The Shape of Identity

Modeling Behavioral Continuity through Time, Context, and Drift.

GrayPass Labs · tools@graypass.org

GPL-TR-2026-01 · 7 May 2026

Abstract. Behavioral signals from ordinary computer use, such as typing rhythm and pointer movement, carry measurable person-specific structure, and systems that treat this structure as an identity signal often report strong laboratory accuracy. Reported accuracy, however, depends on an easily overlooked choice: how the evaluation splits data across time. We reanalyze two canonical public datasets, the CMU keystroke dynamics benchmark (51 subjects, 8 sessions) and the Balabit mouse dynamics challenge (10 users), holding the detector fixed and varying only the evaluation protocol. Hold-time features show intraclass correlations of 0.55 to 0.77 across eight sessions, while all inter-key latencies fall below 0.5; median typing duration shortens by 26% over the eight sessions, and 96% of subjects type faster in the final session than in the first. With a scaled Manhattan detector and 50 template repetitions, mean equal error rate moves from 10.0% (same-session) to 14.3% (session-mixed) to 18.6% (session-disjoint) on identical data, and error grows with session distance from enrollment, from 14.6% to 21.7%. A small mouse-dynamics cohort replicates the pattern of strong trait structure alongside protocol-sensitive verification error. We propose a distributional decomposition of behavioral identity into trait, context, and drift components, and a reporting standard for longitudinal behavioral-identity claims.

Keywords: behavioral biometrics; keystroke dynamics; mouse dynamics; evaluation protocols; template aging; dataset shift; intraclass correlation

1 Introduction

People interact with machines in ways that are measurably their own. The intervals between keystrokes, the shape and tempo of pointer trajectories, and the cadence of scrolling and clicking all differ from person to person, and this observation has supported three decades of work on behavioral biometrics: the use of interaction behavior as evidence about who is at the keyboard [1], [2], [3]. The appeal is practical. Behavioral evidence is available continuously and without dedicated hardware, so it is a natural candidate for continuous or secondary verification behind a conventional login [4].

The field's dominant metaphor, however, is borrowed from physiological biometrics: behavior as a fingerprint, a fixed pattern that a system enrolls once and matches thereafter. A fingerprint's ridges do not reorganize between Monday and Friday. Typing does. People practice, tire, switch chairs, keyboards, and tasks, and their motor behavior changes with them. If behavior is an identity signal, it is a signal whose distribution moves. The distinction matters most at evaluation time: an evaluation that trains and tests within one sitting measures how separable people are at a single moment, while a deployed verifier must recognize a person days or months after enrollment, under conditions the enrollment never sampled. These are different questions, and they have different answers on the same data.

This report quantifies how different those answers are, on two of the most widely used public datasets in the area, using deliberately simple and transparent methods. On the CMU keystroke dynamics benchmark [5] we measure, feature by feature, how much variance lies between subjects rather than within them (a trait measure), how the population's typing changes across eight sessions (a practice measure), and how the measured equal

error rate (EER) of one fixed, canonical detector moves when only the evaluation protocol changes (a protocol-bias measure). On the Balabit mouse dynamics challenge [6] we repeat the trait and protocol measurements in a second modality with a small cohort. Everything is deterministic and rerunnable from the public data; every number in this report is produced by the accompanying analysis scripts.

The results support a simple account. Some behavioral features behave like traits: keystroke hold times carry intraclass correlations between 0.55 and 0.77 even when estimated across sessions collected on different days. Others behave like states: inter-key latencies shorten substantially with practice, and the median total typing time for the fixed password drops by 26% across eight sessions, with 96% of subjects faster at the end than at the beginning. The consequence for verification is direct and large: the same detector on the same subjects yields a mean EER of 10.0% when the template and the test repetitions come from the same session, 14.3% when template repetitions are sampled across all sessions, and 18.6% when the template comes from session 1 and testing is confined to later sessions. Against individual sessions, the error climbs from 14.6% in session 2 to 21.7% in session 8. The evaluation protocol, an experimenter's choice that is often relegated to a methods footnote, moves the headline number by nearly a factor of two.

We argue that the right response is to change the object of study. Identity from behavior should be modeled as a person-specific distribution that shifts with context and drifts with time, not as a fixed template, and evaluations should be designed and reported so that trait, context, and drift contributions are visible rather than folded into a single optimistic number. We formalize this as an additive decomposition of an observed feature vector into a person effect, a context effect, a person-context interaction, a person-time drift, and residual noise, and we use it to organize both the empirical sections and a proposed reporting standard.

This report makes five contributions:

1. A quantified evaluation-protocol penalty on canonical public benchmarks: with the detector, features, template size, and subjects held fixed, moving from a same-session protocol to a session-disjoint protocol raises mean EER from 10.0% to 18.6% on the CMU benchmark, with subject-bootstrap confidence intervals throughout.
2. A trait-versus-drift decomposition map for each feature family: hold times combine high between-subject variance share with moderate drift, while inter-key latencies combine low variance share with large practice-driven drift.
3. A cross-modality analysis on the Balabit mouse cohort, showing high trait structure in movement-shape features and verification error that is far from the optimistic within-session regime, with wide intervals for a ten-user cohort.
4. A distributional framing of behavioral identity that separates trait, context, and drift components and clarifies what "identity evidence" should mean for continuous verification.
5. A compact reporting checklist for longitudinal behavioral-identity claims, derived from the failure modes that the experiments expose.

2 Related work

2.1 Keystroke dynamics and the CMU benchmark

Interest in typing rhythm as an identity signal dates to a 1980 RAND study in which seven professional typists could be distinguished from digraph timing distributions collected in two sittings four months apart [7]. Monrose and Rubin brought keystroke dynamics into the modern authentication literature, collecting profiles under structured and unstructured tasks and framing typing rhythm as a behavioral biometric that could harden password authentication [2], [8]. The area subsequently accumulated many detectors evaluated under incompatible conditions, which is the problem the CMU benchmark was built to solve: Killourhy and Maxion collected 51 subjects typing the same strong password 400 times over eight sessions and evaluated 14 anomaly detectors under one

shared protocol, reporting equal error rates between 9.6% and 10.2% for the top three detectors, among them the scaled Manhattan detector we reuse here [5]. The benchmark's protocol trains each subject's detector on that subject's first 200 repetitions and scores the remaining 200 as genuine attempts, with impostor scores drawn from other subjects' early repetitions. This was a deliberate and well-documented choice for comparability, not a claim about deployment; but the single protocol also fixes a single answer to the time question, and follow-on work has largely inherited it. Our contribution is not a better detector but a measurement of how much the protocol itself moves the number for a fixed detector on the same data.

The benchmark's fourteen detectors span the families that still organize the area: per-feature statistical distances (Manhattan and Euclidean variants, Mahalanobis), one-class learners such as support vector machines and neural networks, and nearest-neighbor methods [5]. One comparability note matters for reading our results against theirs: the benchmark's 9.6% figure for the scaled Manhattan detector rests on 200 training repetitions spanning each subject's first four sessions, while our protocols fix the template budget at 50 repetitions so that the three time-splits are comparable with one another. Nothing in our design optimizes absolute accuracy; the design isolates the protocol as the only moving part, which is why our absolute numbers sit above the benchmark's and why the differences between our protocols, not the levels, carry the finding.

Typing behavior is also known to change over time. Giot, Dorizzi, and Rosenberger showed that keystroke verification performance degrades over sessions unless the template is updated, and evaluated template-update strategies explicitly against that degradation [9]. Adaptive-template work of this kind treats drift as an engineering problem to be compensated; we treat it here as a quantity to be measured and reported, because compensation schemes inherit whatever the evaluation hides.

2.2 Mouse, touch, and mobile modalities

Pointer behavior was proposed as a biometric by Gamboa and Fred, who built a verifier from stroke-level features of web interaction [10], and by Ahmed and Traoré, who modeled mouse actions collected in participants' own environments [3]. Later work reported strong results with fine-grained angular features [11]. Jorgensen and Yu re-examined the area and argued that several reported results were inflated by evaluation artifacts, including pooling data collected on different machines and environments, so that the classifier could exploit environmental rather than behavioral differences [12]. Shen and colleagues then did for mouse dynamics what the CMU benchmark did for keystrokes: a controlled dataset and a systematic evaluation of detectors under one protocol [13]. Public datasets in this modality include the Balabit Mouse Dynamics Challenge, released on GitHub in 2016 with remote-desktop pointer logs for ten users [6], and the more recent SapiMouse collection [14].

Touch and mobile studies provide the clearest prior evidence that the time split drives the number. Frank and colleagues, in the Touchalytics study of touchscreen strokes, reported a median EER of 0% when training and testing within a session, 2% to 3% across sessions, and below 4% when the test followed enrollment by a week [15]. Sitová and colleagues found that behavioral features for smartphone users differ substantially between sitting and walking [16], and Crawford and Ahmadzadeh showed that mobile keystroke patterns measured without regard to body position are barely better than chance, while conditioning on position restores discriminability [17]. These are context effects in the vocabulary of this report. Surveys of continuous mobile authentication likewise identify behavioral variability as a central open problem [4].

2.3 Person-specific error structure

Averaged error rates conceal structured differences between people. Doddington and colleagues, analyzing the 1998 NIST speaker recognition evaluation, established that some speakers are systematically hard to verify and others easy to impersonate, and gave the phenomenon its animal taxonomy [18]. Yager and Dunstone generalized the analysis into a biometric menagerie across modalities [19]. This literature matters here for two reasons. First, it motivates our uncertainty procedure: because error concentrates in subsets of people, confidence intervals must

resample subjects, not attempts. Second, it is itself a distributional observation: the population is a mixture of person-specific distributions with different separations, so a single population EER is a coarse summary even before drift enters.

2.4 Dataset shift and evaluation standards

What the biometric literature calls template aging is, in machine-learning vocabulary, dataset shift: the joint distribution of features and labels differs between training and test [20]. Moreno-Torres and colleagues give a taxonomy in which covariate shift and concept shift are special cases [21]; enrollment-to-deployment drift in behavioral identity is covariate shift within each person's genuine class. The biometric testing standard ISO/IEC 19795-1 already requires that reports disclose the separation between enrollment and test data and cautions against collecting them in a single visit [22]. Our results can be read as an empirical argument for treating those disclosure requirements as first-class results rather than compliance details: on the CMU data, the disclosure in question changes the measured error by a factor of 1.85.

Across these threads, prior work generally reports accuracy under one protocol per study, and cross-study comparisons then confound detector, data, and protocol. The closest antecedents to our design are the within-versus-across-session contrasts in Touchalytics [15] and the evaluation critiques of Jorgensen and Yu [12]. What is new here is the systematic quantification, on the same data with the same fixed detector and template size, of how the protocol alone moves the headline number in two modalities, together with a feature-level decomposition that explains where the movement comes from and a reporting standard that follows from it.

3 Datasets and rights

3.1 CMU keystroke dynamics benchmark

The CMU benchmark [5] contains typing data from 51 subjects, each of whom typed the fixed strong password ".tie5Roanl" followed by Enter 400 times, organized as 8 sessions of 50 repetitions. Sessions were separated by at least one day. The public CSV distributed by the authors at the benchmark website contains 20,400 rows, one per repetition, each with a subject identifier, a session index from 1 to 8, a repetition index from 1 to 50, and 31 timing features in seconds: 11 hold times (H, key press to release of the same key), 10 keydown-to-keydown latencies (DD), and 10 keyup-to-keydown latencies (UD), covering the 11 keystrokes of the password including the Shift chord and the final Enter. The dataset is publicly distributed by its authors for research, with a request to cite the original paper. Known limitations are relevant to everything that follows: subjects were recruited from a university population, the task is fixed-text password entry on laboratory apparatus, and the public data provide a session index rather than calendar dates, so elapsed time between sessions is unknown beyond the one-day minimum. We therefore never state drift in units of days.

3.2 Balabit mouse dynamics challenge

The Balabit dataset [6] contains raw pointer event logs collected through a remote-desktop monitoring system, released on GitHub in 2016 for research use with a citation request in the repository. We use the training portion: 10 users (numbered 7 to 35) with 5 to 7 session files each, 65 sessions in total comprising 2,253,816 events and roughly 177 hours of wall-clock time, with a median session length of about 131 minutes. Each event row carries a server-side and a client-side timestamp in seconds, a button field, a state field (Move, Drag, Pressed, Released, and scroll events), and screen coordinates in pixels. We use the client timestamp throughout. The challenge's test files, which contain deliberately mixed legitimate and illegitimate segments for the original competition, are not used. Limitations mirror the CMU case in aggravated form: ten users is a very small identity population, the users

were IT professionals monitored during ordinary remote work rather than a controlled task, screen resolution and device are unlabeled, and session files are identified by opaque identifiers whose lexicographic order is not documented to be chronological. Both datasets are used here strictly for secondary analysis of the properties their creators released them to study; we perform no demographic analysis (no demographic attributes are distributed with either dataset) and no attempt to re-identify participants. The two datasets are never merged into pooled statistics; every number in this report belongs to exactly one dataset.

4 Methods

4.1 A distributional decomposition

We model an observed feature vector from person i, in context c, at time t as

$$x_{i,c,t} = \mu_i + \gamma_c + \delta_{i,c} + \tau_{i,t} + \varepsilon_{i,c,t}, \tag{1}$$

where μ_i is the person's stable trait component, γ_c is a context effect common to everyone in context c (device, task, posture, environment), $\delta_{i,c}$ is a person-context interaction (how this person in particular changes on a laptop keyboard), $\tau_{i,t}$ is a slowly varying person-time drift (practice, fatigue, motor change), and $\varepsilon_{i,c,t}$ is short-term residual variation. Figure 1 illustrates the three components that matter for identity: separation of the μ_i, common shifts from γ_c, and the wandering of $\mu_i + \tau_{i,t}$ across sessions. The decomposition is a modeling device rather than a claim of additivity in nature, but it clarifies what different experiments estimate. A same-session evaluation measures separation of $\mu_i + \tau_{i,t}$ at a single t with ε as the only noise, the most favorable case possible. A session-disjoint evaluation additionally charges the verifier for the drift $\tau_{i,t} - \tau_{i,t_0}$ accumulated since enrollment at t_0. Cross-context evaluation charges it for $\gamma_c + \delta_{i,c}$. In the CMU data the context is fixed by design (one task, one apparatus), so the dataset isolates trait and drift; in the Balabit data context varies uncontrolled within and across sessions and is absorbed into the within-user variation. No experiment in this report estimates γ_c explicitly; the touch and mobile literature reviewed above documents its importance [16], [17].

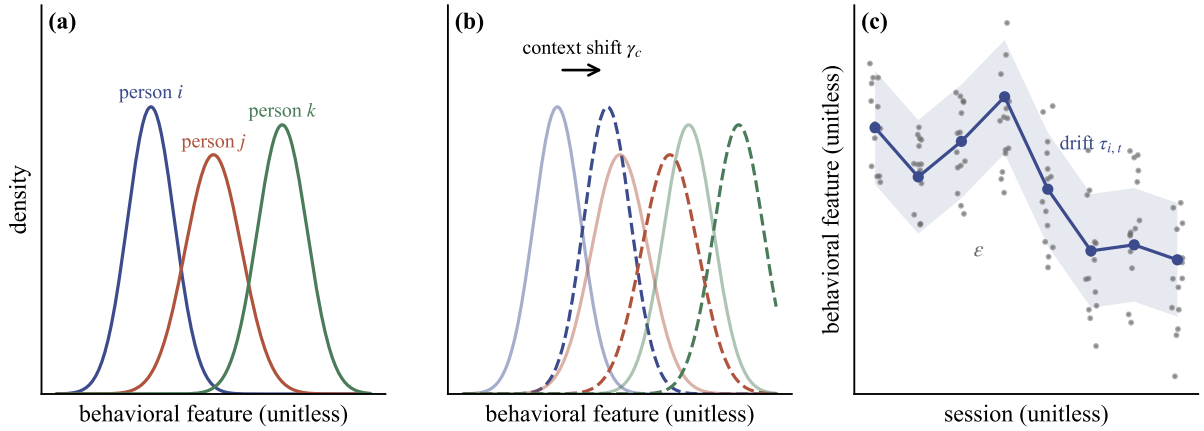

Figure 1: Conceptual schematic of the decomposition in Equation 1. (a) Trait structure: person-specific distributions of a behavioral feature, separated by μ_i. (b) A context effect γ_c shifts every person's distribution (solid to dashed). (c) Within-person drift: one person's session means wander as $\tau_{i,t}$ accumulates, while ε scatters individual observations (dots) around each session mean. Axes are unitless and the curves are synthetic illustrations, not data. Evidence class: conceptual schematic.

4.2 Trait strength: intraclass correlation

For each feature we estimate the share of variance that lies between subjects with the one-way random-effects intraclass correlation ICC(1) [23]. With k subjects, n_i observations for subject i, $N = \sum_i n_i$, and the usual between and within mean squares $\mathrm{MS_B}$ and $\mathrm{MS_W}$ from a one-way analysis of variance,

$$\mathrm{ICC}(1) = \frac{\mathrm{MS_B} - \mathrm{MS_W}}{\mathrm{MS_B} + (k_0 - 1)\,\mathrm{MS_W}}, \quad k_0 = \frac{N - \left(\sum_i n_i^2\right)/N}{k - 1}, \tag{2}$$

where k_0 is the standard correction for unequal group sizes [24]; for the balanced CMU groups k_0 equals the common group size. ICC(1) is the expected correlation between two observations drawn from the same person, and equivalently the between-person share of total variance. We compute it two ways for each CMU feature: on session 1 only (50 repetitions per subject, a within-day estimate) and on all eight sessions pooled (400 repetitions per subject). The comparison is informative about drift: if a feature drifts within person across sessions, pooling sessions inflates the within-person variance and deflates the ICC relative to the single-session estimate.

4.3 Practice and population nonstationarity

As a proxy for total typing duration of a repetition we use the sum of the 10 DD keydown-to-keydown latencies. Because consecutive DD features share endpoints, the sum telescopes to the elapsed time from the keydown of the first character to the keydown of the final Enter, so the proxy is exact for first-to-last keydown duration and ignores only the final key's hold time. For each subject and session we average this duration over the session's 50 repetitions. The population curve reported in the text is the median across subjects of these session means, normalized by its session-1 value. For the figure we normalize each subject's curve by that subject's own session-1 mean before taking the median across subjects, so that the spaghetti of individual curves and the summary line share a scale; both normalizations are reported in the results file and agree to within three percentage points at session 8. We also report the fraction of subjects whose session-8 mean duration is shorter than their session-1 mean.

4.4 Detector and verification protocols

All CMU verification experiments use one fixed detector, the scaled Manhattan distance, which was among the top performers in the original benchmark evaluation [5] and has no tuning parameters. From the template repetitions we compute, per feature j, the mean μ_j and the mean absolute deviation a_j, flooring a_j at 10^{-6}; a probe repetition $\boldsymbol{x}$ receives the score

$$s(\boldsymbol{x}) = -\sum_{j=1}^{31} \frac{|x_j - \mu_j|}{a_j}, \tag{3}$$

so that higher scores indicate closer agreement with the template. For each subject we compute an equal error rate by a full threshold sweep over the subject's genuine and impostor score sets, interpolating linearly between adjacent operating points at the crossing of false accept and false reject rates, and we report the mean over the 51 subjects. Impostor sampling is deterministic under seed 42.

Three protocols are compared. R-same-session uses 25 template repetitions, whereas the other two use 50; this is a non-isolated protocol comparison, so the contrast includes template-budget differences.

R-mixed. The template is 50 repetitions sampled uniformly at random from the subject's 400 (about six per session in expectation). Genuine scores come from the remaining 350 repetitions. Impostor scores come from 5 repetitions sampled uniformly from each of the other 50 subjects' full data, 250 per target subject. This protocol lets enrollment data span the entire study, which is the implicit assumption of any evaluation that shuffles before splitting.

R-disjoint. The template is all 50 repetitions of session 1. Genuine scores are computed separately against each later session $s = 2, ..., 8$ (50 repetitions each), with impostor scores drawn from session s of the other subjects (5 repetitions each, 250 per target and session), yielding an error curve EER(s); the headline number pools the genuine and impostor score sets across sessions 2 to 8 before the threshold sweep. This is the deployment-shaped protocol: enroll once, verify later.

R-same-session. Within each session s, the template is the 25 odd-numbered repetitions, genuine scores come from the 25 even-numbered repetitions, and impostor scores come from 5 repetitions of session s of each other subject. The per-subject EER is averaged over the eight sessions. With only 25 template repetitions this bound is, if anything, handicapped relative to the other protocols, which makes its advantage in the results a conservative estimate of the same-session optimism.

Per-subject EERs use each subject's own genuine and impostor score sets, with the acceptance threshold swept over all observed scores; no threshold is shared across subjects. This follows the benchmark's convention and matches per-user deployment of such detectors, and it is the more favorable convention, since a shared global threshold would add cross-subject calibration error on top of discrimination error. Because the convention is held fixed across all three protocols, the protocol contrasts are unaffected by it; only absolute levels would move under a shared threshold. Where the crossing of the false accept and false reject curves falls between operating points, we interpolate linearly rather than snapping to the nearer point, so small score sets do not bias the estimate toward optimistic steps.

4.5 Drift versus discriminability per feature

For the trait-state map we pair each feature's all-sessions ICC with a temporal drift index. For subject i and feature j, let $\overline{x}_{i,s}$ be the session-s mean and let σ_i^{w} be the subject's pooled within-session standard deviation, the square root of the mean of the eight within-session variances. The drift index is

$$d_j = \frac{1}{51} \sum_i \frac{1}{7} \sum_{s=2}^{8} \frac{|\overline{x}_{i,s} - \overline{x}_{i,1}|}{\sigma_i^{\mathrm{w}}}, \tag{4}$$

the average distance of later session means from the session-1 mean, in units of the subject's own short-term variability. A feature with d_j near 1 wanders, between sessions, by about as much as it scatters within a session.

4.6 Mouse features and protocols

For each Balabit session file we compute one transparent feature vector over movement segments. A segment is a maximal run of consecutive Move or Drag events, split wherever the client-timestamp gap exceeds 0.5 s, wherever the state switches between Move and Drag, and wherever any non-movement event intervenes. Within a segment, points with duplicate timestamps are dropped (the first is kept). Segments with fewer than 8 points or a path length under 100 px are discarded, as are segments containing any point-to-point speed above 5,000 px/s, which we treat as logging artifacts of the remote-desktop capture rather than human motion. Of 144,395 candidate segments, 55,637 are kept (77,874 failed the length threshold, mostly single-jitter runs between clicks; 6,856 failed the path threshold; 4,028, or 2.8%, contained a speed artifact), an average of 856 usable segments per session.

The eight session-level features are: mean and standard deviation over segments of segment mean speed (path length over duration); mean segment straightness (endpoint displacement over path length); mean absolute turn angle per interior point; pause fraction (the share of session wall time inside inter-event gaps longer than 1 s); direction-change rate (sign flips of the turning direction per second of within-segment time, pooled over segments); clicks per minute (Pressed events over wall time); and mean segment duration. Feature vectors are z-normalized per feature using the mean and standard deviation over all 65 sessions of all users; this population normalization uses no identity labels but does use the full corpus, and it is applied identically under both protocols.

The verifier is the Euclidean distance between a session's normalized vector and the mean of the user's template sessions, negated to give a score.

Two protocols mirror the CMU design at session granularity. **O-ordered**: the template is the first half (rounded down) of the user's session files in lexicographic filename order, genuine scores come from the user's remaining sessions, and impostor scores come from the other users' non-template sessions. Because the session identifiers are not documented as chronological, this is a fixed, arbitrary, deterministic split rather than a verified temporal one, and we state it as such. **O-random**: the same construction with the template sessions chosen uniformly at random, repeated 20 times under seed 42, with per-user EERs averaged over repeats. Per-user EERs use the same threshold-sweep estimator as the CMU analysis; with 3 or 4 genuine sessions per user per split, individual EERs are coarse, which the uncertainty analysis makes visible rather than hiding.

4.7 Uncertainty and reproducibility

Because verification error concentrates in subsets of people [18], [19], all confidence intervals resample participants, not attempts: we draw 1,000 bootstrap resamples of the subject set (51 subjects for CMU, 10 users for Balabit) with replacement, recompute the mean of the per-subject EERs for each resample, and report 95% percentile intervals [25]. With ten users the Balabit intervals are wide; this is the honest cost of a small identity population, and we frame that cohort strictly as a small replication. All analyses are deterministic under seed 42. A single script (`analysis/run_all.sh`) regenerates every number in this report from the raw public data into a machine-readable results file, and every figure from the same file.

5 Results

5.1 Trait structure: hold times are traits, latencies are not

Table 1 and Figure 2 summarize the ICC analysis. The three feature families behave differently and consistently. Hold times are the most trait-like: their all-sessions ICCs average 0.64 and range from 0.55 to 0.77, meaning that even with eight sessions of drift pooled into the within-person variance, more than half of the variance of a typical hold-time feature lies between people. The two latency families are substantially less person-specific: DD latencies average 0.31 (range 0.13 to 0.42) and UD latencies 0.34 (range 0.14 to 0.47) on all sessions. Under the all-sessions estimate every one of the 11 hold-time features has an ICC above 0.5 and none of the 20 latency features does. The single most trait-like feature in session 1 is the hold time of the letter t at 0.80; the weakest is the DD latency of the i-to-e digraph at 0.06.

Feature family	n	ICC(1), session 1 only mean (range)	ICC(1), all 8 sessions mean (range)
H (hold times)	11	0.69 (0.59–0.80)	0.64 (0.55–0.77)
DD (keydown–keydown)	10	0.37 (0.06–0.55)	0.31 (0.13–0.42)
UD (keyup–keydown)	10	0.37 (0.06–0.57)	0.34 (0.14–0.47)

Table 1: Between-subject variance share by feature family on the CMU keystroke benchmark (51 subjects; 50 repetitions per subject for the session-1 estimate, 400 for the all-sessions estimate). One-way random-effects ICC(1), Equation 2. Point estimates. Evidence class: reanalysis of public data.

The session-1 and all-sessions estimates also separate in an informative direction: for most features the all-sessions ICC is lower (family means drop from 0.69 to 0.64 for H and from 0.37 to 0.31 for DD), exactly what Equation 1 predicts when $\tau_{i,t}$ moves within person across sessions. A few low-ICC latencies move the other way, which is expected when between-person spread is small enough that even modest stable differences accumulate visibility with eight times the data. The overall picture is that the password-typing signal is a mixture: a genuinely

person-specific core (how long fingers rest on keys) embedded in transition timings that are only weakly person-specific to begin with.

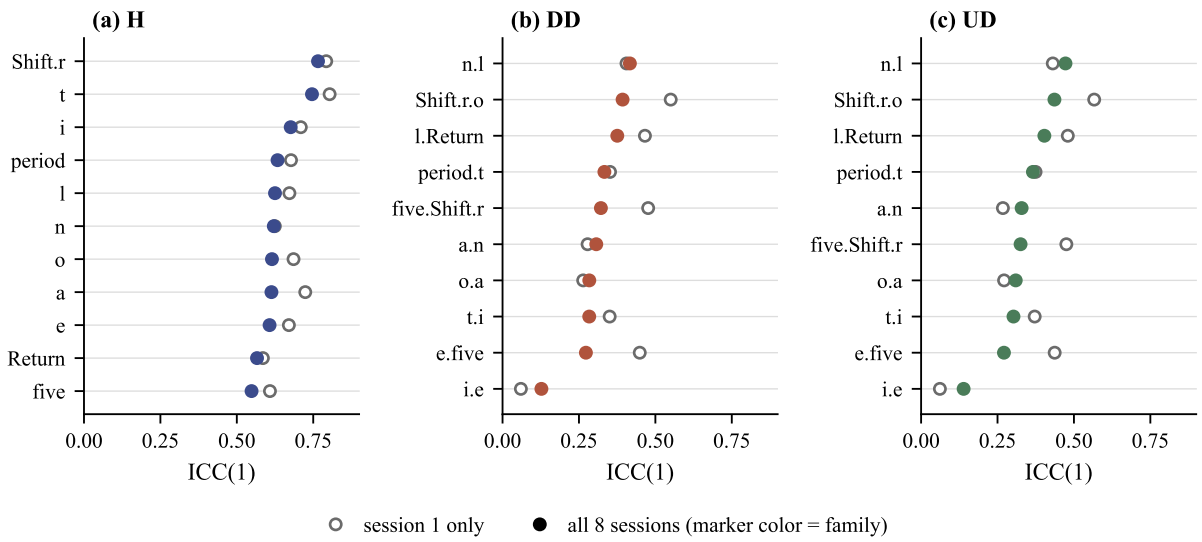

Figure 2: ICC(1) per timing feature on the CMU keystroke benchmark (51 subjects), grouped by family and sorted by the all-sessions estimate within each panel: (a) hold times, (b) keydown–keydown latencies, (c) keyup–keydown latencies. Open markers use session 1 only (50 repetitions per subject); filled markers pool all 8 sessions (400 per subject). Point estimates without intervals. Evidence class: reanalysis of public data.

5.2 The population does not sit still

Figure 3 shows the practice effect. The median across subjects of per-subject mean typing duration falls from 2.66 s in session 1 to 1.97 s in session 8, a 26% reduction; the normalized curve declines monotonically (0.88 by session 2, 0.80 by session 4, 0.74 by session 8), still visibly unsaturated at the final session. The effect is nearly universal: 49 of 51 subjects (96.1%) have a faster session-8 mean than session-1 mean. This is a population-level covariate shift in the sense of the dataset-shift literature [20], [21], produced here by practice on an initially unfamiliar password. Its size relative to the between-subject spread is what makes it consequential: a verifier enrolled on session-1 behavior is matching against a population that has since moved by a quarter of its own timescale.

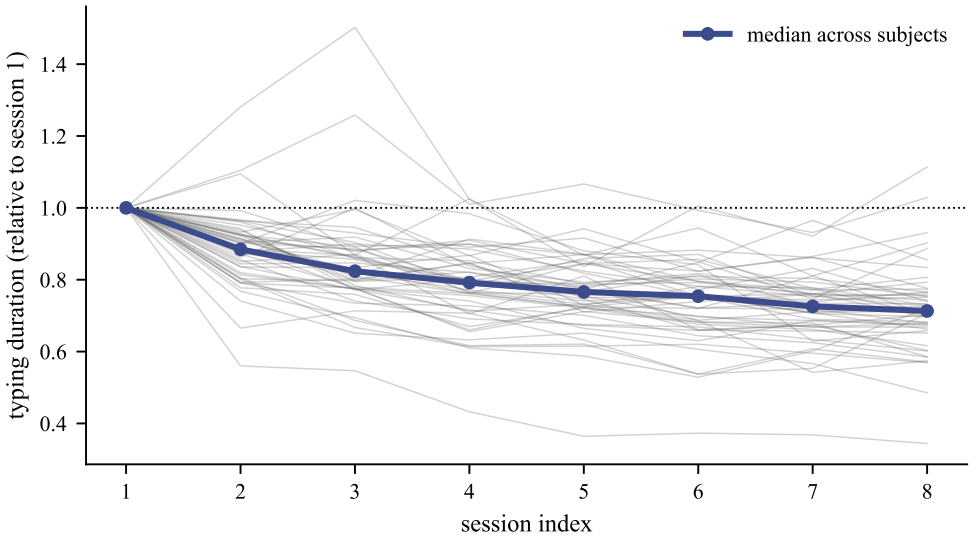

Figure 3: Practice effect on the CMU keystroke benchmark (51 subjects). Total typing duration per repetition (the telescoping sum of the 10 DD latencies, i.e., first keydown to last keydown), averaged per subject and session, then normalized to each subject's session-1 mean. Thin gray lines: individual subjects. Bold line: median across subjects of the normalized curves (reaching 0.71 at session 8; the population median normalized after aggregation reaches 0.74). No interval shown; the spaghetti displays the full population dispersion. Evidence class: reanalysis of public data.

5.3 The protocol moves the number

Table 2 is the central result. With the detector, features, subjects, and template budget fixed, the mean EER of the scaled Manhattan verifier (Equation 3) is 10.0% (95% CI 8.8 to 11.3) under R-same-session, 14.3% (12.5 to 16.2) under R-mixed, and 18.6% (15.2 to 22.0) under R-disjoint. The session-disjoint number is 1.85 times the same-session number. The bootstrap intervals of R-same-session and R-disjoint do not overlap, so the protocol penalty is resolved decisively even at 51 subjects. Note the direction of the handicaps: R-same-session achieves its advantage with only 25 template repetitions, half the budget of the protocols it beats.

Protocol	Dataset	Template	Mean EER	95% CI
R-same-session	CMU keystroke	25 odd reps of session s; genuine: 25 even reps of s; averaged over $s = 1..8$	10.0%	8.8–11.3
R-mixed	CMU keystroke	50 reps sampled across all 8 sessions; genuine: remaining 350	14.3%	12.5–16.2
R-disjoint	CMU keystroke	all 50 reps of session 1; genuine: sessions 2–8 pooled	18.6%	15.2–22.0
O-random	Balabit mouse	random half of sessions (20 seeded repeats); genuine: remaining sessions	18.3%	12.5–24.4
O-ordered	Balabit mouse	first half of session files (filename order); genuine: remaining sessions	16.7%	8.7–24.5

Table 2: Verification error under different evaluation protocols with the detector held fixed within each dataset (CMU: scaled Manhattan over 31 timing features, 51 subjects; Balabit: z-normalized Euclidean distance over 8 session-level features, 10 users). Mean of per-participant EERs; 95% percentile CIs from bootstrap over participants (1,000 resamples). The two datasets are analyzed strictly separately. Evidence class: reanalysis of public data.

Impostor scores per CMU subject: 250 (R-mixed and each R-same-session session), 1,750 (R-disjoint pooled); impostor sampling seeded (42). Balabit per-user splits leave 3–4 genuine sessions, so per-user EERs are coarse; the wide intervals reflect the 10-user cohort honestly.

Figure 4 resolves the session-disjoint error in time. Against session 2 the mean EER is 14.6% (11.9 to 17.6); it stays near that level in session 3 (14.4%), then rises through 16.1%, 17.4%, 18.1%, and 19.7% to 21.7% (17.0

to 26.5) against session 8, about 1.5 times the session-2 level. Consistent with the practice curve, degradation is not an abrupt failure but a steady climb as the population's behavior moves away from its enrollment-time state; and since sessions are only guaranteed to be at least a day apart, the horizontal axis is session order, not elapsed time. The two reference lines make the protocol comparison visual: the R-mixed line sits inside the drift curve's early range because mixing lets the template average over the very drift the disjoint protocol charges for, and the same-session line sits below everything.

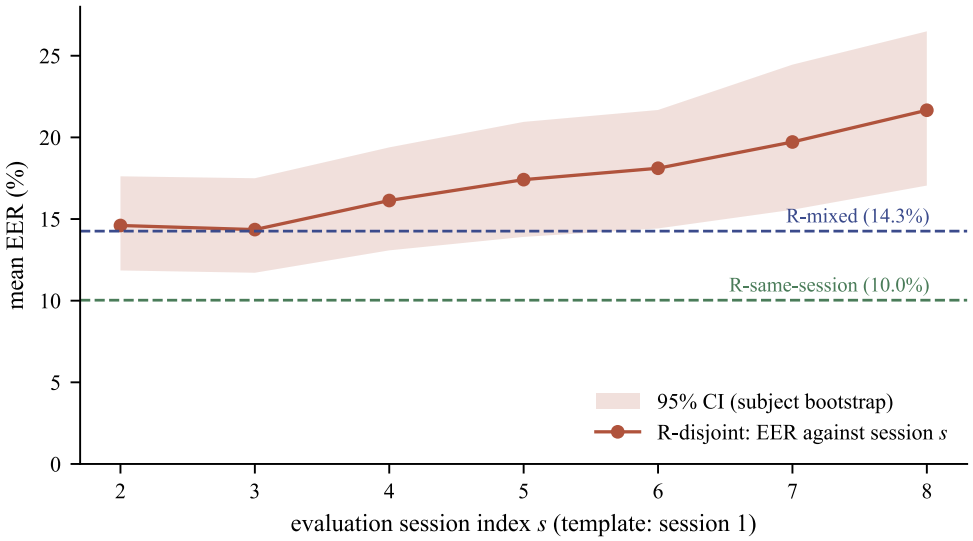

Figure 4: Drift of verification error on the CMU keystroke benchmark (51 subjects). Mean EER of the session-1-enrolled scaled Manhattan detector against each later session (R-disjoint), with the 95% band from bootstrap over subjects (1,000 resamples). Dashed lines: mean EER under R-mixed and R-same-session for the same subjects and detector. The horizontal axis is session index; elapsed time between sessions is not available in the public data. Evidence class: reanalysis of public data.

Figure 5 places the three CMU protocols side by side and adds the Balabit cohort. The qualitative ordering, optimistic protocols below deployment-shaped ones, is the finding that survives the small-cohort noise; the exact Balabit means do not separate statistically, as discussed below.

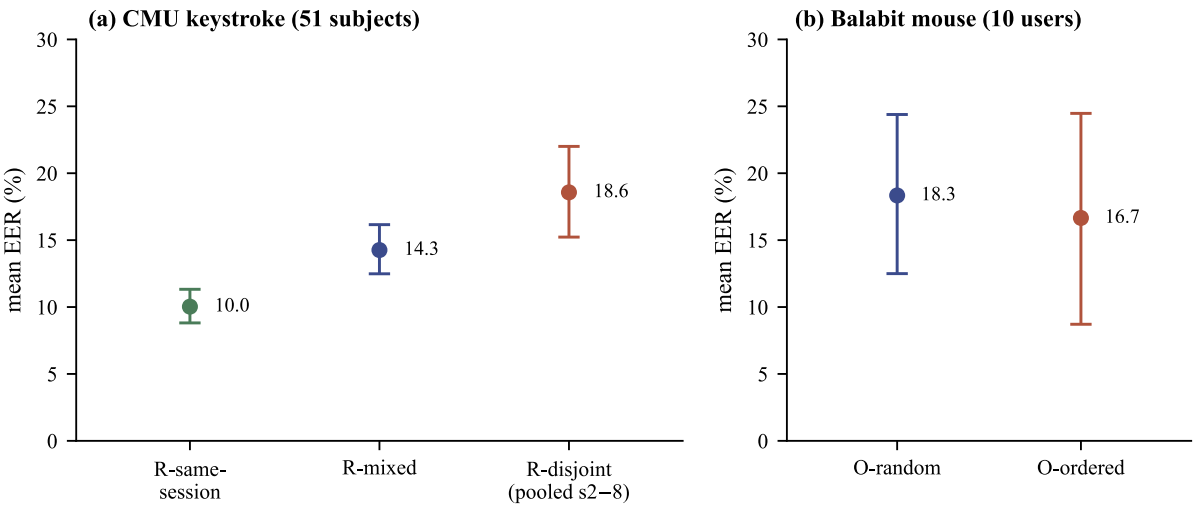

Figure 5: Protocol comparison. (a) CMU keystroke benchmark (51 subjects): mean EER under the three protocols. (b) Balabit mouse cohort (10 users, 65 sessions): mean EER under O-random and O-ordered. Error bars: 95% percentile intervals from bootstrap over participants (1,000 resamples). Numeric labels give the means. Evidence class: reanalysis of public data.

5.4 Which features drift, and which discriminate

Figure 6 crosses the two per-feature quantities: the all-sessions ICC (trait strength) and the drift index of Equation 4. The map has structure. Hold times cluster at high ICC (0.55 to 0.77) with intermediate drift (family mean 0.73 within-subject SDs). Latencies scatter at low ICC with the widest drift range: the two latencies into the digit 5 (DD.e.five and UD.e.five) drift by 1.13 and 1.16 within-subject SDs respectively, the largest values on the map, consistent with subjects initially hunting for the digit and Shift chord and then automatizing that transition; the smallest drift belongs to DD.o.a at 0.47. Family mean drift is nearly flat across families (H 0.73, DD 0.70, UD 0.72), so drift is not simply the complement of trait strength; the rank correlation between ICC and drift across the 31 features is weakly positive and not significant (Spearman rho 0.25, $p = 0.17$). The practical reading: no feature family is free of drift, and the drift-heaviest features are exactly the ones a same-session evaluation will happily exploit, since within a session they are as stable as anything else. The hold time of t is the instructive extreme case, combining the second-highest ICC on the map (0.75) with hold-family-maximal drift (0.82), person-specific and yet moving.

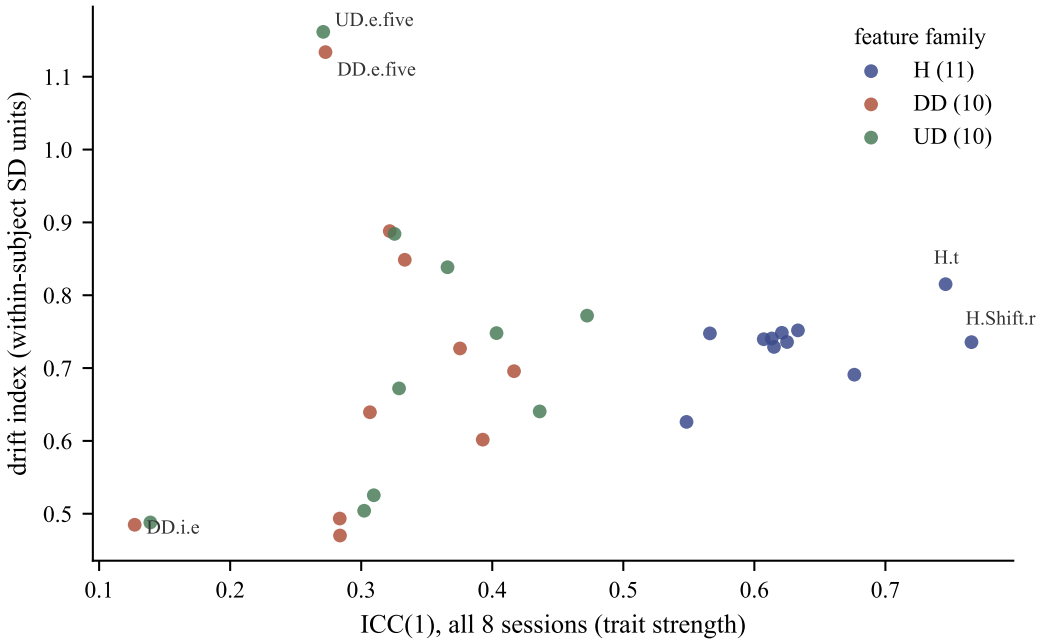

Figure 6: Trait versus state map for the 31 CMU timing features (51 subjects). Horizontal: ICC(1) over all 8 sessions. Vertical: drift index of Equation 4, the mean absolute displacement of later session means from the session-1 mean in units of pooled within-session SD, averaged over subjects. Colors: feature family; labels mark extremes. Point estimates. Evidence class: reanalysis of public data.

5.5 Mouse replication: small cohort, same shape

The Balabit analysis replicates both halves of the story in a second modality, at small scale and with honest imprecision. First, trait structure exists and is strong for movement-shape features: across the 10 users, ICC(1) over the 65 session vectors is 0.95 for direction-change rate, 0.94 for mean absolute turn angle, 0.91 for mean segment duration, and 0.88 for straightness. Speed features are moderately person-specific (0.47 for mean segment speed, 0.48 for its SD), while clicks per minute (0.42) and especially pause fraction (0.22) are dominated by within-user, presumably task-driven, variation. The ordering is sensible on motor grounds: turning behavior and segment shape reflect how a person aims and corrects a pointer, quantities tied to individual motor control, while pause fraction and click rate track what the user is doing in a session more than who the user is; the speed features sit between, person-linked but also load-dependent. It is plausible that part of the shape-feature separation reflects stable per-user environment (device, resolution, pointer settings) rather than motor behavior alone, a confound this dataset cannot resolve and that Jorgensen and Yu warned about [12]; we flag it rather than adjudicate it. A drift analysis parallel to the CMU practice curve is not attempted here, because the session identifiers are opaque and carry no verified order and the task content differs across sessions; the cohort therefore speaks to trait structure and protocol sensitivity, not to drift.

Second, verification error is far from the within-session optimism common in the modality's literature: 18.3% mean EER (12.5 to 24.4) under O-random and 16.7% (8.7 to 24.5) under O-ordered. The two protocols are statistically indistinguishable at this cohort size, and, because the session filename order is not verified as chronological, O-ordered cannot be interpreted as a clean temporal split; we report both without a directional claim. What the cohort does establish is the value of participant-level uncertainty: per-user EERs range from 0% to 34%, so any single-number summary of ten users is fragile, and the bootstrap interval says so. A ten-user result reported as a bare mean would look decisive; reported honestly, it is a replication of direction, not of magnitude.

6 A distributional account of digital identity

The results are individually unsurprising and jointly clarifying: person-specific structure is real (Table 1), the population moves (Figure 3), and the measured quality of identity evidence depends on whether the evaluation lets the verifier see across the movement (Table 2). This section states the account of identity that fits those facts.

Identity evidence is a likelihood comparison between distributions, not a distance to a point. Under Equation 1, what a person contributes to an observation is not a value but a conditional distribution: at time t in context c, person i emits samples centered on $\mu_i + \gamma_c + \delta_{i,c} + \tau_{i,t}$ with spread ε. A behavioral observation is evidence about identity exactly to the extent that it is more probable under one person's current distribution than under the alternatives'. The fingerprint metaphor collapses this to a fixed template plus a tolerance, which is adequate only when every non-trait term is negligible. The CMU data show the terms are not negligible even in the most controlled setting the field has: one task, one apparatus, motivated subjects, and still a 26% median shift in the core timing quantity across eight sessions, with the verifier's error growing in step with it.

The metaphor fails in three characteristic ways. First, drift makes templates stale: the R-disjoint curve is the direct picture, error rising monotonically with session distance from enrollment, and the template-update literature is the corresponding engineering response [9]. Second, context moves everyone at once: a device change or task change shifts distributions wholesale, so evidence collected in one context is systematically miscalibrated in another, as the touch and mobile studies show for posture and motion [16], [17]. Third, people are heterogeneous in a structured way: the menagerie results [18], [19] and our per-user EER spread on ten Balabit users (0% to 34%) both say that separation between distributions varies across people, so population-level error understates individual risk for some users and overstates it for others.

What follows for continuous verification. Three design consequences are worth stating in a product-neutral form. Calibration must be temporal: a score threshold tuned at enrollment time implicitly fixes the τ term at zero, so score-to-confidence mappings need recalibration on a timescale informed by measured drift curves rather than assumed stability. Abstention is a first-class output: when an observation is improbable under the enrolled distribution, "this is someone else" and "this person has drifted or changed context" are competing explanations, and a system that cannot say "insufficient evidence" will convert drift into false rejections and context novelty into false alarms. Context should be conditioned on, not averaged over: where context labels exist (device class, input modality, task type), maintaining per-context distributions or explicit context offsets replaces an inflated pooled variance with smaller conditional ones, recovering discriminability that pooling destroys. None of these require novel machinery; they require treating the identity model as an estimated, moving distribution with uncertainty, which is what the data say it is.

The distributional view also reframes enrollment. If the genuine-class distribution moves, enrollment is not an event but an estimation process that continues for as long as the identity is used: each verified interaction is also a candidate observation of the current distribution, and the template-update literature is best read as sequential estimation of $\mu_i + \tau_{i,t}$ under label uncertainty, with the acceptance gate deciding which observations may update the estimate [9]. The evaluation consequence is the one this report keeps returning to: a system that updates its enrolled distribution must be evaluated under a protocol in which the update process, not only the initial template, faces time-disjoint data; otherwise the update machinery is credited for surviving drift it never encountered. Reported numbers for adaptive systems inherit all of the reporting requirements of the next section, applied to the full update loop rather than to the first enrollment alone.

The account also disciplines interpretation of accuracy claims. A same-session EER is a measurement of trait separation at one moment, an interesting quantity for feasibility, and an upper bound on nothing. A session-disjoint EER at horizon s is a measurement of trait separation net of drift at that horizon. A cross-context EER measures transportability. These are different quantities with different uses, and a single undifferentiated number is most often the first one being mistaken for the third.

7 A reporting standard for longitudinal behavioral-identity claims

The empirical sections motivate a short list of reporting requirements. None are novel individually; ISO/IEC 19795-1 already requires several disclosures in spirit [22], and the dataset-shift literature supplies the vocabulary [21]. The point of the checklist is that on public benchmark data the difference between complying and not complying is the difference between reporting 10% and reporting 19% (Table 2), so these items belong in results, not appendices. Table 3 states the seven items; a claim that satisfies all seven is one whose headline number a reader can place on the trait-context-drift map without guessing.

Item	Requirement	Why it matters here
1. Person-disjoint splits	No person contributes to both detector development (feature selection, tuning) and reported evaluation.	Development-time leakage inflates separation of μ_i; standard practice, listed for completeness.
2. Time-disjoint evaluation	Template and test data separated in collection time (distinct sessions at minimum); same-session results labeled as within-session bounds.	Same-session evaluation hides τ: 10.0% vs 18.6% mean EER on identical CMU data.
3. Protocol disclosure	Template size and composition, genuine/impostor counts and sampling rule, and threshold policy stated exactly.	The protocol alone moves mean EER from 10.0% to 18.6% on CMU with everything else fixed.
4. Variance decomposition	Between/within-person variance share (ICC or equivalent) reported per feature or feature family.	Separates trait-like from state-like features; hold times (0.55–0.77) vs latencies (≤ 0.47) on CMU.
5. Drift curve	Error reported as a function of time or session distance from enrollment, not only pooled.	CMU error climbs 14.6% to 21.7% from session 2 to 8; a pooled number hides the slope.
6. Participant-bootstrap uncertainty	Confidence intervals from resampling participants, with the participant count stated next to every headline number.	Per-user EERs span 0–34% on 10 Balabit users; attempt-level intervals would be spuriously tight.
7. Context transfer	Where context varies (device, task, posture), report cross-context error or state explicitly that context was held fixed.	γ_c and $\delta_{i,c}$ are invisible in single-context data (CMU) and uncontrolled in the wild (Balabit).

Table 3: Reporting checklist for longitudinal behavioral-identity claims. Numeric justifications refer to the analyses of this report (CMU: 51 subjects; Balabit: 10 users; CIs by participant bootstrap, 1,000 resamples). Evidence class: reanalysis of public data (justification column).

Two remarks on use. First, items 2, 5, and 6 are cheap: they reuse data the study already has and require only a different split and a resampling loop. Item 7 is the expensive one, since it constrains collection; but a study that cannot afford context variation can still afford the sentence saying so. Second, the checklist is deliberately metric-agnostic: it constrains how numbers are produced and qualified, not which detector or feature set produces them, and it applies symmetrically to claims that behavioral identity works and claims that it does not.

8 Limitations

The CMU benchmark is a fixed-text password task performed by 51 subjects from a university population on laboratory apparatus; both the absolute error levels and the practice curve are specific to that setting, and free-text or familiar-text typing plausibly drifts differently (the initial password was unfamiliar by design, so part of the measured drift is the automatization of a novel motor sequence). The public CSV provides a session index and a one-day minimum separation, not calendar dates; we therefore make no claims in units of elapsed time, and neither should downstream users of these numbers. The Balabit cohort has ten users, uncontrolled tasks

and environments, no device or resolution labels, and session files whose lexicographic order is not verified as chronological, so its protocol contrast is indicative only, and part of its trait structure may be environmental rather than behavioral. Our detector choices are deliberately simple and untuned; a stronger or adaptive detector would lower absolute EERs and could narrow the protocol gap, but cannot remove the underlying distributional movement, which is a property of the data. Relatedly, per-participant EERs estimated from 25 to 350 genuine scores are coarse, and the interpolated threshold sweep cannot manufacture resolution that small score sets do not contain; the participant bootstrap propagates this granularity into the reported intervals rather than hiding it. The practice analysis identifies population nonstationarity without attributing cause: practice, fatigue, motivation, and time-of-day are confounded in session order, and we make no causal claim among them. Finally, the drift index of Equation 4 measures location shift of session means and is blind to variance changes within person, which the ICC comparison only partially captures.

What this paper does not claim. This report makes no claim that any behavioral signal uniquely identifies a person; ICCs well below 1 say the opposite. It reports no performance of any product or deployed system, and none of its numbers describe one; every EER in this report characterizes a specific public dataset under a specific stated protocol with a deliberately simple detector. It proposes no universal accuracy figure for keystroke or mouse biometrics, and it should not be cited as evidence that behavioral verification achieves any particular error rate in deployment.

9 Conclusion

On the two most familiar public datasets in behavioral identity, we measured three things with one fixed detector and deterministic, rerunnable analyses: how much of each feature's variance is a person (hold times 0.55 to 0.77 across sessions; latencies at most 0.47), how the population moves (median typing duration down 26% over eight sessions, 96% of subjects faster), and what the evaluation protocol alone does to the headline number (10.0% to 14.3% to 18.6% mean EER on identical data, with error rising from 14.6% to 21.7% as testing moves further from enrollment). A ten-user mouse cohort provides an imprecise cross-session illustration with wide intervals. The constructive conclusion is a change of object: behavioral identity is a per-person distribution with trait, context, and drift components, and both systems and studies should treat it as one. For systems, that means temporal calibration, abstention, and context conditioning. For studies, it means the seven-item reporting standard of Table 3, whose entire content is that the quantities this report had to measure the hard way should be visible in every longitudinal behavioral-identity claim from the start.

References

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004, doi: 10.1109/TCSVT.2003.818349.

[2] F. Monrose and A. D. Rubin, "Keystroke Dynamics as a Biometric for Authentication," *Future Generation Computer Systems*, vol. 16, no. 4, pp. 351–359, 2000, doi: 10.1016/S0167-739X(99)00059-X.

[3] A. A. E. Ahmed and I. Traore, "A New Biometric Technology Based on Mouse Dynamics," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, no. 3, 2007, doi: 10.1109/TDSC.2007.70207.

[4] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, "Continuous User Authentication on Mobile Devices: Recent Progress and Remaining Challenges," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 49–61, 2016, doi: 10.1109/MSP.2016.2555335.

[5] K. S. Killourhy and R. A. Maxion, "Comparing Anomaly-Detection Algorithms for Keystroke Dynamics," in *Proceedings of the 2009 IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, IEEE, 2009, pp. 125–134. doi: 10.1109/DSN.2009.5270346.

[6] Á. Fülöp, L. Kovács, T. Kurics, and E. Windhager-Pokol, “Balabit Mouse Dynamics Challenge Data Set.” [Online]. Available: https://github.com/balabit/Mouse-Dynamics-Challenge

[7] R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro, “Authentication by Keystroke Timing: Some Preliminary Results,” Santa Monica, CA, technical report R-2526-NSF, 1980. [Online]. Available: https://www.rand.org/pubs/reports/R2526.html

[8] F. Monrose and A. D. Rubin, “Authentication via Keystroke Dynamics,” in *Proceedings of the 4th ACM Conference on Computer and Communications Security (CCS)*, ACM, 1997, pp. 48–56. doi: 10.1145/266420.266434.

[9] R. Giot, B. Dorizzi, and C. Rosenberger, “Analysis of Template Update Strategies for Keystroke Dynamics,” in *Proceedings of the 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, IEEE, 2011. doi: 10.1109/CIBIM.2011.5949216.

[10] H. Gamboa and A. Fred, “A Behavioral Biometric System Based on Human-Computer Interaction,” in *Proceedings of SPIE 5404, Biometric Technology for Human Identification*, 2004, pp. 381–392. doi: 10.1117/12.542625.

[11] N. Zheng, A. Paloski, and H. Wang, “An Efficient User Verification System via Mouse Movements,” in *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS)*, ACM, 2011, pp. 139–150. doi: 10.1145/2046707.2046725.

[12] Z. Jorgensen and T. Yu, “On Mouse Dynamics as a Behavioral Biometric for Authentication,” in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, ACM, 2011, pp. 476–482. doi: 10.1145/1966913.1966983.

[13] C. Shen, Z. Cai, X. Guan, and R. A. Maxion, “Performance Evaluation of Anomaly-Detection Algorithms for Mouse Dynamics,” *Computers & Security*, vol. 45, pp. 156–171, 2014, doi: 10.1016/j.cose.2014.05.002.

[14] M. Antal, N. Fejér, and K. Buza, “SapiMouse: Mouse Dynamics-Based User Authentication Using Deep Feature Learning,” in *Proceedings of the 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, IEEE, 2021, pp. 61–66. doi: 10.1109/SACI51354.2021.9465583.

[15] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, “Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 136–148, 2013, doi: 10.1109/TIFS.2012.2225048.

[16] Z. Sitová *et al.*, “HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 877–892, 2016, doi: 10.1109/TIFS.2015.2506542.

[17] H. Crawford and E. Ahmadzadeh, “Authentication on the Go: Assessing the Effect of Movement on Mobile Device Keystroke Dynamics,” in *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS)*, USENIX Association, 2017, pp. 163–173. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/crawford

[18] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, “SHEEP, GOATS, LAMBS and WOLVES: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation,” in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, 1998. doi: 10.21437/ICSLP.1998-244.

[19] N. Yager and T. Dunstone, “The Biometric Menagerie,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220–230, 2010, doi: 10.1109/TPAMI.2008.291.

[20] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.

[21] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A Unifying View on Dataset Shift in Classification,” *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[22] International Organization for Standardization, “ISO/IEC 19795-1:2021, Information Technology: Biometric Performance Testing and Reporting, Part 1: Principles and Framework.” [Online]. Available: https://www.iso.org/standard/73515.html

[23] P. E. Shrout and J. L. Fleiss, “Intraclass Correlations: Uses in Assessing Rater Reliability,” *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979, doi: 10.1037/0033-2909.86.2.420.

[24] K. O. McGraw and S. P. Wong, “Forming Inferences About Some Intraclass Correlation Coefficients,” *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996, doi: 10.1037/1082-989X.1.1.30.

[25] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.