives estimated from fleet history. The record contains no comparison method for conflict designation.

Discussion

Contrary to the prescription of Fay and Herriot (1979) and Bühlmann and Straub (1970), in which the variance ratio fixes the shrinkage weight because it minimizes prediction error, the results for case W indicate that detection requires its own lower bound on κ_0. The quantity that squared-error minimization removes from a unit estimate is $\varkappa 0/(\varkappa 0 + n)$ times the gap between the unit and the population, and that quantity carries the signal of the deviation test: 0.29 at n = 18 in case W, which holds z above 2.0 for κ_0 above 22.3 at $\hat{\sigma} = 0.31$ (23.8 under Equation 2). Any variance ratio below that bound removes the flag from the 3.18 segment. Dominance of shrinkage in squared error is a total over units (Efron & Morris, 1973), and detection is judged one unit at a time in the tail, where the shrunk estimate carries its largest bias. A κ_0 above the variance ratio costs accuracy in the unit mean and buys visibility of tail drivers; the detection requirement and the variance ratio are two constraints on one parameter, and the larger one governs.

The pooled variance specification introduces a structural vulnerability: the between-mean separation component constitutes over half of the total variance in sparse regimes, peaking when accumulated segments match the pseudo-count. While this penalty prevents premature confidence when profiling outlier drivers with limited mileage, it deflates standardized scores, dropping the test statistic from 3.11 to 2.11 in the highlighted trial. The resulting inflation in tolerance paradoxically shields habitual tail offenders from targeted supervisory alerts. The predictive variance of a normal-normal model, $\sigma w^2(1 + 1/(\kappa_0 + n))$, carries no such term, and the choice between the two forms is a design decision with a quantified cost of 1.0 z unit in case W.

Because individual risk indices rely on group-specific moving references, recorded performance gains conflate genuine behavioral correction with the mechanical

convergence of the profile toward an individual driver's historical average. Consequently, the observed 23.7% reduction in index scores within the adaptive cohort reflects this dual etiology. Isolating the coaching effect requires contrasting cohorts evaluated against identical references; under this constraint, automated cadence accounted for a modest 1.3 point score reduction, whereas the remaining divergence stemmed primarily from baseline adjustment dynamics. The absolute value in $|z|$ also scores a braking segment smoother than the profile as a deviation; the target-range excess of Equation (6) separates direction and counts movement toward risk alone. Any run that reports a safety effect has to log reference-neutral outcomes: raw feature means by class and the exceedance rate of T_{hard}.

The responsiveness estimator carries the selection structure that produces regression to the mean in before-after safety evaluation (Hauer *et al.*, 2002). A message fires on a segment with $|z| \geq 2.0$, the expected $|z|$ of a normal feature is 0.80, and the post-delivery window drifts toward that value whatever the response; the control estimate from message-free periods is selected independently of any flag and carries no such drift. The cadence rule generates a better control by itself: flagged deviations that occur in intervals where a message is suppressed share the selection of delivered messages and lack the delivery. Comparing the change in $|z|$ between delivered and suppressed flagged deviations separates the response from the drift. Shrinking a driver's responsiveness toward the message-type mean with the hierarchy of Equation (1) stabilizes the decision after $N_{min} = 3$ deliveries, where a rolling mean of three values needs a variance estimate to separate -0.09 from -0.42. The causal reading that Mannering *et al.* (2020) set against predictive accuracy depends on this control.

Calibration of the z layer needs the count of features evaluated per segment. Assuming independent normal distributions across twelve features, a standard significance threshold would theoretically trigger false positives at a rate of 0.55 flags per trip, well above the empirical frequency of 0.18 observed in the primary cohort. This divergence arises partly because distance-conditioned features operate only during specific maneuvers, reducing the effective feature dimensionality. Furthermore, empirical deceleration and jerk distributions exhibit pronounced leptokurtosis, rendering standard normal tail probabilities an uncalibrated proxy within this simulation environment.

Case W also marks the boundary of the absolute layer. A stable mean of 2.82 under $T_{hard} = 3.0$ is absorbed by the profile and passes both layers ($z = 0.95$ at $n = 18$, 0.17 at $n = 300$). A manager-facing check on the standardized driver effect, $(\bar{x} - \mu_0)/\tau$, sees that pattern at 0.47 units above the population mean without a configured threshold, and it stays on the manager channel because it ranks drivers.

The absorption regime lies mostly outside the simulated horizon: an average cell holds 12.4 segments and reaches $n = \kappa_0$ after about 29 weeks, so the contrasts under RC test context conditioning and cadence more than they test protection against absorption, which rests on Equation (5) and case W. The measurement that closes the gap is the share of flags raised by the absolute layer alone, by accumulated-count bin, in a run longer than 29 weeks. Time since shift start belongs among the context dimensions, since commercial-driver data model event intensity over the shift (Zhang *et al.*, 2022), and the closed-form updates of Equations (1) and (2) need a benchmark: a full multilevel Bayesian fit in brms (Bürkner, 2017) provides the reference posterior, and Pareto-smoothed importance-sampling leave-one-out cross-validation compares candidate κ_0 values by predictive accuracy from one fit (Vehtari *et al.*, 2017).

CONCLUSION

A fleet can run an individual shrinkage profile and an absolute safety boundary as independent references, and the cost of individualization is set by a single parameter: for a driver with a stable braking-composite mean of 2.82 against a population mean of 2.35, the adaptive flag on a 3.18 segment exists only when the pseudo-count exceeds 22.3, while the absolute flag holds at every accumulated segment count. In one simulated twelve-week fleet of 42 drivers, the adaptive feedback cadence ended 5.2% below a static cadence in mean Safety Behavior Index under a shared adaptive reference, and mean time headway met the conflict definition for 11 of the 42 drivers. These results matter for fleet safety management because a baseline built from a driver's own history gradually absorbs a habitual unsafe pattern, and a fixed boundary keeps severe deviations visible regardless of accumulated mileage. The evidence has clear limits: it comes from a single simulated realization, measures no effect in a physical fleet, and ends before an average driver-class cell reaches the pseudo-count at about 29 weeks. Fleet operators who adopt this design should therefore set the pseudo-count above the detection lower bound, log reference-neutral outcomes such as raw feature means and exceedances of the absolute threshold, and test in a field run longer than 29 weeks how the share of flags raised by the absolute layer alone changes as cells pass the pseudo-count.

REFERENCES

Ayuso, M., Guillén, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation, 46*(3), 735–752. https://doi.org/10.1007/s11116-018-9890-7

Bell, J. L., Taylor, M. A., Chen, G.-X., Kirk, R. D., & Leatherman, E. R. (2017). Evaluation of an in-vehicle monitoring system (IVMS) to reduce risky driving behaviors in commercial drivers: Comparison of in-cab warning lights and supervisory coaching with videos of driving behavior. *Journal of Safety Research, 60,* 125–136. https://doi.org/10.1016/j.jsr.2016.12.008

Bühlmann, H., & Straub, E. (1970). Glaubwürdigkeit für

Schadensätze [Credibility for loss ratios]. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker, 70*(1), 111–133.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors: An empirical Bayes approach. *Journal of the American Statistical Association, 68*(341), 117–130. https://doi.org/10.1080/01621459.1973.10481350

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association, 74*(366), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Gao, G., Meng, S., & Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal, 2019*(2), 143–162. https://doi.org/10.1080/03461238.2018.1523068

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., & Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal, 24*(1), 141–152. https://doi.org/10.1080/10920277.2019.1627406

Hauer, E., Harwood, D. W., Council, F. M., & Griffith, M. S. (2002). Estimating safety by the empirical Bayes method: A tutorial. *Transportation Research Record, 1784*(1), 126–131. https://doi.org/10.3141/1784-16

Hickman, J. S., & Hanowski, R. J. (2011). Use of a video monitoring approach to reduce at-risk driving behaviors in commercial vehicle operations. *Transportation Research Part F: Traffic Psychology and Behaviour, 14*(3), 189–198. https://doi.org/10.1016/j.trf.2010.11.010

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research, 25,* Article 100113. https://doi.org/10.1016/j.amar.2020.100113

Simons-Morton, B. G., Zhang, Z., Jackson, J. C., & Albert, P. S. (2012). Do elevated gravitational-force events while driving predict crashes and near crashes? *American Journal of Epidemiology, 175*(10), 1075–1079. https://doi.org/10.1093/aje/kwr440

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the GPS-enabled smartphones of regular drivers. *Accident Analysis & Prevention, 115,* 160–169. https://doi.org/10.1016/j.aap.2018.03.005

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal, 7*(1), 89–108. https://doi.org/10.1007/s13385-017-0149-z

Zhang, Y., Chen, M.-H., & Guo, F. (2022). Bayesian criterion-based assessments of recurrent event models with applications to commercial truck driver behavior studies. *Statistics in Medicine, 41*(23), 4607–4628. https://doi.org/10.1002/sim.9528