

Single-camera pose estimation for quantitative assessment of balance deficits: a proof-of-concept using graded acute alcohol intake

Nadine Domnik

nadine.domnik@hest.ethz.ch

ETH Zurich

Nerea Lopez

ETH Zurich

Tim Unger

ETH Zurich

Suno Diekmann

ETH Zurich

Alexander Andrea Tarnutzer

Cantonal Hospital Baden

Olivier Lambercy

ETH Zurich

Research Article

Keywords: Postural control, Balance assessment, Markerless motion capture, Human pose estimation, Berg Balance Scale, Responsiveness, Ceiling effect, Acute alcohol intake, Digital outcome measures, Neurorehabilitation

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11164923/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. O.L. is a member of the Editorial Board of the Journal of NeuroEngineering and Rehabilitation. He was not involved in the editorial handling or peer review of this manuscript. All other authors declare no competing interests.

Single-camera pose estimation for quantitative assessment of balance deficits: a proof-of-concept using graded acute alcohol intake

Nadine Domnik^{1*}, Nerea Lopez¹, Tim Unger^{1,4}, Suno Diekmann¹,
Alexander Andrea Tarnutzer^{2,3}, Olivier Lambercy¹

^{1*}Rehabilitation Engineering Laboratory, ETH Zurich, Gloriastrasse
37/39, Zurich, 8092, Switzerland.

²Neurology, Cantonal Hospital Baden, Im Ergel 1, Baden, 5404,
Switzerland.

³Faculty of Medicine, University of Zurich, Rämistr. 71, Zurich, 8006,
Switzerland.

⁴Data Analytics & Rehabilitation Technology (DART), Lake Lucerne
Institute, Vitznau, Switzerland.

*Corresponding author(s). E-mail(s): nadine.domnik@hest.ethz.ch;
Contributing authors: lopezn@student.ethz.ch; tim.unger@hest.ethz.ch;
diekmanns@student.ethz.ch; alexander.tarnutzer@ksb.ch;
olivier.lambercy@hest.ethz.ch;

Abstract

Background: Clinical balance scales are ordinal, depend on a rater's judgment, and saturate in people who function well, which limits their responsiveness. Whether a camera-derived measure can register fine changes in balance that such scales miss has not been tested, because designs that stratify participants by an existing scale make that scale the reference standard. We therefore used acute alcohol intake as a graded, reversible within-person perturbation against which the responsiveness of both measures could be judged.

Methods: Twenty-four healthy adults were assessed at three blood alcohol concentrations, 0.0, 0.6 and 1.0 ‰, on six standing tasks: two-legged stance, tandem stance and one-legged stance, each performed with eyes open and with eyes closed. All trials were recorded with a single consumer camera and processed by markerless pose estimation, yielding 167 candidate digital metrics, from which a small per-task subset was selected and combined into a single digital score.

The same recordings were rated from video by trained assessors on a 0–4 scale adapted from individual Berg Balance Scale (BBS) items, applying a uniform 30 s criterion to every task. The primary endpoint was whether a measure could rank a participant’s own three trials by dose, evaluated by leave-one-subject-out cross-validation, in which each participant is scored by a model built without their data, with feature selection repeated inside each fold.

Results: The digital score ordered the three doses more accurately than the concurrent rating in all participants scorable by both, and on all six tasks, significantly so in four. Mean dose-ordering accuracy was 0.86 against 0.59, and the pooled paired advantage 0.28 (95% confidence interval 0.23 to 0.32). Across the cohort, 77% of ratings were the maximum score. Among participants awarded the maximum both when sober and at 1.0 ‰, all nine testable digital comparisons showed a significant sober-to-high-dose change, with rank-biserial effect sizes of 0.61 to 1.00. The response was monotonic across doses in the subset rated at ceiling throughout. Inter-rater agreement was 83% exact overall but 30% once ceiling ratings were excluded, and the second rater’s scores reproduced the result.

Conclusions: Continuous metrics from a single camera resolved graded postural impairment that a single-item BBS rating did not register, including in participants the rating judged unaffected. A maximum BBS item rating indicates that performance fell inside the top category, not that balance is intact.

Keywords: Postural control; Balance assessment; Markerless motion capture; Human pose estimation; Berg Balance Scale; Responsiveness; Ceiling effect; Acute alcohol intake; Digital outcome measures; Neurorehabilitation

Background

Restoring and monitoring postural control is a central goal of neurorehabilitation [1, 2]. Balance depends on the continuous central integration of visual, vestibular, and proprioceptive feedback into coordinated motor output [3, 4], and this sensorimotor loop is disrupted across frequent neurological conditions such as stroke [5, 6] and Parkinson’s disease (PD) [7, 8], as well as in rare disorders affecting the cerebellum such as the spinocerebellar ataxias [9, 10].

To assess postural control in clinical practice, the standard approach is for a clinician to observe a patient performing a set of balance tasks, such as standing still, and to assign ordinal scores based on a subjective scoring framework. The most widely used and best-validated instrument for this purpose is the Berg Balance Scale (BBS) [11]. It comprises 14 tasks spanning both static and dynamic balance, with difficulty increasing as the base of support narrows (for example, the static items progress from a two-leg stance to tandem to single-leg stance). Each task is scored on a five-point ordinal scale from 0 to 4 against defined criteria, where 0 denotes inability to perform the task and 4 denotes safe, independent performance. One of the standing items is performed with eyes closed, following the Romberg principle that removing vision forces balance to rely on vestibular and proprioceptive feedback alone, so that disproportionate instability without sight points to a deficit in these non-visual channels [12].

While the BBS is well established in clinical use, it has several known limitations that fall into two categories. The first concerns its responsiveness [13]. Scoring depends on a clinician’s visual judgment and is therefore subjective and rater-dependent [14]. The scale is ordinal rather than continuous, so a given change in score does not correspond to a fixed change in balance ability [15]. It also shows pronounced ceiling and floor effects that compress variation at the ends of the range [16]. The second category concerns what the score reveals. A single ordinal value records how well a task was performed but not how it was performed, giving no insight into the underlying movement, such as trunk or arm sway, center of mass (CoM) excursion, or stepping, that produced it. Together these properties limit both the responsiveness of the BBS and the mechanistic insight it offers, motivating a move toward instrumented, kinematic assessment that measures and represents the movement itself.

Instrumented, kinematic assessment can be realized through several technologies, each with its own trade-off between accuracy and clinical feasibility. Optical motion capture (OMC) is the established gold standard for movement quantification, but its cost, dedicated laboratory space, marker placement, and lengthy setup make it impractical for routine clinical use [17]. Inertial measurement units (IMUs) are portable and inexpensive, and a single trunk- or pelvis-mounted sensor is widely used to quantify postural sway and stability [8]. Deriving full-body kinematics, however, requires a complete sensor array that is time-consuming to don and calibrate and therefore poorly suited to clinical routine [18]. Further, orientation estimates drift over time, and sensor-to-segment calibration error accumulates [19]. Force plates provide precise, objective measurement of center-of-pressure displacement, but capture only the net ground-reaction signal, leaving the trunk, arm and limb motion behind it unresolved. Computer-vision approaches recover kinematics directly from video, but the

validated work is dominated by depth (red-green-blue-depth (RGB-D)) cameras and multi-camera 3D reconstruction, which can be accurate yet again require specialized hardware and controlled setups that are not universally available [20, 21]. This leaves a clear opportunity in monocular pose estimation. A single consumer red-green-blue (RGB) camera, such as the one in any smartphone, requires no markers, no body-worn sensors, and no specialized hardware, and can in principle measure and assess movement anywhere. Recent progress in deep-learning pose estimation has made this feasible. Compact 2D pose models can track body landmarks reliably while remaining light enough to run on edge devices [22], and higher-capacity full-body models such as Sapiens [23] push accuracy further. This combination of accessibility and accuracy is why we evaluate the potential of single-camera markerless pose estimation in the present work.

Designing a protocol that systematically relates level of impairment to postural stability requires more than an accurate representation of movement. It also requires knowledge of the impairment itself, which means a controlled mechanism for inducing postural instability in a graded, reversible way. This is what establishing responsiveness demands. A responsive metric is one that changes when balance changes, and this property cannot be inferred from cross-sectional or observational data, where between-person differences confound the within-person change a clinical measure must track [13]. Stratifying participants by an existing clinical score such as the BBS does not resolve this but compounds it: using the BBS to define the levels of impairment makes it the reference standard, so candidate metrics can at best be shown to agree with it and necessarily inherit its ceiling, ordinal resolution, and rater dependence rather than improving on them. Establishing responsiveness instead calls for a controlled perturbation of the balance system. The perturbation must be graded, so that the dose-response relationship can be characterized. It must be reversible, so that participants serve as their own controls across conditions. It must also be safe to apply in healthy volunteers. Acute alcohol intake meets these criteria. It is a reversible, dose-dependent perturbation that transiently disrupts the cerebellar and vestibular control of posture, systems also affected in clinical balance disorders: acute alcohol intake produces transient spinocerebellar and vestibulocerebellar signs, with the postural signature of moderate doses closely resembling that of a chronic lesion of the cerebellar anterior lobe [24, 25]. Critically for a balance assessment, the relationship between blood-alcohol concentration and postural sway is well replicated and monotonic: body sway increases with alcohol dose under both eyes-open and eyes-closed conditions, and higher blood-alcohol levels are reliably distinguished from lower ones by quantitative posturography [26–29]. Acute alcohol intake therefore provides a controlled, graded, reversible within-person model of postural instability against which the responsiveness of candidate digital metrics can be quantified.

Despite the maturity of each of these components, no prior work has brought them together. Markerless pose estimation has been validated for balance assessment but largely with depth or multi-camera systems [20, 21], and single-camera and instrumented approaches alike have been built to reproduce the clinician’s score [30–32]. Such measures are validated by their agreement with the rating and therefore inherit its ceiling by construction, rather than being evaluated for responsiveness.

The dose-dependent effect of alcohol on balance, in turn, has been characterized only with instrumented force plates, never with markerless pose, and never benchmarked against the BBS. To address these gaps, this work evaluates a single-camera, markerless approach that derives continuous, multi-domain balance metrics, quantifies their responsiveness under a controlled and reversible within-person perturbation, and compares them directly against the concurrent clinical standard.

We hypothesize that continuous, pose-derived balance metrics are more responsive to a graded perturbation of postural control than the BBS. To test this, we use a single consumer RGB camera and monocular 2D pose estimation to derive continuous, multi-domain metrics, including CoM and trunk sway, joint angles, and discrete stepping, from a battery of static stances drawn from the BBS: the two-legged stance (TLS), tandem, and one-legged stance (OLS), each performed with eyes open and, extending the scale, also with eyes closed, yielding six tasks. Twenty-four healthy adults were assessed within-subject at three controlled blood alcohol concentration (BAC) levels (0.0, 0.6, and 1.0 ‰), with concurrent BBS scoring by two independent trained raters. For each metric we quantify responsiveness to the dose gradient by the size and monotonicity of the dose-response [14]. The most responsive metrics are then combined into a single digital score, which is benchmarked against the concurrent BBS item rating.

Methods

An overview of the approach and analysis is presented in Fig. 1. From single-camera video of three balance tasks, markerless pose estimation yields per-frame body key-points, from which a large set of candidate digital metrics is computed, drawn with few exceptions from established posturographic and clinical balance measures (Supplementary Table S2). A selection based on responsiveness to the alcohol-dose gradient is then applied in two ways. Applied once to all participants, it reduces the candidates to a single representative metric per task, which characterizes the dose-response descriptively. Re-applied inside each fold of a leave-one-subject-out (LOSO) cross-validation, it yields a per-fold set of metrics that an ordinal model, fit on the training participants with the ordered BAC level as the outcome, combines into a single digital score for the held-out participant. These held-out scores are compared head-to-head against the BBS item rating of the same trials.

Study Design and Participants

Twenty-six healthy, able-bodied adults took part in the balance experiment reported here, recruited within a larger study at the Rehabilitation Engineering Laboratory, ETH Zurich. Eligible participants were aged 18–70 years, free of cognitive deficits or concomitant disease that might affect motor performance, and able to follow the study procedures and give informed consent. Exclusion criteria were alcohol dependence or known alcohol intolerance, use of anxiolytic, antidepressant, neuroleptic, or sedative medication, and pregnancy or breastfeeding (all female participants of childbearing age were screened with a voluntary pregnancy test). Each participant performed the same six-task battery at three target BAC levels (0.0, 0.6 and 1.0 ‰) within a single session of ascending dose, so that the design is fully within-subject with each participant

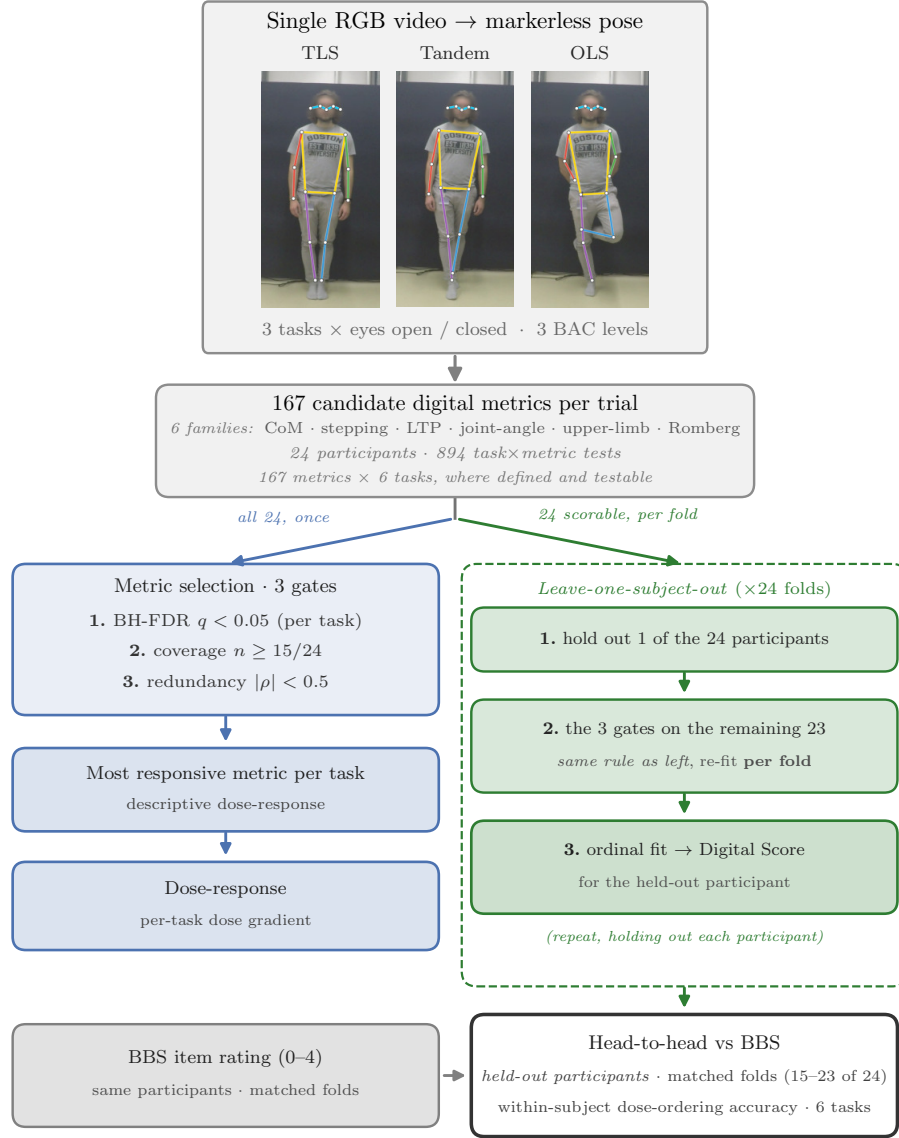


Fig. 1 Overview of the study and analysis. *Top:* three balance tasks, each performed with eyes open and closed at three BAC levels, shown for a representative participant with the 17-keypoint markerless-pose skeleton overlaid on the raw video frame. Markerless pose over all trials yields 167 candidate digital metrics (across CoM, stepping, LTP, joint-angle, upper-limb, and Romberg families). The same three-gate selection is then used two ways. *Left (descriptive, blue):* applied once to all participants, it keeps the most responsive metric per task and characterizes its dose-response. *Right (inference, green):* the same selection procedure is re-fitted inside leave-one-subject-out cross-validation, and an ordinal model combines the selected metrics into a single Digital Score for the held-out participant. *Bottom:* the held-out Digital Scores are compared against the concurrent BBS item rating by within-subject dose-ordering accuracy across the six task-vision conditions. Each comparison uses only participants that both the digital score and the rating could score, which is 15 to 23 of the 24 depending on the task. *Abbreviations:* BAC, blood alcohol concentration; BBS, Berg Balance Scale; CoM, center of mass; LTP, lower-trunk proxy; OLS, one-legged stance; TLS, two-legged stance.

serving as their own control. Each combination of one task at one BAC level is referred to as a trial, giving 18 trials analyzed per participant, and trial is used in this sense throughout. Every trial was video-recorded for scoring on the BBS, and the present analysis applies markerless pose estimation to these recordings, which were captured for review rather than for kinematic analysis.

Participants abstained from food for at least two hours before the session. The initial dose required to reach each target concentration was estimated per participant with the Widmark formula [33, 34]. Participants then ingested alcohol under supervision while breath alcohol was monitored every 5 min with a Dräger Alcotest 6000 (Dräger, Lübeck, Germany). Before each measurement, participants drank water to clear residual mouth alcohol, which can otherwise inflate the readout. The device returns a breath-based estimate of BAC, converting breath alcohol with a 2100:1 partition ratio. Monitoring continued until this estimate came within ± 0.1 ‰ of the target concentration, which defined the start of the assessment. If the target was not reached within 40 min of the initial consumption, or if two consecutive readouts plateaued or declined, an additional dose was administered. A fourth recording, obtained approximately two hours after the 1.0 ‰ trial, falls on the descending limb of the BAC-time curve. Because the relationship between BAC and impairment differs between limbs (acute tolerance, the Mellanby effect [35]), it was not analyzed.

All trials were captured with a single consumer-grade camera positioned in front of the participant, recording at 1920×1080 pixels and 30 frames per second (fps).

Balance Assessments and Clinical Reference

The six static-balance tasks were administered in a fixed order of increasing difficulty: TLS, tandem stance (one foot placed directly in front of the other), and OLS, each performed with eyes open and with eyes closed and held for up to 30s. These tasks derive from the BBS: TLS (eyes open and eyes closed) and the eyes-open tandem and OLS items are standard BBS items, whereas the eyes-closed Tandem and OLS conditions extend the BBS to probe balance under withdrawn vision.

Each trial was scored on a 0–4 ordinal rating scheme adapted from the BBS (higher scores indicate better balance), yielding a per-task rating rather than the 14-item BBS total. We refer to this throughout as the BBS item rating, and reserve BBS for the full 14-item scale. It should be noted that this differs from the conventional BBS: it applies a uniform 30s criterion where the BBS specifies a different duration for each item, and two of the six conditions are not BBS items at all. All comparisons reported here therefore use this adapted per-task rating as the clinical reference. The 14-item BBS total was never scored or used as a comparator. The rating scheme was graded against the 30s window as 4 (position held safely for the full 30s), 3 (held at least 20s but with supervision, an assessor standing by without physical contact), 2 (held at least 10s unsupported on the first attempt), 1 (reached 10s only after several attempts), and 0 (unable to hold 10s). The same thresholds applied to each task, substituting the required posture (TLS, tandem, or OLS) and vision condition. Each trial was scored from video by a trained rater, and a second trained rater independently re-scored all trials to quantify reliability. The adapted rating scheme was reviewed and approved by a consultant neurologist.

231 Markerless Pose Estimation

232 Body pose was estimated with Sapiens-1B, a one-billion-parameter vision transformer
233 pretrained on a large corpus of human images [23], applied frame by frame to each
234 trial video. Sapiens operates top-down: an off-the-shelf detector (RTMDet, [36]) first
235 localizes each person, and Sapiens then regresses 17 body keypoints per detection in
236 the COCO layout [37] (nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and
237 ankles), giving a two-dimensional trajectory for each keypoint across the trial.

238 Signal Processing and Metric Extraction

239 The first processing steps repaired frame geometry and checked signal quality to
240 yield, for each keypoint, a cleaned trajectory expressed in body-relative units. These
241 steps comprised isolation of the participant from multi-person detections, reconstruc-
242 tion of ankles that left the frame, compensation of camera motion, conversion to
243 body-height-normalized coordinates, outlier removal, gap interpolation and low-pass
244 filtering, and rejection of keypoints that remained too sparse over a trial. Each trial
245 was then reduced to the balance task itself by a task-window detection, and OLS tri-
246 als were mirrored where necessary so that the stance leg is always the right leg. Full
247 details, including all threshold values, are given in the Supplementary Methods and
248 Supplementary Table S1.

249 From the cleaned keypoints we derived 167 candidate digital metrics, organized
250 into six families. With few exceptions, these are established quantities from the pos-
251 turography and balance-assessment literature [38–43] rather than newly introduced
252 definitions. The full list, with definitions and sources, is provided in Supplementary
253 Table S2.

254 The CoM family (46 metrics) tracks the whole-body CoM, computed as a mass-
255 weighted average of the body segments [38], and summarizes its sway with standard
256 posturographic quantities, namely amplitude, mean velocity, path length, the 95%
257 prediction ellipse, and movement smoothness. They are computed primarily on the
258 mediolateral axis, the direction most sensitive to alcohol and best resolved by a frontal
259 camera [26, 39–41]. The lower-trunk proxy (LTP) family (23 metrics) applies the same
260 sway measures to a lower-trunk landmark so it largely duplicates the CoM family. The
261 joint-angle family (22 metrics) comprises frontal-plane segment inclinations (shoulder,
262 pelvis, head, trunk, and shank), each summarized by range, variability, and angular
263 velocity, together with inter-segment coordination measures [44, 45]. The stepping
264 and task-endpoint family (37 metrics) captures discrete instability and failure events,
265 OLS time-to-failure and touch-downs, step counts and timing, and tandem stance
266 width, following established clinical-balance and field-sobriety endpoints [42, 46–48].
267 The Romberg family (18 metrics) takes eyes-closed-to-eyes-open ratios of the sway
268 measures, indexing reliance on vision [49]. Finally, an upper-limb family (21 metrics)
269 quantifies arm motion as a compensatory response [43].

270 Because a single frontal camera resolves mediolateral but not anteroposterior
271 sway, all measures are restricted to the frontal plane. A subset of the digital metrics
272 described above, namely the nonlinear measures (sample entropy and recurrence quan-
273 tification), the inter-segment coordination measures, and the continuous upper-limb

measures, have not previously been used in acute alcohol or markerless-pose research and are therefore treated as exploratory [50–53].

Data Analysis

Responsiveness and Metric Selection

All analyses were performed on the N participants who completed all three dose levels. The analysis proceeds in three stages: each metric’s responsiveness to the dose gradient is quantified and used to select a compact set, that set is combined into a digital score inside cross-validation, and that score is compared against the concurrent BBS item rating on the same trials. Given the sample size, this procedure is treated as exploratory and descriptive [14]. A global effect of BAC on each (task, metric) pair was first tested with the Friedman test, and the resulting p-values were corrected within each condition (the family of metrics tested in that condition) by the Benjamini–Hochberg false discovery rate (FDR) ($q < 0.05$) [54].

Not every measure is defined in every task, so the 167 candidates yield 896 rather than 1002 task–metric combinations, of which 894 could be tested (Fig. 1). Measures of CoM and lower-trunk sway, joint angles and upper-limb motion are defined in all six tasks, as are 24 of the 37 stepping measures. Of the remaining 13 stepping measures, ten presuppose a particular posture — the OLS endurance, hold and touch-down measures and the two arm-abduction measures, the tandem ankle separation, and the standing displacement asymmetry — while three describe the timing and direction of steps and are therefore defined only where steps occurred; each was computed only in the two conditions where it is defined. The 18 Romberg ratios contrast eyes-closed against eyes-open performance and are therefore defined once per stance rather than once per task, giving three values each. Multiplicity correction was applied within each of the nine resulting test families: the six tasks and the three Romberg stance pairs. Two upper-limb abduction measures were constant in eyes-open two-legged stance, where no participant raised an arm, and so admitted no test, leaving 894 tests over the 896 defined combinations. No minimum sample size was imposed at this stage; the coverage requirement described below applied only to metric selection.

For every metric we summarized responsiveness as a within-subject dose-ordering score. For each of the three dose pairs (0 vs 0.6, 0 vs 1.0, and 0.6 vs 1.0 ‰) we computed the fraction of participants whose values moved in the metric’s dose direction (ties counted as one-half), a within-subject area under the receiver-operating-characteristic curve (AUC). These three were averaged, so that 0.5 marks chance ordering and 1 marks all participants ordered correctly at every dose step. Page’s trend test for an ordered alternative was retained as a confirmatory check rather than a selection gate [55]. To be eligible for selection, an FDR-significant metric additionally had to be quantifiable in at least 15 of the N participants. Because a metric is only computed when the pose is tracked, and tracking fails most in the demanding eyes-closed and high-BAC conditions, a low-coverage metric reflects mainly the less-impaired participants who could sustain the pose. The most affected participants are precisely the ones missing, so such a metric appears artificially insensitive to BAC. This coverage floor is a criterion applied to metrics during selection, and is unrelated

317 to the number of participants available for the head-to-head comparison reported
318 later. The eligible metrics within each condition were then pruned for redundancy by
319 greedy elimination of any pair correlated at $|\text{Spearman } \rho| \geq 0.5$, keeping the more
320 responsive metric of each pair, and the single most responsive surviving metric per
321 task (TLS, tandem, OLS) was retained as that task’s representative metric, so that
322 the reported quantity remains a physical measure of movement in interpretable units.
323 Selection stability was separately characterized, but not used as a selection gate, by
324 repeating the pipeline on 200 random subsamples of the participants [56, 57].

325 Digital Score versus BBS Item Rating

326 Responsiveness was further expressed as the ability to discriminate the three BAC
327 levels, comparing the digital metrics against the concurrent BBS item rating on the
328 same trials, separately for each of the six tasks. Models were compared by LOSO
329 cross-validation over the N participants, in which each participant was held out in
330 turn for testing while the model was fit on the remaining $N - 1$. Within each fold,
331 the selection pipeline was re-fitted on the training participants only, using the same
332 eligibility and redundancy criteria as the main selection, but without reducing to a
333 single representative: every metric surviving the gates in that fold entered the model.
334 One criterion was added: among equally responsive redundant metrics the one quan-
335 tifiable in more participants was preferred, since a sparse metric costs a matched fold
336 for each held-out participant it cannot score. The selected metrics were z -scored and
337 combined by an ordinal logistic regression (all-threshold model [58]) with the ordered
338 BAC level as outcome, whose latent linear predictor served as a single digital score.
339 Keeping this set compact limits the number of coefficients estimated per fold relative
340 to the sample size. For the task with the lowest coverage we additionally repeated
341 the comparison using the most responsive metric quantifiable in all participants. For
342 each held-out participant we computed a within-subject dose-ordering accuracy: the
343 fraction of the three dose pairs the score ranked in the correct order, ties counted as
344 one-half. With three ordered levels and one trial each this is equivalent to a within-
345 subject AUC, and to Kendall’s τ_a rescaled to the unit interval as $(1 + \tau_a)/2$, so that
346 0.5 marks chance ordering. Because the comparison is made within each participant,
347 it controls for between-person differences. This ordering accuracy was averaged across
348 held-out participants and reported with a percentile 95% confidence interval (CI) from
349 a bootstrap over participants, resampling participants rather than individual trials so
350 that repeated observations within a participant are kept together.

351 Ordering accuracy is the primary endpoint because the question is whether a mea-
352 sure can rank one person’s three trials by dose. A one-versus-rest AUC is misleading
353 at the intermediate level. Because the digital score rises with dose, the 0.6 % trials
354 sit between the sober and high-dose trials. Pooling the outer two levels into one com-
355 parison group therefore puts half of that group below the intermediate trials and half
356 above, and a score that ranked all three of a participant’s trials perfectly would still
357 reach only 0.5 on that comparison. The macro-average over the three levels is pulled
358 down with it. Ordering accuracy is not subject to this constraint, because it is the

mean of the three dose-pair comparisons and each of those involves only two dose levels. It therefore decomposes exactly into them. The one-versus-rest macro-AUC and its per-level components are reported as a secondary summary.

The identical pipeline was run with the per-task BBS item rating as the sole predictor. Because ordering accuracy depends only on the ranking of scores, it is invariant to any monotone transform of a single predictor, so the ordinal fit is no advantage to the digital arm where one metric was selected. The coarse ordinal rating and the continuous digital score are directly comparable, and the two arms were evaluated on matched folds (participants present in both).

The two arms were then compared within participants on those matched folds. For each participant and task we formed the paired difference between the digital and rating ordering accuracies. The pooled mean paired difference was estimated by the same participant-clustered bootstrap (10,000 resamples) and is reported with a percentile 95% CI. Significance was assessed by a two-sided Wilcoxon signed-rank test on participant-level mean differences, and per-task differences by the same test, corrected across the six tasks by the Benjamini–Hochberg FDR. Finally, we asked how many metrics were required. The digital score was rebuilt from an increasing number of the most responsive metrics (the single best, then the best two, and so on), leaving the primary comparison unchanged (Supplementary Table S3). This analysis was specified after the primary comparison.

Concurrent Validity, Ceiling, and Inter-Rater Reliability

Concurrent validity was assessed by correlating each selected digital metric with its concurrent BBS item rating (Spearman), with subject-cluster bootstrap CIs, plus the participant-total composite correlation (Pearson and Spearman) as an overall summary. Because the digital scores are pooled from per-fold models, they are comparable in relative rather than absolute terms, so these correlations are interpreted as relative agreement.

The BBS item rating ceiling was quantified as the fraction of trials awarded the maximum score of 4, per task and per BAC level. Two analyses then asked whether the digital metrics resolved impairment where the BBS item rating could not. In the subset of participants whose BBS item rating sat at ceiling both when sober and at 1.0 ‰, the sober-to-high-dose change in the digital metric was tested with the Wilcoxon signed-rank test and summarized by the rank-biserial effect size. In the stricter subset whose rating sat at ceiling across all three doses, a monotonic dose-response was tested with the Friedman test and Page’s trend test. Both analyses were corrected by the Benjamini–Hochberg FDR. Metrics for which too few participants were at ceiling to form a paired comparison were not testable and are omitted, and comparisons with fewer than six non-zero differences were flagged, since the exact two-sided Wilcoxon test cannot reach $p < 0.05$ below that sample size.

Inter-rater reliability between the two raters was reported as exact agreement, agreement within one point, mean absolute difference, and linearly and quadratically weighted Cohen’s κ [59] on the full set, with exact agreement additionally reported for the non-ceiling subset.

Results

Participants and Alcohol Intake

We first confirmed that the intended dose gradient was reached. Of the 26 participants who underwent the balance battery, two reported feeling unwell during the session and dropped out. Because their balance data might not be comparable with the rest, their data were excluded without analysis, leaving a final cohort of 24 healthy adults (11 men, 13 women). Participants were young (mean age 26.2 ± 4.6 years, range 21–43), with a mean body-mass index (BMI) in the normal range ($22.5 \pm 2.3 \text{ kg m}^{-2}$). Their characteristics are summarized in Table 1.

Table 1 Participant characteristics
($N = 24$). Values are mean $\pm$ SD (range).
Abbreviations: BMI, body-mass index.

Characteristic	Value
Sex (men / women)	11 / 13
Age (years)	26.2 ± 4.6 (21–43)
Height (cm)	173.1 ± 9.2 (157–189)
Weight (kg)	67.7 ± 10.9 (46.0–90.2)
BMI (kg m^{-2})	22.5 ± 2.3 (18.0–26.9)

BAC, measured immediately before each test sequence, confirmed that the intended dose gradient was reached. Readouts averaged 0.00 ‰ at the sober baseline, 0.59 ± 0.05 ‰ (range 0.51–0.69) at the low target, and 0.99 ± 0.04 ‰ (range 0.90–1.08) at the high target. Every readout fell within the predefined ± 0.1 ‰ acceptance window of its target, so the three conditions’ measured ranges did not overlap and each participant was at their intended dose level. The nominal targets are therefore used as an ordered grouping factor throughout.

A Responsive Sway Metric

We then asked how broadly the dose gradient was detectable across the metric set. Balance kinematics responded extensively to the dose gradient. Of the 894 tested task-metric combinations, 532 (59%) showed a significant overall effect of BAC after FDR correction (Fig. 1). These responses were consistently ordered: 97% ran in the direction of greater instability, and a monotonic $0.0 \rightarrow 0.6 \rightarrow 1.0$ trend was significant by Page’s test in 97% of them. Significant metrics spanned the CoM, LTP, joint-angle, stepping, and upper-limb families. Many of these metrics were mutually correlated. The two remaining gates in Fig. 1, coverage and redundancy, reduce them to a compact non-redundant set of 11, one to five per task. Taking the single most responsive metric per task yielded three representatives, and all three quantified the same construct, mediolateral sway velocity, though by two different metrics, namely whole-body CoM velocity in TLS and OLS, and the variability of lower-trunk mediolateral velocity in

431 tandem stance. Each is reported in its eyes-open condition for comparability across
 432 tasks.

433 The three representatives are shown across the dose gradient in Fig. 2. All rose
 434 gradually with dose. Each ordered the sober below the high-dose condition almost
 435 perfectly (ordering accuracy 0.91–1.00) and was monotonic within 73–80% of partici-
 436 pants. Sway velocity also resolved the low dose, where discrete stepping does not: the
 437 sober-to-0.6 ‰ ordering accuracy was 0.77–0.83, so a majority of participants already
 438 swayed faster at 0.6 ‰. The gradient was nonetheless cleanest at the high dose.

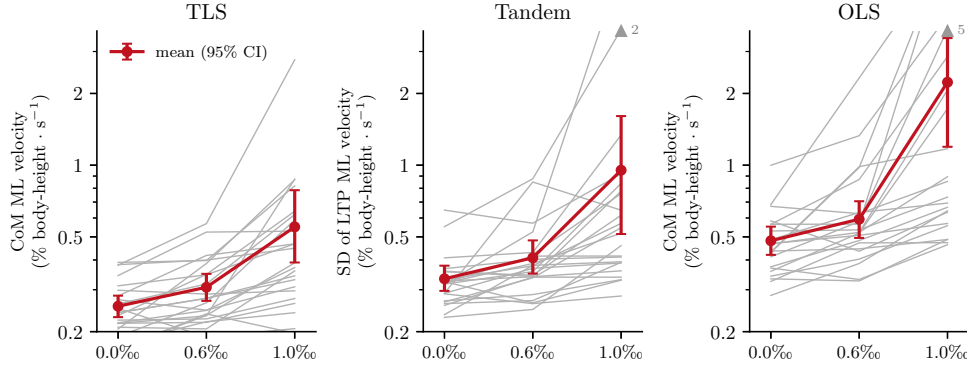


Fig. 2 Dose response of the representative sway metric for each task across the sober, low (0.6 ‰) and high (1.0 ‰) conditions. The three panels show two distinct metrics, both quantifying mediolateral sway velocity. Gray lines are individual participants, the red line is the mean (95% bootstrap CI). All three rise across the three doses, with a smaller separation already at the low dose. Each metric is shown in its eyes-open condition for comparability. Y-axes are logarithmic and clipped to the bulk of the distribution; gray triangles mark participants above the axis at each dose. *Abbreviations:* CI, confidence interval; CoM, center of mass; LTP, lower-trunk proxy; ML, mediolateral; OLS, one-legged stance; SD, standard deviation; TLS, two-legged stance.

439 As a check on the coverage floor, we confirmed that the exact threshold was not
 440 critical: the selected representatives were identical for floors from 12 to 18, each quan-
 441 tifiable in 20 to 23 participants, while the sparsely tracked metrics removed by the
 442 floor never ranked among the most responsive in any task.

443 As a complementary analysis, we re-ran the full pipeline on 200 random partici-
 444 pant subsamples. The single best metric was re-selected in 38–98% of subsamples and
 445 the selected sets overlapped one another with a mean Jaccard index of 0.17 to 0.48,
 446 depending on the task. In five of the six conditions the most frequently selected metric
 447 was a continuous mediolateral sway metric, the alternatives being other sway metrics
 448 such as spectral arc length (SPARC), path length, sway-frequency dispersion, or the
 449 mean and standard deviation (SD) of velocity (Supplementary Table S2). The excep-
 450 tion was eyes-closed OLS, where non-sway metrics competed. We therefore report
 451 the metric by its construct, treating the named metric as a representative example
 452 rather than a uniquely optimal choice. The upper-limb metrics were likewise respon-
 453 sive (84 of the 124 tested) but were pruned as redundant with trunk and CoM sway.

One metric family was the exception. The eyes-closed-to-eyes-open Romberg ratios showed no dose effect at all (0 of 54 significant), although metrics computed within the eyes-closed conditions themselves were responsive.

Digital Score versus BBS Item Rating

For each task we compared how well the digital score and the concurrent BBS item rating could order a participant’s three BAC levels under LOSO cross-validation. The primary endpoint was the within-subject dose-ordering accuracy: for each held-out participant, the fraction of their three dose pairs that the score placed in the correct order (ties one-half), averaged over participants. The one-versus-rest macro-AUC is reported as a secondary summary. The digital score and the BBS item rating are evaluated on matched folds, restricted to the participants that both measures could score, so every comparison is made within the same people. This is the primary comparison. The analyses that follow check whether its result depends on how the digital score was built.

The digital score out-ordered the BBS item rating on every task (Fig. 3). Its ordering accuracy fell in a high band of 0.78–0.93 across the six tasks (mean of the six task values, 0.86), whereas the BBS item rating ranged from chance (0.50) to 0.73 (mean 0.59 on the same average). The advantage was largest where the BBS item rating was lowest: in the eyes-open TLS and tandem-stance tasks and eyes-closed TLS, where the BBS item rating sat at or near chance (0.50–0.55), the digital score reached 0.86–0.91, a gain of 0.33–0.40. It was smallest in eyes-closed OLS, the task on which the rating carried the strongest ordering signal (0.73). Even there the digital score was higher (0.78).

One participant contributed no matched fold in any of the six tasks, leaving 23 participants and 122 participant–task observations for this comparison. Pooled across tasks, the mean paired within-participant advantage of the digital score over the BBS item rating was 0.279 (95% CI 0.230 to 0.323) weighting each participant–task observation equally. Every one of the 23 participants was ordered more accurately by the digital score (two-sided Wilcoxon signed-rank test on participant-level mean differences, $p = 2.4 \times 10^{-7}$). Per task the advantage was positive in all six tasks and reached significance in four after Benjamini–Hochberg correction: eyes-open TLS (+0.399, 95% CI 0.312 to 0.471), eyes-closed TLS (+0.357, 0.278 to 0.437), eyes-open OLS (+0.333, 0.250 to 0.417) and eyes-open tandem stance (+0.326, 0.220 to 0.424), all $q < 0.001$. The two remaining differences were smaller and not significant: eyes-closed Tandem stance (+0.135, 0.008 to 0.254; $q = 0.200$) and eyes-closed OLS (+0.044, −0.100 to 0.200; $q = 0.812$), the latter also having the fewest matched folds ($n = 15$).

The ordering accuracy decomposes exactly into the three dose-pair comparisons (Supplementary Table S5). The digital score exceeded the BBS item rating on 16 of the 18 task–pair combinations. The two exceptions fell in the two tasks with the least ceiling: the arms tied on the sober-versus-0.6‰ comparison in eyes-closed Tandem stance (0.71 each), and the rating was higher on the 0.6-versus-1.0‰ comparison in eyes-closed OLS (0.77 vs. 0.73). Ordering was most accurate for the sober-versus-high-dose pair (digital 0.86–1.00, mean 0.92), intermediate for the 0.6-versus-1.0‰ pair

(0.73–1.00, mean 0.88), and least accurate for the sober-versus-0.6 ‰ pair (0.71–0.83, mean 0.77).

The secondary one-versus-rest macro-AUC reproduced the advantage on five of the six tasks (digital 0.65–0.84 vs. rating 0.50–0.74; Supplementary Table S6). In eyes-closed OLS the digital macro-AUC (0.73) fell just below the rating (0.74), while the digital score led on the primary ordering accuracy (0.78 vs. 0.73). As anticipated in the Methods, the macro-average is held down by the intermediate level: the digital score’s 0.6 ‰ component sat near chance (0.36–0.67 across tasks) while its sober and high-dose components reached 0.79–0.90 and 0.80–1.00.

Repeating the comparison with the whole-body CoM mediolateral velocity measure, which was quantifiable in all 24 participants, the digital score exceeded the BBS item rating on ordering accuracy (0.81 vs. 0.72). Because this measure could score every participant, these matched folds comprise all 24 rather than the 15 available for the primary metric, which is why the rating’s accuracy here (0.72) differs from its value on the 15 matched folds (0.73).

Two analyses confirmed the comparison did not hinge on modeling choices. Capping the per-fold non-redundant set at an increasing number of the most responsive metrics (Supplementary Table S3) left the digital score above the BBS item rating at every reachable setting but one: with a single metric in eyes-closed OLS it reached 0.70 against the rating’s 0.71 on the same folds. In four of the six tasks the per-fold set contained at most two metrics and the cap was rarely binding, so the score was effectively a single metric throughout. In the two tasks where more metrics were available, adding a second raised the ordering accuracy from 0.79 to 0.92 in eyes-closed TLS, settling at 0.90 with the cap released. The corresponding change in eyes-closed OLS (0.70 to 0.78) is not directly comparable, because the matched-fold set falls from 22 participants at one metric to 15 at two or more.

Expressing each metric relative to the participant’s own sober trial did not improve discrimination and slightly reduced it (Supplementary Section S2, Supplementary Table S4), so absolute values were retained.

BBS Ceiling and Inter-Rater Reliability

We next quantified how much of the rating scale was in use. Across the 24 participants and six tasks, 77% of BBS item ratings were the maximum score of 4. The proportion at ceiling fell gradually with dose, from 88% at the sober baseline to 76% at 0.6 ‰ and 66% at 1.0 ‰, and differed more between tasks (Fig. 4A). Every participant scored 4 on eyes-open TLS at all three doses, and 96% of ratings were at ceiling in eyes-closed TLS and in eyes-open tandem stance. The proportion was 85% in eyes-open OLS, 62% in eyes-closed Tandem stance, and 21% in eyes-closed OLS, where the mean rating was 2.1 and 58% of ratings were 2 or below. The full distribution of ratings by task and dose is given in Supplementary Fig. S1, the counterpart to the dose-response of the digital metrics in Fig. 2.

We then examined the digital metrics in the participants for whom the BBS item rating recorded no change, that is, those awarded the maximum score of 4 at both of the doses being compared. Among participants rated at ceiling both when sober

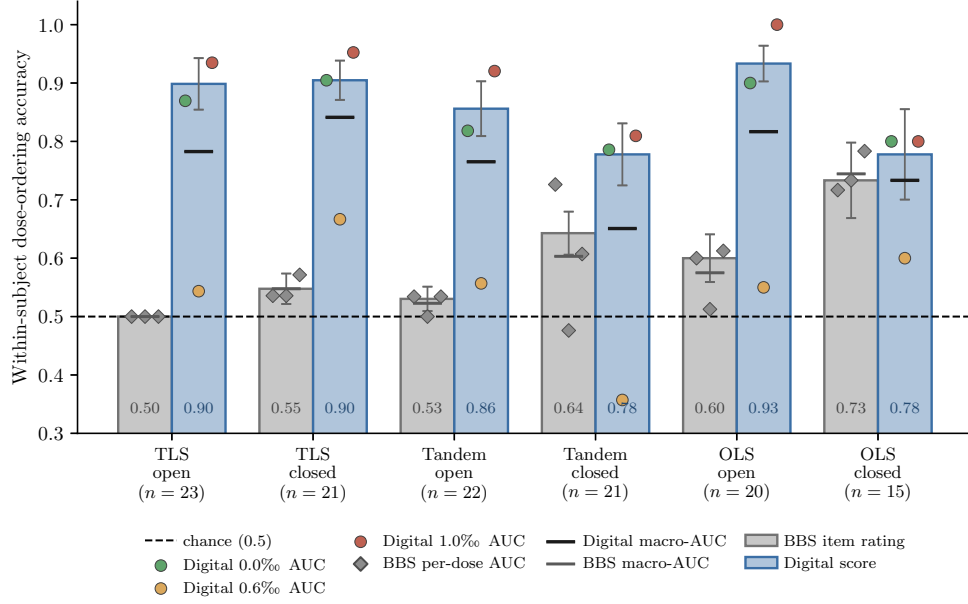


Fig. 3 Digital score versus BBS item rating: ordering of the three BAC levels across the six tasks. Bars show the primary endpoint, the within-subject dose-ordering accuracy on matched folds (participants scorable by both arms; matched n below each task) for the digital score (blue) and the per-task BBS rating (gray): per held-out participant, the fraction of the three dose pairs the score ordered correctly, averaged over participants (LOSO). Whiskers are 95% bootstrap confidence intervals. Overlaid on each bar is the secondary one-versus-rest macro-AUC decomposition, three per-dose AUCs (digital: colored circles, green 0.0, amber 0.6, red 1.0 ‰; BBS: gray diamonds; spread left-to-right by dose) and their mean, the macro-AUC (horizontal dash: black digital, gray BBS). For the digital score the weak intermediate 0.6 ‰ dose (as low as 0.36 in eyes-closed Tandem stance) pulls the macro-AUC below the near-ceiling 1.0 ‰ separation, whereas the BBS per-dose AUCs cluster near chance except on the two hardest tasks. Both metrics share the ordering-accuracy axis and the 0.50 chance line (dashed). The digital score exceeds the BBS on every task on the primary ordering accuracy. *Abbreviations:* AUC, area under the receiver-operating-characteristic curve; BAC, blood alcohol concentration; BBS, Berg Balance Scale; LOSO, leave-one-subject-out; OLS, one-legged stance; TLS, two-legged stance.

and at 1.0 ‰, all nine testable task-metric comparisons showed a significant sober-to-high-dose change after FDR correction, with rank-biserial effect sizes of 0.61 to 1.00 (mean 0.92) in 10 to 23 participants (Supplementary Fig. S2). In the stricter subset rated at ceiling at all three doses, all nine comparisons showed a significant monotonic dose-response by Page’s trend test after FDR correction, with 43 to 79% of participants monotonic in 9 to 23 participants. The two remaining metrics, both in eyes-closed OLS, could not be tested in either analysis, only one participant being at ceiling at both doses in that task. Repeating both analyses with the second rater’s ratings gave the same result.

The two raters agreed exactly on 83% of the 432 paired ratings and within one point on 92%, with a mean absolute difference of 0.27 points, Cohen’s $\kappa = 0.51$, and

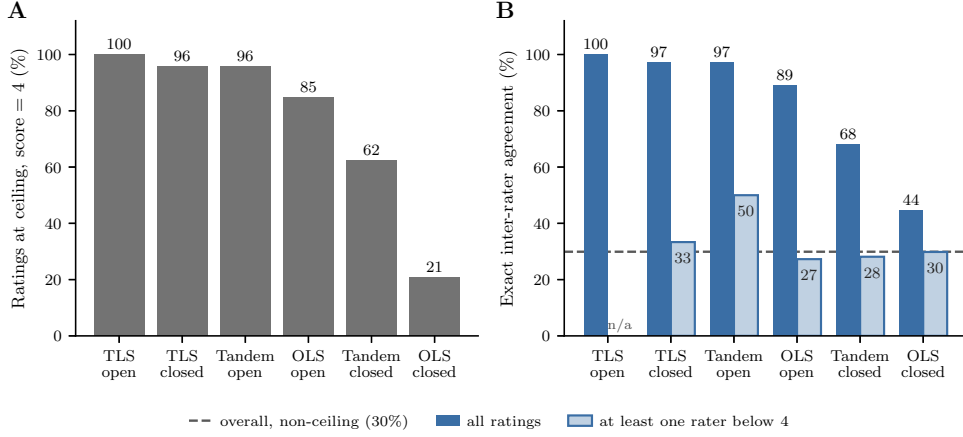


Fig. 4 BBS rating ceiling and inter-rater agreement by task. **A** Percentage of BBS ratings awarded the maximum score of 4, per task, across all participants and doses. **B** Exact agreement between the two raters, over all ratings and over only those ratings for which at least one rater scored below 4. The latter comprise 0, 3, 4, 11, 32 and 57 ratings for the six tasks in the order shown. Tasks are ordered by ceiling in both panels. Agreement over all ratings falls with the ceiling, whereas agreement on non-ceiling ratings is similar across tasks. *Abbreviations:* BBS, Berg Balance Scale; OLS, one-legged stance; TLS, two-legged stance.

linearly and quadratically weighted κ of 0.65 and 0.76. Agreement over all ratings fell with the proportion at ceiling (Fig. 4), from 100% in eyes-open TLS to 44% in eyes-closed OLS. Ratings for which both raters awarded a 4 agree by definition, so we also computed agreement over the 107 ratings for which at least one rater scored below 4. On these, agreement was 27 to 50% across tasks, with no ordering by task, and 30% overall. Eyes-closed OLS contributed 57 of these ratings, and there the two raters differed by 0.90 points on average. Repeating the head-to-head comparison of the previous section with the second rater's ratings reproduced the result on all six tasks, with a mean ordering accuracy of 0.86 for the digital score and 0.57 for the BBS item rating.

Concurrent Validity

Finally, we asked whether the digital metrics track the same construct the rating measures. Aggregated per participant, the digital score and the BBS item rating were strongly correlated (Spearman $\rho = 0.88$, Pearson $r = 0.94$, both $p < 0.001$). Because the digital scores are pooled from per-fold models, this reflects relative rather than absolute agreement.

At the level of individual metrics, 7 of the 10 testable metric-task pairs correlated significantly with the concurrent BBS item rating under a subject-cluster bootstrap (Fig. 5). Every significant correlation was negative, greater instability corresponding to a lower rating. Magnitudes across the ten pairs ranged from 0.10 to 0.75. The three largest were mediolateral CoM velocity in eyes-closed Tandem stance ($\rho = -0.75$),

573 head angular velocity in eyes-closed OLS ($\rho = -0.62$), and mediolateral CoM velocity
574 in eyes-open OLS ($\rho = -0.55$). The three pairs that did not reach significance were
575 all in eyes-closed TLS, where 96% of ratings were at ceiling. In eyes-open TLS the
576 correlation was not computable, the BBS item rating there having zero variance.

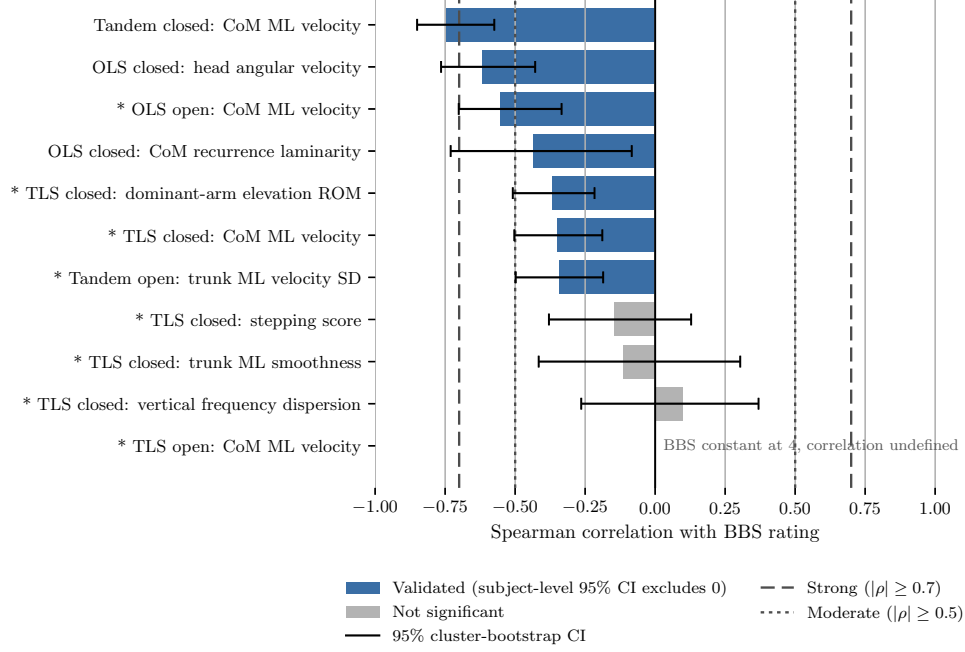


Fig. 5 Correlation between each selected digital metric and the concurrent BBS rating. Bars are Spearman correlations across all participant-dose observations, whiskers are 95% subject-cluster bootstrap CIs, and color indicates whether the interval excludes zero. A negative correlation means that greater instability corresponds to a lower BBS rating. An asterisk marks tasks in which the BBS reference was near ceiling, which limits the attainable correlation. Eyes-open TLS is shown without a bar: every rating there is a 4, so the correlation is undefined rather than small. *Abbreviations:* BBS, Berg Balance Scale; CI, confidence interval; CoM, center of mass; ML, mediolateral; OLS, one-legged stance; ROM, range of motion; SD, standard deviation; TLS, two-legged stance.

577 Discussion

578 We investigated whether continuous, pose-derived balance metrics are more responsive
579 than the BBS item rating to a graded perturbation of postural control. Our results
580 show that they are. A digital score derived from a single consumer camera ordered
581 each participant's three BAC levels more accurately than the concurrent BBS item
582 rating, and did so on all six tasks. Each participant served as their own control, and
583 both the digital score and the BBS item rating were derived from the same trials, so
584 between-person differences in baseline balance cannot account for the gap. Because

the impairment was induced rather than observed, graded and reversible within each participant, both measures could be judged against the administered dose of alcohol instead of against each other. That dose is a change of known size and direction, which is exactly what measuring responsiveness requires [13].

The digital score’s advantage over the BBS item rating was not uniform across tasks. It was largest in the three tasks where the BBS item rating sat at or near chance, reaching a gain of 0.33 to 0.40, and smallest in eyes-closed OLS, the one task in which the BBS item rating used its full range. Those tasks differ in how much of the BBS item rating was in use. In our cohort 77% of task ratings were the maximum score of 4, and in eyes-open TLS every rating at every dose was a 4, which fixes discrimination at chance by construction.

The clinical relevance of this ceiling lies in whom it affects. The BBS saturates not in the severely impaired but in those who function well: in independently living older adults with above-average function [60], in patients with balance disorders, where the Mini Balance Evaluation Systems Test (Mini-BESTest) is the more responsive instrument [16], and at discharge from subacute stroke rehabilitation, where 43% of patients score the maximum [61]. Those reports concern the 14-item total rather than the per-task ratings used here. This is the range in which a responsive measure would be most useful, because it is where recovery is tracked, where discharge decisions are made, and where a constant score is most easily read as a stable balance. It is also the range in which responsiveness is hardest to establish, since the comparator against which a candidate measure would be validated is the one that has stopped moving.

Acute alcohol intake resolves that circularity. It produces a graded and reversible disturbance of postural control of comparable severity to the impairment in which the scale saturates, acting on the cerebellar and vestibular systems that are also disrupted in clinical balance disorders [24, 25], and the size and direction of the change are set by the administered dose rather than inferred from a scale. The healthy cohort studied here is therefore not a substitute for patients. It is the setting in which the change can be known exactly, and in which the behavior of the rating can be observed against it. What this demonstrates is a property of the instrument, that a maximum score bounds impairment without excluding it, and that property is carried by the scale rather than by the population.

The clearest evidence that the digital metrics capture what the BBS item rating misses comes from the participants it judged to be unaffected. In those awarded the maximum score both when sober and at 1.0 ‰, every one of the nine testable digital comparisons showed a significant sober-to-high-dose change. In the stricter subset rated at ceiling at all three doses the response was monotonic across all nine, so what the constant rating concealed was not a single step but a graded change tracking the dose. Because this held for the second rater as well, the limitation lies in the BBS item rating scale rather than in one observer’s judgment. A maximum rating therefore bounds the impairment without excluding it.

The reliability of the BBS item rating followed the same pattern. Agreement between the two raters was high overall, 83% exact and 92% within one point, with a quadratically weighted κ of 0.76. Most of that agreement is trivial. Ratings in which

629 both raters awarded a 4 cannot disagree, so they contribute agreement without con-
630 tributing discrimination. Once those are removed, exact agreement falls to 30% and no
631 longer varies systematically across tasks. Agreement therefore looked better on some
632 tasks than others only because those tasks had more ratings at ceiling, not because the
633 raters judged them better. On the ratings that required a real decision, they agreed
634 less than a third of the time. This level of agreement concerns single-item ratings and
635 is not comparable with reliability values published for the 14-item BBS total. It does,
636 however, bear on how the scale behaves whenever a single task is scored by eye.

637 Those metrics still track what the BBS rates. A measure can out-resolve a compara-
638 tor by capturing something else entirely, which would make the gap uninformative,
639 but wherever the rating retained enough variance to correlate, the two moved together
640 in the expected direction, with more sway corresponding to a lower rating. The digital
641 metrics add resolution within the same construct. The correlations were modest, and
642 they were weakest in the tasks where the rating was least variable, which is the ceil-
643 ing appearing for a third time after the ordering accuracies and the rater agreement.
644 In eyes-open TLS it could not be computed. This is the difficulty with validating any
645 new balance measure against the BBS in people who function well. Agreement can
646 only be shown over the range in which the comparator still moves, and that range nar-
647 rows as function rises, so the evidence thins exactly where a more sensitive measure
648 would be most useful.

649 The Mini-BESTest is more sensitive than the BBS in people with milder impair-
650 ment [16, 60], largely because it includes more demanding items. Such a scale could
651 recover some of the resolution lost here. But it would not remove the problem that
652 every bounded ordinal scale saturates somewhere, including the Mini-BESTest [60].
653 A wider effective range moves the ceiling to a higher level of function. It does not
654 eliminate it. Ordinal observational scales carry no fixed metric value [15], depend on
655 a rater’s judgment, and record how well a task was performed without describing the
656 movement that produced it. The digital metrics used here address each of those limi-
657 tations. They are continuous, expressed in physical units, computed without a rater,
658 and they describe the movement itself rather than a judgment about its outcome.
659 Together these properties are what make a camera-based measure a candidate for
660 responsive balance assessment in rehabilitation: it registers change where a bounded
661 ordinal scale cannot, and it does so without adding equipment to the clinical setting.

662 That leaves the question of which movements a camera should be asked to mea-
663 sure. The answer depends on the impairment, not on the method. Here the selection
664 converged on mediolateral sway velocity. Alcohol acts on cerebellar and vestibular
665 control [24, 25], and the resulting instability is dominated by continuous sway, not by
666 discrete events, so the metrics that ranked highest here are the ones that capture it.
667 The candidate set was not restricted to sway. It spanned CoM and trunk kinemat-
668 ics, joint angles, stepping, upper-limb motion and Romberg ratios, and 532 of the 894
669 task-metric combinations responded to the dose gradient, in five of those six families.
670 Upper-limb metrics were responsive but were pruned as redundant with trunk and
671 CoM sway, and that redundancy reflects how alcohol destabilizes posture rather than
672 a fixed relationship between arm and trunk motion. The selection procedure does not
673 prespecify which metrics matter, so an impairment of another kind would surface a
674 different subset of the same candidate set.

These patterns are consistent with what instrumented studies of alcohol have long reported. Force platforms and marker-based motion capture have characterized alcohol’s effects on standing balance for decades. The Romberg quotient is unchanged across a wide range of BACs [24], as our eyes-open to eyes-closed ratios also showed. Three-dimensional motion analysis of five body segments at the same two doses used here found the destabilization substantially larger mediolaterally than anteroposteriorly [26], and mediolateral sway velocity is what our selection converged on. An instrumented balance platform detected elevated sway at 0.8 ‰ but not at 0.4 ‰ BAC [62], and our 0.6 ‰ falls between those two doses. Ordering was most accurate for the sober-versus-high-dose pair and least accurate for sober-versus 0.6 ‰, which follows the size of the underlying change, since stability declines far more steeply between 0.6 and 1.0 ‰ than between 0.0 and 0.6 ‰ [63]. Markerless capture has already been shown to measure balance accurately [64]. Simpler video methods reach the same conclusion [65]. These studies establish that a camera can measure balance. They do not ask whether it resolves change a clinical rating misses.

What has not been done is to set a camera-derived measure against a clinical scale on the same trials, with a graded within-person change as the reference. Camera-based balance work has been cross-sectional, benchmarking pose-derived features against a clinical index in one session [66], and reviews of computer vision in movement disorders report the same pattern, with most studies addressing gait instead of quiet stance [67]. Sensor-based studies of cerebellar disease use a nearly identical set of stance conditions and reach comparable discrimination, but they separate patients from controls instead of following graded change within a person [68]. The scale is either the reference standard or absent from the comparison.

In a clinical setting the graded change supplied here by the dose could instead be a course of rehabilitation, or change over time. Extending the approach to the remaining BBS items, which involve transfers, turning and reaching rather than quiet stance, would also call for a different candidate set, but that is an extension of the procedure rather than a redesign. The approach does not have to cover the whole scale to be useful. The clinical value of a responsive measure lies where the rating does not move. A patient whose BBS item score is unchanged between two assessments may have improved, deteriorated, or stayed the same, and the score cannot distinguish these. That is the situation in which decisions about continuing, adapting or ending therapy are made, and it is the situation this study reproduces: in the participants rated at the maximum at every dose, the digital metrics still changed monotonically with the perturbation. A continuous measure does not replace the clinical judgment the scale encodes, but it can indicate change the scale cannot register.

What such a measure would cost to obtain is modest. Recording required a single consumer camera on a tripod, positioned in front of the participant, with no markers, no force platform, no wearable sensors and no rater. The camera was not calibrated and its position was not standardized between participants, since expressing kinematics as a fraction of body height makes them invariant to camera distance and image resolution. The recordings analyzed here were made for visual scoring rather than for kinematic analysis, so the approach can be applied to video already captured in routine practice. The analysis is not yet lightweight: Sapiens-1B is a one-billion-parameter

720 model and does not run in real time on ordinary hardware. This is a property of the
721 model rather than of the approach, and current pose estimators reach real-time inference
722 on a laptop Central Processing Unit (CPU) or a mobile device at comparable
723 accuracy [69], so nothing in the pipeline requires specialized computing infrastructure
724 in principle.

725 Which metrics would matter in a given patient population is not settled by this
726 study. Alcohol disturbs cerebellar and vestibular control, and the selection converged
727 on mediolateral sway velocity accordingly. The closest clinical parallels are therefore
728 the cerebellar ataxias and vestibular disorders, where digital sway measures already
729 discriminate patients from controls [9, 68]. In conditions whose postural signature
730 differs, such as the asymmetric loading after stroke or the freezing and festination of
731 PD, a different subset of the same candidate set would be expected to emerge. The
732 procedure rather than the resulting metric is what transfers.

733 Limitations

734 The cohort was 24 healthy young adults, and alcohol at 0.6 and 1.0 ‰ produces
735 impairment at the mild end of what a balance scale is built to measure. The advantage
736 shown here is therefore an advantage in the regime where the BBS sits at or
737 near its ceiling. It says nothing about how the BBS performs where its range is in
738 use, and the responsiveness reported for the BBS in progressive ataxia suggests it performs
739 well there [70]. Alcohol is also a particular model of impairment, acting acutely
740 and reversibly without the compensation, adaptation and fluctuation that accompany
741 chronic pathology.

742 Dose was administered in ascending order within a single session, so time on task,
743 fatigue and repeated practice all covary with BAC and cannot be separated from it.
744 The two act in opposite directions: practice on a repeated stance task would tend
745 to improve performance as dose rose, working against the observed decline, whereas
746 fatigue would add to it.

747 The camera recorded the frontal plane only, which suits a perturbation that destabilizes
748 the body mediolaterally more than anteroposteriorly [26], but an impairment acting
749 mainly in the sagittal plane would require a second view. The metrics were
750 selected from this cohort. Selection was repeated inside each cross-validation fold,
751 so the reported accuracies are not inflated by it, but it is the underlying constructs
752 rather than the exact measures that should be expected to replicate. Finally, the end-
753 point throughout was the ordering of doses within a participant, so what a given sway
754 velocity means for a person seen once remains an open question.

755 Conclusions

756 Under a graded and reversible alcohol intake, continuous metrics derived from a single
757 consumer camera ordered each participant’s blood alcohol levels more accurately than
758 a concurrent BBS item rating of the same trials, and did so on all six tasks. The
759 gap was largest where the rating sat at its ceiling, and impairment was measurable
760 even in the participants whose rating did not change at all. A rating at the maximum
761 records that performance fell inside the top category, not that balance was unimpaired.

762 Because the selection procedure is not told in advance which metrics matter, the
763 same approach applies wherever a change of known size and direction can serve as the
764 reference, and the recording it requires is no more than a camera.

765 Abbreviations

766	<i>AUC</i>	area under the receiver-operating-characteristic curve
767	<i>BAC</i>	blood alcohol concentration
768	<i>BBS</i>	Berg Balance Scale
769	<i>BMI</i>	body-mass index
770	<i>CI</i>	confidence interval
771	<i>CoM</i>	center of mass
772	<i>CPU</i>	Central Processing Unit
773	<i>FDR</i>	false discovery rate
774	<i>fps</i>	frames per second
775	<i>IMU</i>	inertial measurement unit
776	<i>LOSO</i>	leave-one-subject-out
777	<i>LTP</i>	lower-trunk proxy
778	<i>Mini-BESTest</i>	Mini Balance Evaluation Systems Test
779	<i>OLS</i>	one-legged stance
780	<i>OMC</i>	optical motion capture
781	<i>PD</i>	Parkinson's disease
782	<i>RGB</i>	red-green-blue
783	<i>RGB-D</i>	red-green-blue-depth
784	<i>SD</i>	standard deviation
785	<i>SPARC</i>	spectral arc length
786	<i>TLS</i>	two-legged stance

787 Declarations

788 Ethics approval and consent to participate

789 The study was approved by Kantonale Ethikkommission Zürich KEK with number
790 2021-00602 and all participants provided written informed consent.

791 Consent for publication

792 All participants provided written informed consent for the publication of their
793 anonymized data and results.

794 Data availability

795 The datasets generated and/or analyzed during the current study are not publicly
796 available due to data protection reasons but are available from the corresponding
797 author on reasonable request.

798 Competing interests

799 O.L. is a member of the Editorial Board of the Journal of NeuroEngineering and
800 Rehabilitation. He was not involved in the editorial handling or peer review of this
801 manuscript. All other authors declare no competing interests.

802 **Funding**

803 This study did not receive funding.

804 **Authors' contributions**

805 Study design: O.L., A.A.T., N.D.. Data collection: N.D.. Participant recruitment:
806 N.D.. Data analysis: N.D., N.L., S.D.. Data interpretation: N.D., N.L., T.U., O.L..
807 Manuscript writing: N.D., T.U.. Manuscript review: N.D., T.U., N.L., O.L., A.A.T..
808 All authors read and approved the final manuscript.

809 **Acknowledgements**

810 We thank Anna Knill and Christoph Kanzler for their contribution to study prepara-
811 tion and ethics documentation, and Stefan Fuchs and Danja Conconi for their support
812 with data collection. We further thank all participants for their time and valuable
813 contribution.

References

- [1] Horak, F.B.: Postural orientation and equilibrium: what do we need to know about neural control of balance to prevent falls? *Age and Ageing* **35**(suppl_2), 7–11 (2006) <https://doi.org/10.1093/ageing/afl077> . Accessed 2026-06-26
- [2] Pollock, A., Baer, G., Campbell, P., Choo, P.L., Forster, A., Morris, J., Pomeroy, V.M., Langhorne, P.: Physical rehabilitation approaches for the recovery of function and mobility following stroke. *The Cochrane Database of Systematic Reviews* **2014**(4), 001920 (2014) <https://doi.org/10.1002/14651858.CD001920.pub3>
- [3] Horak, F.B., Macpherson, J.M.: Postural Orientation and Equilibrium. *Comprehensive Physiology* **1995**(4S29), 255–292 (1995) <https://doi.org/10.1002/j.2040-4603.1995.tb01704.x> . eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2040-4603.1995.tb01704.x>. Accessed 2026-06-26
- [4] Peterka, R.J.: Sensorimotor Integration in Human Postural Control. *Journal of Neurophysiology* **88**(3), 1097–1118 (2002) <https://doi.org/10.1152/jn.2002.88.3.1097> . Accessed 2026-06-26
- [5] Geurts, A.C.H., Haart, M., Nes, I.J.W., Duysens, J.: A review of standing balance recovery from stroke. *Gait & Posture* **22**(3), 267–281 (2005) <https://doi.org/10.1016/j.gaitpost.2004.10.002>
- [6] Haart, M., Geurts, A.C., Huidekoper, S.C., Fasotti, L., Limbeek, J.: Recovery of standing balance in postacute stroke patients: a rehabilitation cohort study. *Archives of Physical Medicine and Rehabilitation* **85**(6), 886–895 (2004) <https://doi.org/10.1016/j.apmr.2003.05.012>
- [7] Mancini, M., Carlson-Kuhta, P., Zampieri, C., Nutt, J.G., Chiari, L., Horak, F.B.: Postural sway as a marker of progression in Parkinson’s disease: a pilot longitudinal study. *Gait & Posture* **36**(3), 471–476 (2012) <https://doi.org/10.1016/j.gaitpost.2012.04.010>
- [8] Horak, F.B., Mancini, M.: Objective Biomarkers of Balance and Gait for Parkinson’s Disease using Body-worn Sensors. *Movement disorders : official journal of the Movement Disorder Society* **28**(11), 1544–1551 (2013) <https://doi.org/10.1002/mds.25684> . Accessed 2026-06-26
- [9] Velázquez-Pérez, L., Rodríguez-Labrada, R., González-Garcés, Y., Arrufat-Pie, E., Torres-Vega, R., Medrano-Montero, J., Ramirez-Bautista, B., Vazquez-Mojena, Y., Auburger, G., Horak, F., Ziemann, U., Gomez, C.M.: Prodromal Spinocerebellar Ataxia Type 2 Subjects Have Quantifiable Gait and Postural Sway Deficits. *Movement Disorders: Official Journal of the Movement Disorder Society* **36**(2), 471–480 (2021) <https://doi.org/10.1002/mds.28343>

- [10] Zhou, H., Nguyen, H., Enriquez, A., Morsy, L., Curtis, M., Piser, T., Kenney, C., Stephen, C.D., Gupta, A.S., Schmähmann, J.D., Vaziri, A.: Assessment of gait and balance impairment in people with spinocerebellar ataxia using wearable sensors. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* **43**(4), 2589–2599 (2022) <https://doi.org/10.1007/s10072-021-05657-6>
- [11] Berg, K.O., Wood-Dauphinee, S.L., Williams, J.I., Maki, B.: Measuring balance in the elderly: validation of an instrument. *Canadian Journal of Public Health = Revue Canadienne De Sante Publique* **83 Suppl 2**, 7–11 (1992)
- [12] Lanska, D.J.: The Romberg Sign and Early Instruments for Measuring Postural Sway. *Seminars in Neurology* **22**(04), 409–418 (2002) <https://doi.org/10.1055/s-2002-36763> . Accessed 2026-06-26
- [13] Guyatt, G., Walter, S., Norman, G.: Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases* **40**(2), 171–178 (1987) [https://doi.org/10.1016/0021-9681\(87\)90069-5](https://doi.org/10.1016/0021-9681(87)90069-5)
- [14] Pardasaney, P.K., Latham, N.K., Jette, A.M., Wagenaar, R.C., Ni, P., Slavin, M.D., Bean, J.F.: Sensitivity to change and responsiveness of four balance measures for community-dwelling older adults. *Physical Therapy* **92**, 388–397 (2012) <https://doi.org/10.2522/ptj.20100398>
- [15] La Porta, F., Caselli, S., Susassi, S., Cavallini, P., Tennant, A., Franceschini, M.: Is the Berg Balance Scale an internally valid and reliable measure of balance across different etiologies in neurorehabilitation? A revisited Rasch analysis study. *Archives of Physical Medicine and Rehabilitation* **93**(7), 1209–1216 (2012) <https://doi.org/10.1016/j.apmr.2012.02.020>
- [16] Godi, M., Franchignoni, F., Caligari, M., Giordano, A., Turcato, A.M., Nardone, A.: Comparison of reliability, validity, and responsiveness of the mini-BESTest and Berg Balance Scale in patients with balance disorders. *Physical Therapy* **93**(2), 158–167 (2013) <https://doi.org/10.2522/ptj.20120171>
- [17] Unger, T., Sal Moslehian, A., Peiffer, J.D., Ullrich, J., Gassert, R., Lamercy, O., James Cotton, R., Awai Easthope, C.: Differentiable Biomechanics for Markerless Motion Capture in Upper Limb Stroke Rehabilitation: A Comparison With Optical Motion Capture. *IEEE Transactions on Medical Robotics and Bionics* **8**(1), 90–97 (2026) <https://doi.org/10.1109/TMRB.2025.3605962> . Accessed 2026-06-26
- [18] Unger, T., Yao, X., Schmitt, A., Cebulla, L., Awai, C.E.: Towards Routine Biomechanical Data Collection in Neurorehabilitation: A Usability Comparison of IMU and Markerless Motion Capture Systems for Upper-Limb Assessments. *Research Square*. ISSN: 2693-5015 (2026). <https://doi.org/10.21203/rs.3.rs-8711156/v1> . <https://www.researchsquare.com/article/rs-8711156/v1> Accessed 2026-06-26

- [19] Unger, T., Sousa Ribeiro, R., Mokni, M., Weikert, T., Pohl, J., Schwarz, A., Held, J.P.O., Sauerzopf, L., Kühnis, B., Gavagnin, E., Luft, A.R., Gassert, R., Lambercy, O., Awai Easthope, C., Schönhammer, J.G.: Upper limb movement quality measures: comparing IMUs and optical motion capture in stroke patients performing a drinking task. *Frontiers in Digital Health* **6** (2024) <https://doi.org/10.3389/fdgth.2024.1359776> . Accessed 2026-06-26
- [20] Clark, R.A., Pua, Y.-H., Fortin, K., Ritchie, C., Webster, K.E., Denehy, L., Bryant, A.L.: Validity of the Microsoft Kinect for assessment of postural control. *Gait & Posture* **36**(3), 372–377 (2012) <https://doi.org/10.1016/j.gaitpost.2012.03.033>
- [21] Federolf, P., Kühne, M., Schiel, K., Reimeir, E., Debertin, D., Calisti, M., Mohr, M.: Validation of markerless (Theia3D™) against marker-based (Vicon™) motion capture data of postural control movements analyzed through principal component analysis. *Journal of Biomechanics* **189**, 112831 (2025) <https://doi.org/10.1016/j.jbiomech.2025.112831>
- [22] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M.: MediaPipe: A Framework for Building Perception Pipelines. *arXiv. arXiv:1906.08172 [cs.DC]* (2019). <https://doi.org/10.48550/arXiv.1906.08172> . <http://arxiv.org/abs/1906.08172> Accessed 2026-06-26
- [23] Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569* (2024)
- [24] Nieschalk, M., Ortmann, C., West, A., Schmä, F., Stoll, W., Fechner, G.: Effects of alcohol on body-sway patterns in human subjects. *International Journal of Legal Medicine* **112**, 253–260 (1999) <https://doi.org/10.1007/s004140050245>
- [25] Diener, H.C., Dichgans, J., Bacher, M., Hülser, J., Liebich, H.: Mechanisms of postural ataxia after intake of alcohol. *Zeitschrift Fur Rechtsmedizin. Journal of Legal Medicine* **90**(3), 159–165 (1983) <https://doi.org/10.1007/BF02116226>
- [26] Modig, F., Fransson, P.-A., Magnusson, M., Patel, M.: Blood alcohol concentration at 0.06 and 0.10% causes a complex multifaceted deterioration of body movement control. *Alcohol* **46**, 75–88 (2012) <https://doi.org/10.1016/j.alcohol.2011.06.001>
- [27] Goebel, J.A., Dunham, D.N., Rohrbaugh, J.W., Fischel, D., Stewart, P.A.: Dose-related effects of alcohol on dynamic posturography and oculomotor measures. *Acta Oto-Laryngologica. Supplementum* **520 Pt 1**, 212–215 (1995) <https://doi.org/10.3109/00016489509125232>
- [28] Ledin, T., Odkvist, L.M.: Effect of alcohol measured by dynamic posturography.

- 928 Acta Oto-Laryngologica. Supplementum **481**, 576–581 (1991) [https://doi.org/](https://doi.org/10.3109/00016489109131475)
929 10.3109/00016489109131475
- 930 [29] Kubo, T., Sakata, Y., Koshimune, A., Sakai, S., Ameno, K., Ijiri, I.: Posi-
931 tional nystagmus and body sway after alcohol ingestion. American Journal of
932 Otolaryngology **11**(6), 416–419 (1990) [https://doi.org/10.1016/0196-0709\(90\)](https://doi.org/10.1016/0196-0709(90)90121-b)
933 90121-b
- 934 [30] Romaniszyn, P., Kawa, J., Stepień, P., Nawrat-Szołtysik, A.: Video-based time
935 assessment in 360 degrees turn berg balance test. Computerized Medical Imaging
936 and Graphics **80**, 101689 (2020) [https://doi.org/10.1016/j.compmedimag.2019.](https://doi.org/10.1016/j.compmedimag.2019.101689)
937 101689
- 938 [31] Cho, J., Ha, S., Lee, J., Kim, M., Kim, H.: Standing balance test predicts the berg
939 balance scale score in patients with stroke using principal component analysis.
940 Scientific Reports **15**, 17653 (2025) <https://doi.org/10.1038/s41598-025-99710-5>
- 941 [32] Lu, Z., Zhou, H., Lyu, H., Wu, H., Tian, S., Yang, G.: Berg balance scale scoring
942 system for balance evaluation by leveraging attention-based deep learning with
943 wearable imu sensors. Bioengineering **12**, 395 (2025) [https://doi.org/10.3390/](https://doi.org/10.3390/bioengineering12040395)
944 bioengineering12040395
- 945 [33] Widmark, E.M.P.: Die Theoretischen Grundlagen und die Praktische Verwend-
946 barkeit der Gerichtlich-medizinischen Alkoholbestimmung. Urban & Schwarzen-
947 berg, Berlin (1932)
- 948 [34] Posey, D., Mozayani, A.: The estimation of blood alcohol concentration : Wid-
949 mark revisited. Forensic science, medicine, and pathology **3**, 33–9 (2007) <https://doi.org/10.1385/FSMP:3:1:33>
- 950
- 951 [35] Holland, M.G., Ferner, R.E.: A systematic review of the evidence for acute tol-
952 erance to alcohol – the “mellanby effect”. Clinical Toxicology **55**, 545–556 (2017)
953 <https://doi.org/10.1080/15563650.2017.1296576>
- 954 [36] Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen,
955 K.: RtmDet: An empirical study of designing real-time object detectors (2022)
- 956 [37] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár,
957 P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context, pp. 740–755.
958 Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
- 959 [38] Winter, D.A.: Biomechanics and Motor Control of Human Movement. Wiley,
960 Hoboken (2009). <https://doi.org/10.1002/9780470549148>
- 961 [39] Prieto, T.E., Myklebust, J.B., Hoffmann, R.G., Lovett, E.G., Myklebust, B.M.:
962 Measures of postural steadiness: differences between healthy young and elderly
963 adults. IEEE Transactions on Biomedical Engineering **43**, 956–966 (1996) <https://doi.org/10.1109/10.491301>

- 964 [//doi.org/10.1109/10.532130](https://doi.org/10.1109/10.532130)
- 965 [40] Schubert, P., Kirchner, M.: Ellipse area calculations and their applicability in
966 posturography. *Gait & Posture* **39**, 518–522 (2014) [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.gaitpost.2013.09.001)
967 [gaitpost.2013.09.001](https://doi.org/10.1016/j.gaitpost.2013.09.001)
- 968 [41] Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A., Burdet, E.:
969 On the analysis of movement smoothness. *Journal of NeuroEngineering and*
970 *Rehabilitation* **12**, 112 (2015) <https://doi.org/10.1186/s12984-015-0090-9>
- 971 [42] Springer, B.A., Marin, R., Cyhan, T., Roberts, H., Gill, N.W.: Normative
972 values for the unipedal stance test with eyes open and closed. *Journal of Geriatric Physical Therapy* **30**, 8–15 (2007) [https://doi.org/10.1519/](https://doi.org/10.1519/00139143-200704000-00003)
973 [00139143-200704000-00003](https://doi.org/10.1519/00139143-200704000-00003)
- 974
- 975 [43] Maki, B.E., McIlroy, W.E.: The role of limb movements in maintaining upright
976 stance: The “change-in-support” strategy. *Physical Therapy* **77**, 488–507 (1997)
977 <https://doi.org/10.1093/ptj/77.5.488>
- 978 [44] Emmerik, R.E.A.V., Rosenstein, M.T., McDermott, W.J., Hamill, J.: A nonlinear
979 dynamics approach to human movement. *Journal of Applied Biomechanics* **20**,
980 396–420 (2004) <https://doi.org/10.1123/jab.20.4.396>
- 981 [45] Byrne, J.E., Stergiou, N., Blanke, D., Houser, J.J., Kurz, M.J., Hageman, P.A.:
982 Comparison of gait patterns between young and elderly women: An examination
983 of coordination. *Perceptual and Motor Skills* **94**, 265–280 (2002) [https://doi.org/](https://doi.org/10.2466/pms.2002.94.1.265)
984 [10.2466/pms.2002.94.1.265](https://doi.org/10.2466/pms.2002.94.1.265)
- 985 [46] Mansfield, A., Peters, A.L., Liu, B.A., Maki, B.E.: Effect of a perturbation-based
986 balance training program on compensatory stepping and grasping reactions in
987 older adults: A randomized controlled trial. *Physical Therapy* **90**, 476–491 (2010)
988 <https://doi.org/10.2522/ptj.20090070>
- 989 [47] Gibbons, C.T., Amazeen, P.G., Likens, A.D.: Effects of foot placement on pos-
990 tural sway in the anteroposterior and mediolateral directions. *Motor Control* **23**,
991 149–170 (2019) <https://doi.org/10.1123/mc.2017-0074>
- 992 [48] McKnight, A.J., Langston, E.A., McKnight, A.S., Lange, J.E.: Sobriety tests for
993 low blood alcohol concentrations. *Accident Analysis & Prevention* **34**, 305–311
994 (2002) [https://doi.org/10.1016/S0001-4575\(01\)00027-6](https://doi.org/10.1016/S0001-4575(01)00027-6)
- 995 [49] Lacour, M., Barthelemy, J., Borel, L., Magnan, J., Xerri, C., Chays, A., Ouak-
996 nine, M.: Sensory strategies in human postural control before and after unilateral
997 vestibular neurotomy. *Experimental Brain Research* **115**, 300–310 (1997) [https:](https://doi.org/10.1007/PL00005698)
998 [//doi.org/10.1007/PL00005698](https://doi.org/10.1007/PL00005698)

- 999 [50] Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approx-
1000 imate entropy and sample entropy. *American Journal of Physiology-Heart and*
1001 *Circulatory Physiology* **278**, 2039–2049 (2000) [https://doi.org/10.1152/ajpheart.](https://doi.org/10.1152/ajpheart.2000.278.6.H2039)
1002 2000.278.6.H2039
- 1003 [51] Webber, C.L., Zbilut, J.P.: Dynamical assessment of physiological systems and
1004 states using recurrence plot strategies. *Journal of applied physiology* (Bethesda,
1005 Md. : 1985) **76**, 965–73 (1994) <https://doi.org/10.1152/jappl.1994.76.2.965>
- 1006 [52] Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., Kurths, J.: Recurrence-
1007 plot-based measures of complexity and their application to heart-rate-variability
1008 data. *Physical Review E* **66**, 026702 (2002) [https://doi.org/10.1103/PhysRevE.](https://doi.org/10.1103/PhysRevE.66.026702)
1009 66.026702
- 1010 [53] Riley, M.A., Clark, S.: Recurrence analysis of human postural sway during the
1011 sensory organization test. *Neuroscience Letters* **342**, 45–48 (2003) [https://doi.](https://doi.org/10.1016/S0304-3940(03)00229-5)
1012 [org/10.1016/S0304-3940\(03\)00229-5](https://doi.org/10.1016/S0304-3940(03)00229-5)
- 1013 [54] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and
1014 powerful approach to multiple testing. *Journal of the Royal Statistical Society*
1015 *Series B: Statistical Methodology* **57**, 289–300 (1995) [https://doi.org/10.1111/j.](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)
1016 2517-6161.1995.tb02031.x
- 1017 [55] Page, E.B.: Ordered hypotheses for multiple treatments: A significance test for
1018 linear ranks. *Journal of the American Statistical Association* **58**, 216–230 (1963)
1019 <https://doi.org/10.1080/01621459.1963.10500843>
- 1020 [56] Meinshausen, N., Bühlmann, P.: Stability selection. *Journal of the Royal Sta-*
1021 *tistical Society Series B: Statistical Methodology* **72**, 417–473 (2010) [https:](https://doi.org/10.1111/j.1467-9868.2010.00740.x)
1022 [//doi.org/10.1111/j.1467-9868.2010.00740.x](https://doi.org/10.1111/j.1467-9868.2010.00740.x)
- 1023 [57] Nogueira, S., Brown, G.: Measuring the Stability of Feature Selection, pp. 442–
1024 457 (2016). https://doi.org/10.1007/978-3-319-46227-1_28
- 1025 [58] McCullagh, P.: Regression models for ordinal data. *Journal of the Royal Sta-*
1026 *tistical Society Series B: Statistical Methodology* **42**, 109–127 (1980) [https:](https://doi.org/10.1111/j.2517-6161.1980.tb01109.x)
1027 [//doi.org/10.1111/j.2517-6161.1980.tb01109.x](https://doi.org/10.1111/j.2517-6161.1980.tb01109.x)
- 1028 [59] Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled
1029 disagreement or partial credit. *Psychological bulletin* **70**, 213–20 (1968) [https:](https://doi.org/10.1037/h0026256)
1030 [//doi.org/10.1037/h0026256](https://doi.org/10.1037/h0026256)
- 1031 [60] Ban, B., Sevšek, F., Rugelj, D.: A comparison of the ceiling effect between berg
1032 balance scale and mini-bestest in a group of balance trained community-dwelling
1033 older adults. *Physiotherapy Quarterly* **25**, 3–9 (2018) [https://doi.org/10.5114/](https://doi.org/10.5114/pq.2018.73368)
1034 [pq.2018.73368](https://doi.org/10.5114/pq.2018.73368)

- [61] Inoue, S., Takagi, H., Tan, E., Oyama, C., Otaka, E., Kondo, K., Otaka, Y.: Comparison of usefulness between the Mini-Balance Evaluation Systems Test and the Berg Balance Scale for measuring balance in patients with subacute stroke. *Frontiers in Rehabilitation Sciences* **4**, 1308706 (2024) <https://doi.org/10.3389/fre.2023.1308706>
- [62] Marcziński, C.A., Mearns, C.L.: Automated assessment of alcohol-induced impairment of balance in male and female social drinkers. *Experimental and Clinical Psychopharmacology* **28**, 131–136 (2020) <https://doi.org/10.1037/pha0000309>
- [63] Modig, F., Patel, M., Magnusson, M., Fransson, P.-A.: Study I: Effects of 0.06% and 0.10% blood alcohol concentration on human postural control. *Gait & Posture* **35**(3), 410–418 (2012) <https://doi.org/10.1016/j.gaitpost.2011.10.364>
- [64] Chaumeil, A., Lahkar, B.K., Dumas, R., Muller, A., Robert, T.: Agreement between a markerless and a marker-based motion capture systems for balance related quantities. *Journal of Biomechanics* **165**, 112018 (2024) <https://doi.org/10.1016/j.jbiomech.2024.112018>
- [65] Ota, M., Tateuchi, H., Hashiguchi, T., Fujiwara, K., Sasaki, A., Okumura, K., Ichihashi, N.: Validity of the frame subtraction method in dynamic postural stability. *BMC Sports Science, Medicine and Rehabilitation* **14**, 174 (2022) <https://doi.org/10.1186/s13102-022-00570-7>
- [66] Lin, K.-C., Wai, R.-J., Chien, H.-Y.C.: A lightweight temporal convolutional network for contactless sppb-aligned functional fall-risk stratification in older adults using monocular rgb video. *Sensors* **26**, 3894 (2026) <https://doi.org/10.3390/s26123894>
- [67] Biase, L., Pecoraro, P.M., Bugamelli, F.: Ai video analysis in parkinson’s disease: A systematic review of the most accurate computer vision tools for diagnosis, symptom monitoring, and therapy management. *Sensors* **25**, 6373 (2025) <https://doi.org/10.3390/s25206373>
- [68] Shah, V.V., Muzyka, D., Jagodinsky, A., McNamara, J., Casey, H., El-Gohary, M., Sowalsky, K., Safarpour, D., Carlson-Kuhta, P., Schmähmann, J.D., Rosenthal, L.S., Perlman, S., Horak, F.B., Gomez, C.M.: Digital measures of postural sway quantify balance deficits in spinocerebellar ataxia. *Movement Disorders* **39**, 663–673 (2024) <https://doi.org/10.1002/mds.29742>
- [69] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RTM-Pose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399* (2023) <https://doi.org/10.48550/arXiv.2303.07399>
- [70] Milne, S.C., Kim, S.H., Murphy, A., Larkindale, J., Farmer, J., Malapira, R., Danoudis, M., Shaw, J., Ramakrishnan, T., Rasouli, F., Yiu, E.M.,

1073 Georgiou-Karistianis, N., Tai, G., Zesiewicz, T., Delatycki, M.B., Corben, L.A.:
1074 The responsiveness of gait and balance outcomes to disease progression in
1075 friedreich ataxia. *The Cerebellum* **21**, 963–975 (2021) [https://doi.org/10.1007/](https://doi.org/10.1007/s12311-021-01348-2)
1076 [s12311-021-01348-2](https://doi.org/10.1007/s12311-021-01348-2)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PoseEstimationBBSJNERSupplementary08092026V2.pdf](#)