Statistical Supplementary Material following the Bayesian Analysis Reporting Guidelines (BARG)

Harald Penasso[1,2]

[1] Saphenus Medical Technology GmbH, Hauptplatz 9–13, 2500 Baden, Austria

[2] Ludwig Boltzmann Institut für Experimentelle und Klinische Traumatologie, Donaueschingenstraße 13, 1200 Wien, Austria

Contents

1 BARG Steps

The statistics report follows the steps of the recently published Bayesian analysis reporting guidelines (BARG) [1].

1.0 Preamble

1.0.A Why Bayesian

Small sample sizes are a common challenge in studies that include specific populations, such as persons with unilateral transtibial amputation in non-belligerent countries. A small sample is likely to be non-normally distributed and to contain outliers even if randomly drawn from a normal distribution. Bayesian generalized linear mixed models can model non-normal data by simulating the underlying distribution through Markov chain Monte Carlo simulations and consider fixed (e.g., an intervention) and random effects (e.g., individuality). Bayesian methods do not rely on asymptotics, require careful selection of the prior information contained, are better suited for analyzing small sample sizes [2], and explain outcomes with intuitive probabilities. It allows the interpretation of robust, credible intervals of effect sizes against regions of practical equivalence (ROPE). Thus, Bayesian analysis provides intuitive tools for accepting or rejecting the null hypothesis while quantifying uncertainties.

1.0.B Goals of the analysis

I analyzed the data of this randomized controlled AB|BA cross-over (Figure 1) pilot study with regard to three goals:

- **On the global level**: Estimate the effect size (Hedge's g) for $\mathrm{sequence}$, $\mathrm{intervention}$, and period to quantify the effectiveness of focal vibration feedback and the study design, while controlling for the repeated measures design with random intercepts for each participant in interaction with each test outcome measure.
- **On the test level**: Estimate the effect size (Hedge's g) for each test and $\mathrm{intervention}$ level to quantify the sensitivity of the tests for assessing effects of focal vibration feedback, while controlling for random slopes of sequence and period with random intercepts for each person per test.
- **On the individual level**: Estimate the effect size (Hedge's g) for each ID and $\mathrm{intervention}$ level to differentiate between responders and non-responders, while controlling for random slopes of sequence and period with random intercepts for each person per test. The follow-up receiver-operator characteristic analysis will assess the ability of baseline tests to differentiate between responders and non-responders and will provide related threshold regions.

1.1 Explain the model

Each model estimates the dependent variable by considering fixed effects, random slopes, and random intercepts shown in Table 1 and Figures 2, 3, 4. Each effect was assessed by subtracting the test scores obtained after a period from the test scores obtained before a period. While Figures 5, 6, 8, 9, 10 show the average over three tries per session, 7 and 11 were assessed only once per session. However, all tries were kept (not averaged) for the analysis.

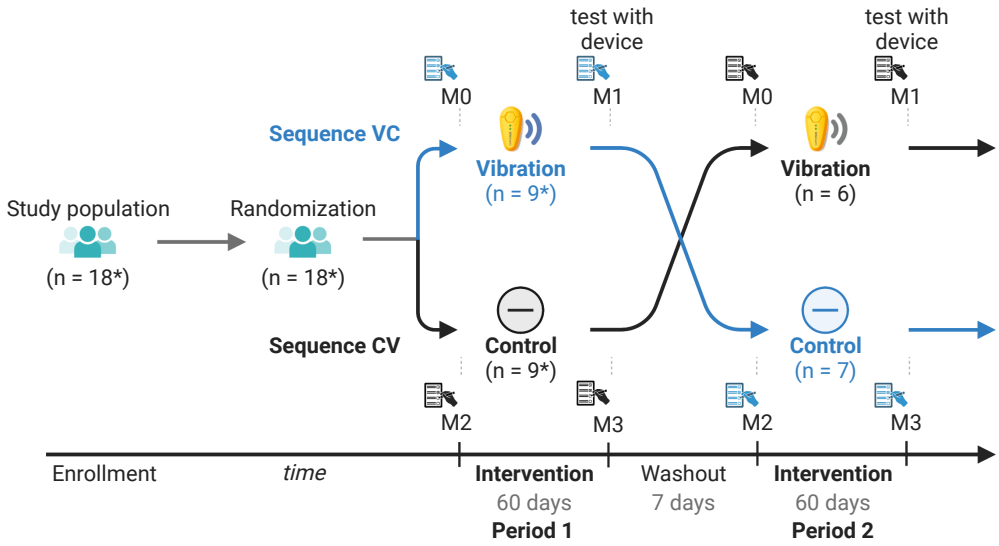

Figure 1: Randomized controlled cross-over design with sequences vibration → control (CV) and control → vibration (CV). In the two 60 day periods measurements were taken before vibration (M0), after vibration (M1), before control (M2), and after control (M3). * denotes the number of enrolled participants before two dropped out in the VC sequence and three in the CV sequence. We analyzed the remaining n = 13 participants at all periods. The participants did wear the device during M1 tests. Created with BioRender.com under paid subscription.

Table 1: Parameters of the Bayesian generalized linear mixed models. Graphical model illustrations are shown in Figures 2, 3, 4.

		fixed effects			random slopes		random
model	intercept	1	2	3	1	2	intercept
global*	0	intervention	sequence	period	1	-	ID : test
test	0	test : intervention	-	-	period	sequence	ID : test
individual	0	ID : intervention	-	-	period	sequence	ID : test

*Three global model submodels with differently ordered fixed effects gave the posterior distributions for all levels of all fixed effects, while not affecting any result. The operator : denotes an interaction between operands.

1.1.A Data Variables

The distributions of the outcome measure period changes mostly follow normal distributions except for outlines which can be handled by generalized linear mixed models (cf. right side Figures 5 to 11).

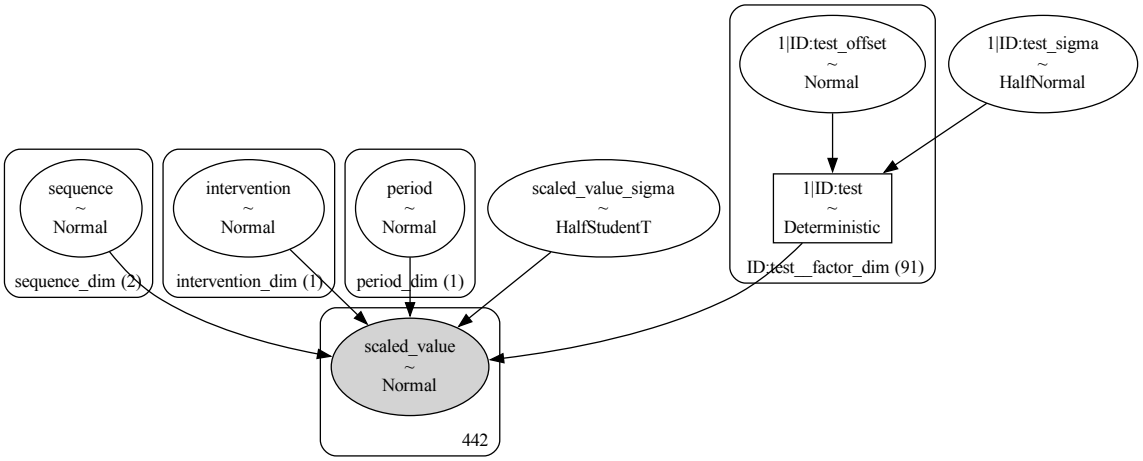

Figure 2: Global model: $\text{scaled\_score} \sim 0 + \text{sequence} + \text{intervention} + \text{period} + (1|\text{ID} : \text{test})$

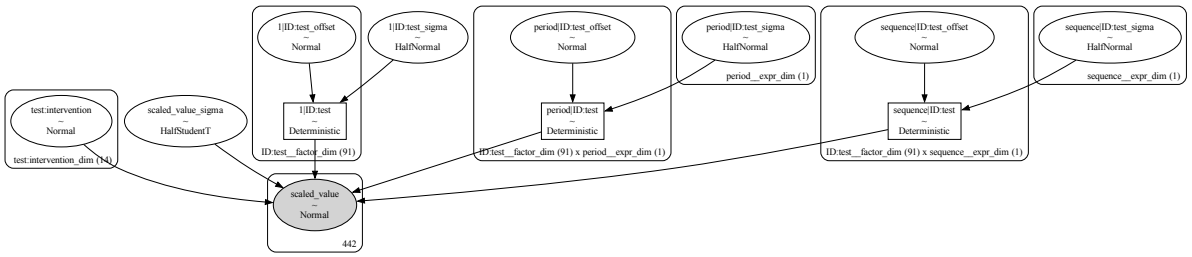

Figure 3: Test-level model: $\text{scaled\_score} \sim 0 + \text{test} : \text{intervention} + (\text{period} + \text{sequence}|\text{ID} : \text{test})$

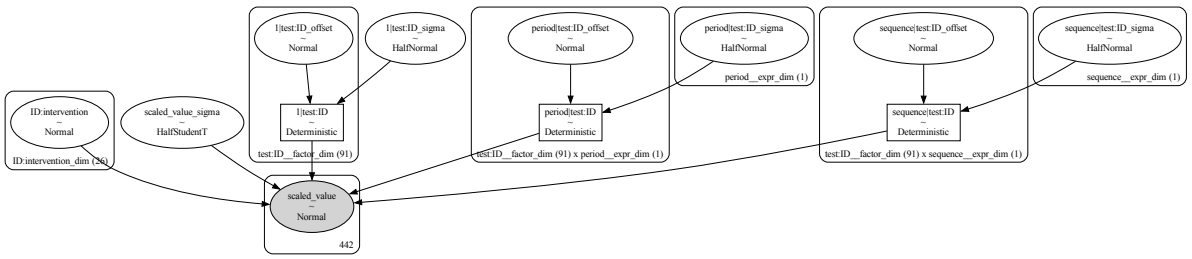

Figure 4: ID-level model: $\text{scaled\_score} \sim 0 + \text{ID} : \text{intervention} + (\text{period} + \text{sequence}|\text{ID} : \text{test})$

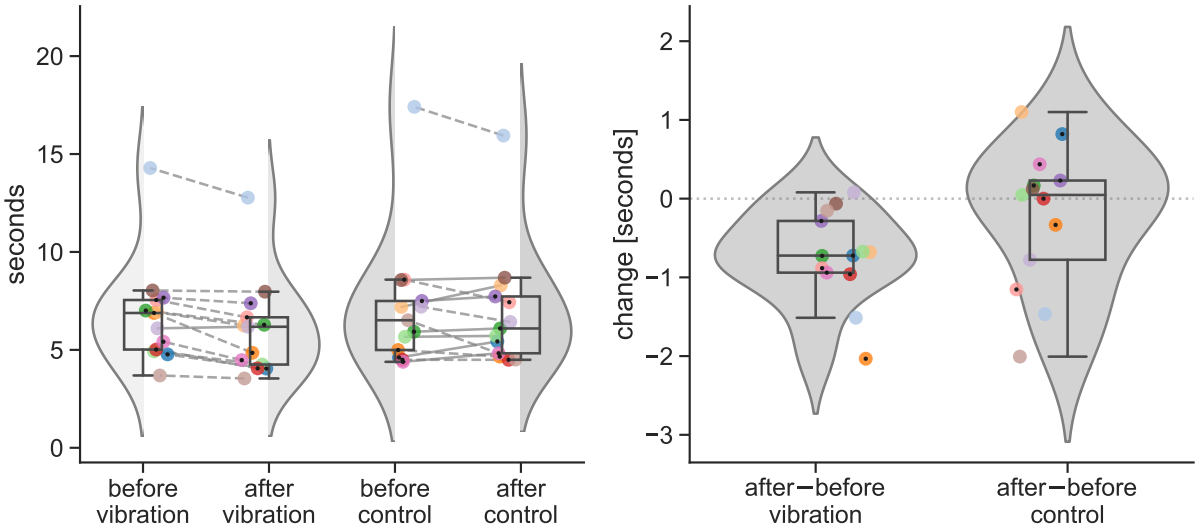

Figure 5: Illustration of the **four square step test** score changes. Marker colors encode participants. *left*: distribution, boxplot, individual scores averaged over tries: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes averaged over tries. Central black dots mark participants in the VC sequence (vibration fist, then control).

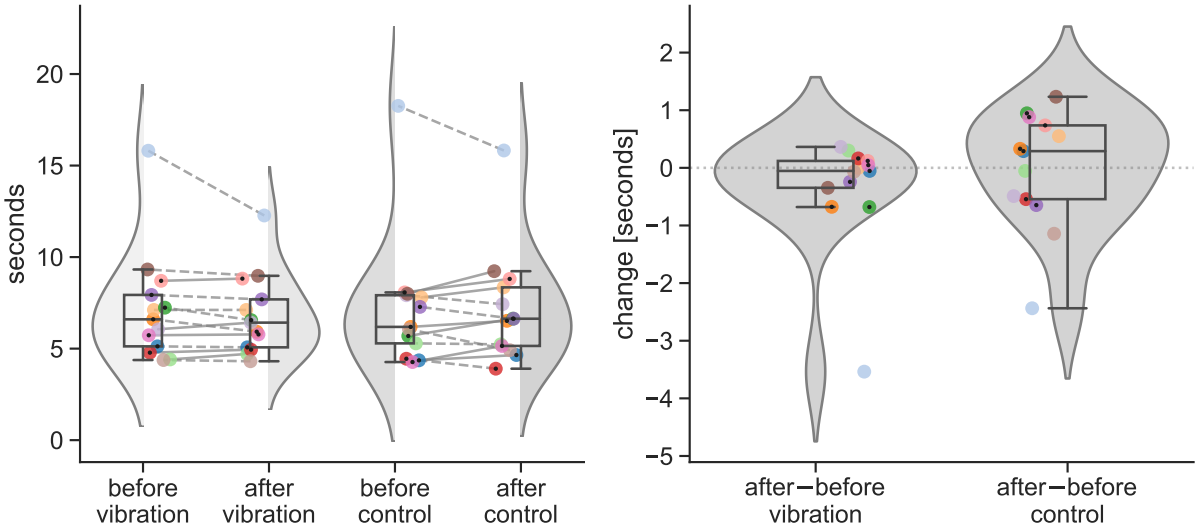

Figure 6: Illustration of the **timed up and go test** score changes. Marker colors encode participants. *left*: distribution, boxplot, individual scores averaged over tries: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes averaged over tries. Central black dots mark participants in the VC sequence (vibration fist, then control).

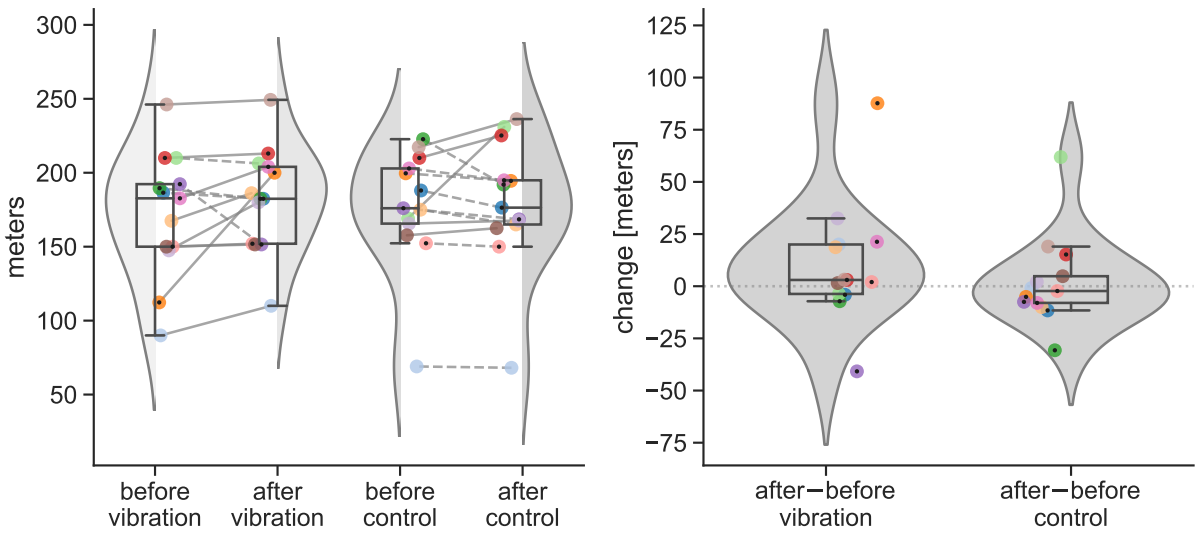

Figure 7: Illustration of the **two minutes walk test** score changes. Marker colors encode participants. *left*: distribution, boxplot, individual scores: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes. Central black dots mark participants in the VC sequence (vibration fist, then control).

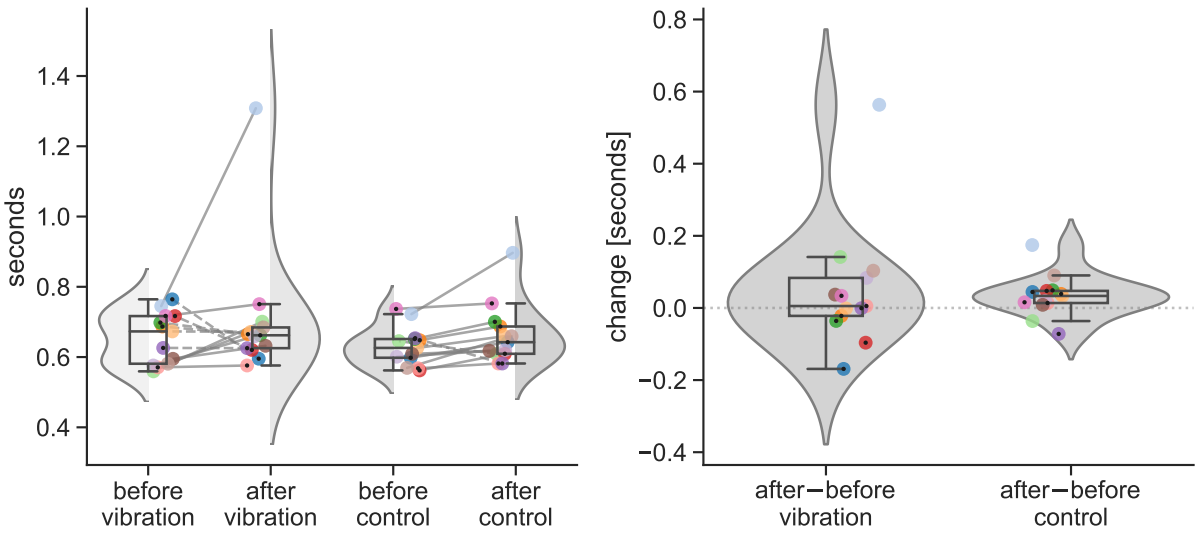

Figure 8: Illustration of the **affected leg stance time** changes measured using GAITRite. Marker colors encode participants. *left*: distribution, boxplot, individual scores averaged over tries: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes averaged over tries. Central black dots mark participants in the VC sequence (vibration fist, then control).

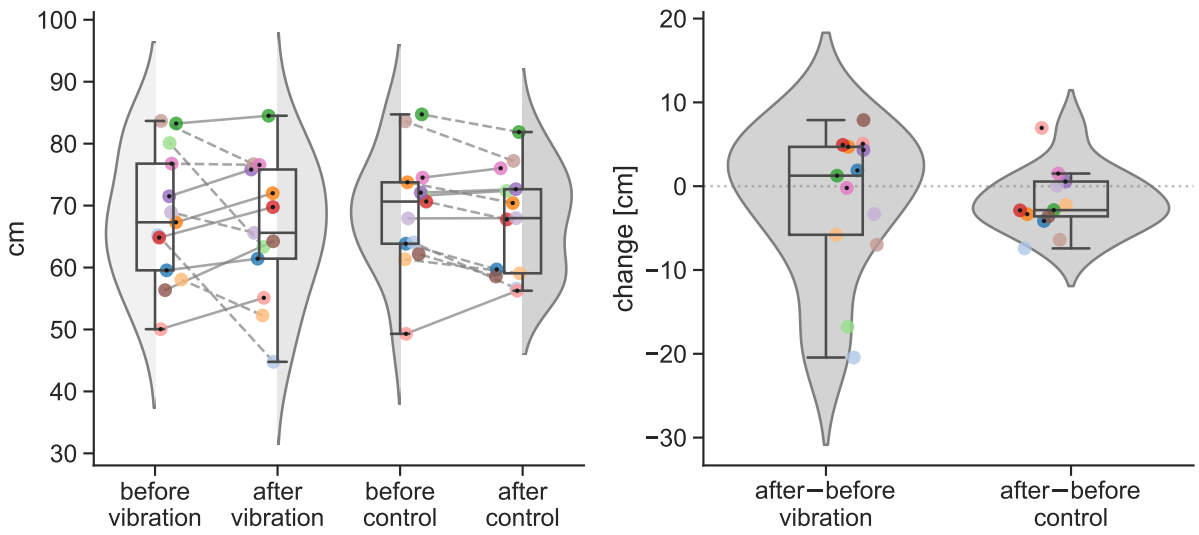

Figure 9: Illustration of the **unaffected leg step length** changes measured using GAITRite. Marker colors encode participants. *left*: distribution, boxplot, individual scores averaged over tries: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes averaged over tries. Central black dots mark participants in the VC sequence (vibration fist, then control).

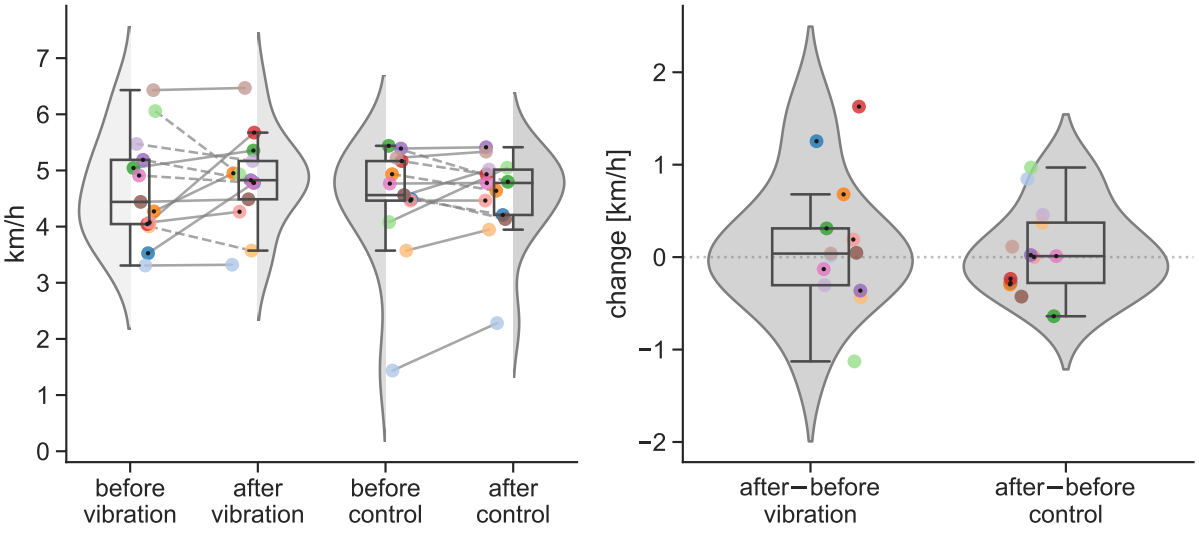

Figure 10: Illustration of the **gait speed** changes measured using GAITRite. Marker colors encode participants. *left*: distribution, boxplot, individual scores averaged over tries: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes averaged over tries. Central black dots mark participants in the VC sequence (vibration fist, then control).

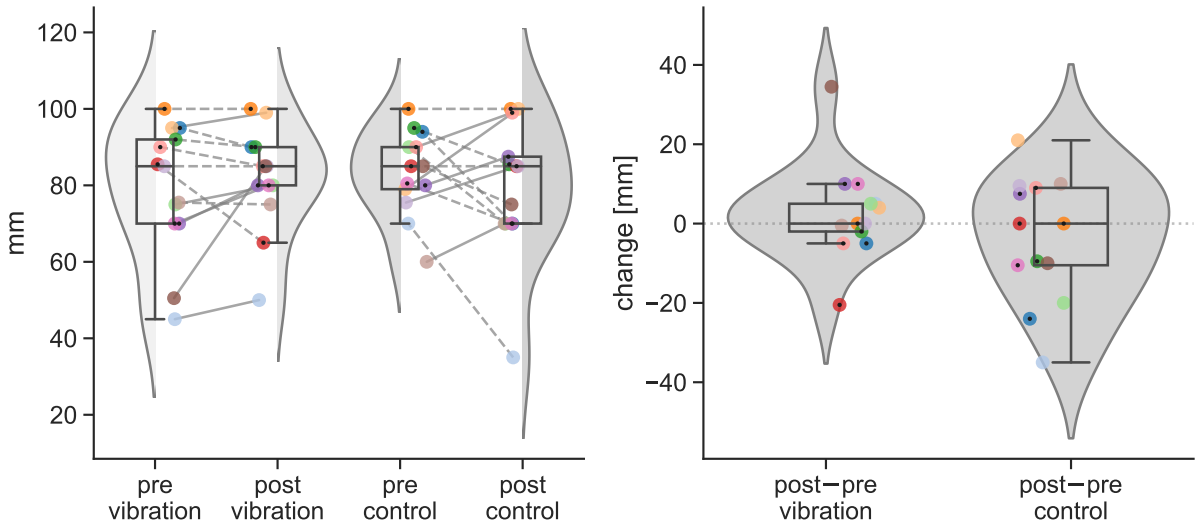

Figure 11: Illustration of the **health-related visual analog scale** score changes. Marker colors encode participants. *left*: distribution, boxplot, individual scores: before vibration (M0) vs. after vibration (M1) and before control (M2) vs. after control (M3). A solid line indicates increase and a dashed line equal or decrease; *right*: distribution, boxplot, and individual within-period changes. Central black dots mark participants in the VC sequence (vibration fist, then control).

1.1.B Likelihood function and parameters

The likelihood function was the Bambi default Gaussian family with identity link [3].

The dependent variable is the scaled test score change ($\Delta\,\text{scaled}_{\text{score}}$) during vibration ($\text{M1} - \text{M0}$) and control ($\text{M3} - \text{M2}$) (Figure 1). Because the test score changes have different magnitudes, they were scaled. The scaling divided the already mostly zero-centered outcome measure changes by the within-test standard deviation (SD). A score change was multiplied by -1 if a positive outcome would be reflected in a negative change such as a faster time, i.e., for the four-square step test (FSST) and timed up and go test (TUG):

$$\Delta\,\text{scaled}_{\text{score}} = \begin{cases} \frac{\text{scores}}{\text{SD}_{\text{scores}}} \cdot (-1), & \text{if a positive outcome would be reflected in a negative change,} \\ \frac{\text{scores}}{\text{SD}_{\text{scores}}}, & \text{otherwise , and was} \end{cases} \tag{1}$$

further scaled for numerical reasons to lie between zero and one by dividing each scaled score by the largest global absolute score, where

$$\Delta\,\text{scaled}_{\text{score}} = \frac{\Delta\,\text{scaled}_{\text{score}}}{\max\{\text{abs}\{\min\{\Delta\,\text{scaled}_{\text{scores}}\}, \max\{\Delta\,\text{scaled}_{\text{scores}}\}\}\}}. \tag{2}$$

The independent variables were intervention with levels V (vibration) and C (control), sequence of enrollment with levels representing vibration first then control (VC) and control first then vibration (CV) evaluating the carryover effect [4], period with levels 1 and 2 encoding the timely order, test with seven levels for the tests (FSST, TUG, two minutes walk test, affected leg stance time, unaffected leg step length, gait speed, health related visual analog score (VAS)), and ID with 13 levels for each participant (Figure 1). Because of unsolved numerical challenges tries were not used in the random effects structure.

1.1.C Prior distribution

The default normal distribution priors of the BAyesian Model-Building Interface (BAMBI) were automatically centered around zero, and their width exceeded the scaling of Equation 2. Thus, the priors did not introduce any information.

- The global model (Figure 2) used default wide weakly informative normal distribution priors with mean $\mu = 0$ indicating no change and sigma $\sigma = 1.04$ for $\text{intervention}, \text{period},$ and sequence allowing for changes within the whole range of the scaled scores. The group-level used a normal distribution prior with $\mu = 0$ and $\sigma = \text{HalfNormal}\,(\sigma = 1.18)$. The auxiliary parameter prior was $\text{HalfStudentT}\,(\nu = 4, \sigma = 0.21)$.
- The test-level model (Figure 3) used default wide weakly informative normal distribution priors with mean $\mu = 0$ and sigma $\sigma \in [1.83, 3.07]$ for each of the eight interactions for $\text{test} : \text{intervention}$. The group levels used normal distribution priors with $\mu = 0$ and $\sigma = \text{HalfNormal}\,(\sigma \in [0.75, 1.04])$. The auxiliary parameter prior was $\text{HalfStudentT}\,(\nu = 4, \sigma = 0.21)$.
- The ID-level model (Figure 4) used default wide weakly informative normal distribution priors with mean $\mu = 0$ and sigma $\sigma = 2.7$ for each of the eight interactions for $\text{ID} : \text{intervention}$. The group levels used normal distribution priors with $\mu = 0$ and $\sigma = \text{HalfNormal}\,(\sigma \in [0.74, 1.04])$. The auxiliary parameter prior was $\text{HalfStudentT}\,(\nu = 4, \sigma = 0.21)$.

1.1.D Formal specification

The technical report of Capretto et al. 2022 includes the formal definition of the used default Bambi priors [3]. The package documentation and code are available on https://bambinos.github.io/bambi/.

1.1.E Prior and posterior predictive check

- Figure 12 shows the prior (left column) and posterior (right column) predictive checks for the global model for $\text{intervention}, \text{period},$ and sequence demonstrating that the prior and posterior distributions were consistent with the observed data.

- Figures 13 and 14 show the prior (left column) and posterior (right column) predictive checks for the test-level model for for all 7 levels of test for and $\mathrm{intervention}$ levels demonstrating that the prior and posterior distributions were consistent with the observed data.
- Figures 15, 16,17, and 18 show the prior (left column) and posterior (right column) predictive checks for the test-level model for for all 13 levels of ID for and $\mathrm{intervention}$ levels demonstrating that the prior and posterior distributions were consistent with the observed data.

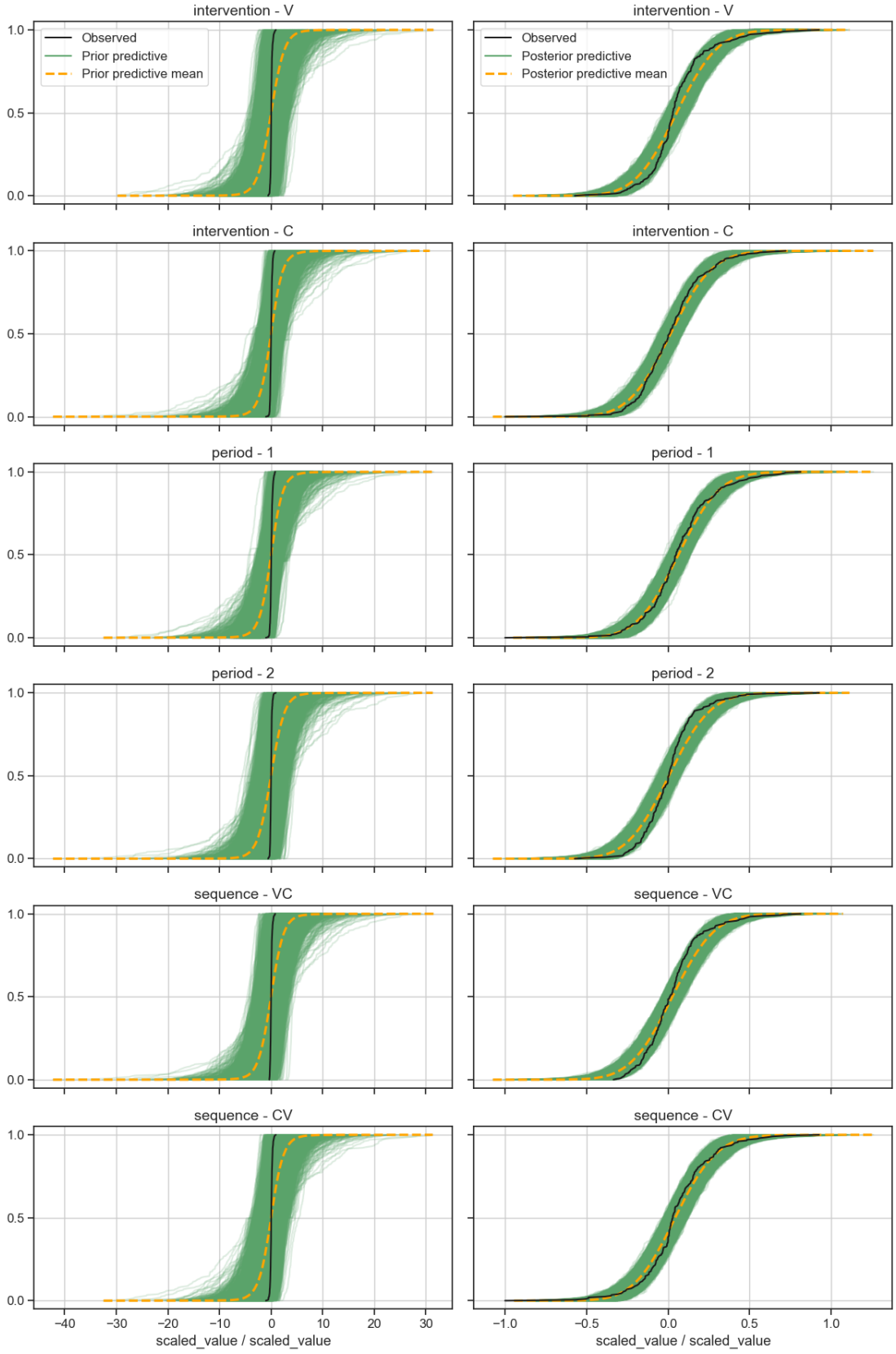

Figure 12: Prior (left) and posterior (right) predictive checks for all levels of intervention, period, and sequence as empirical cumulative distribution functions. V ... vibration intervention, C ... intervention control, 1 ... first period in time, 2 .. second period in time.

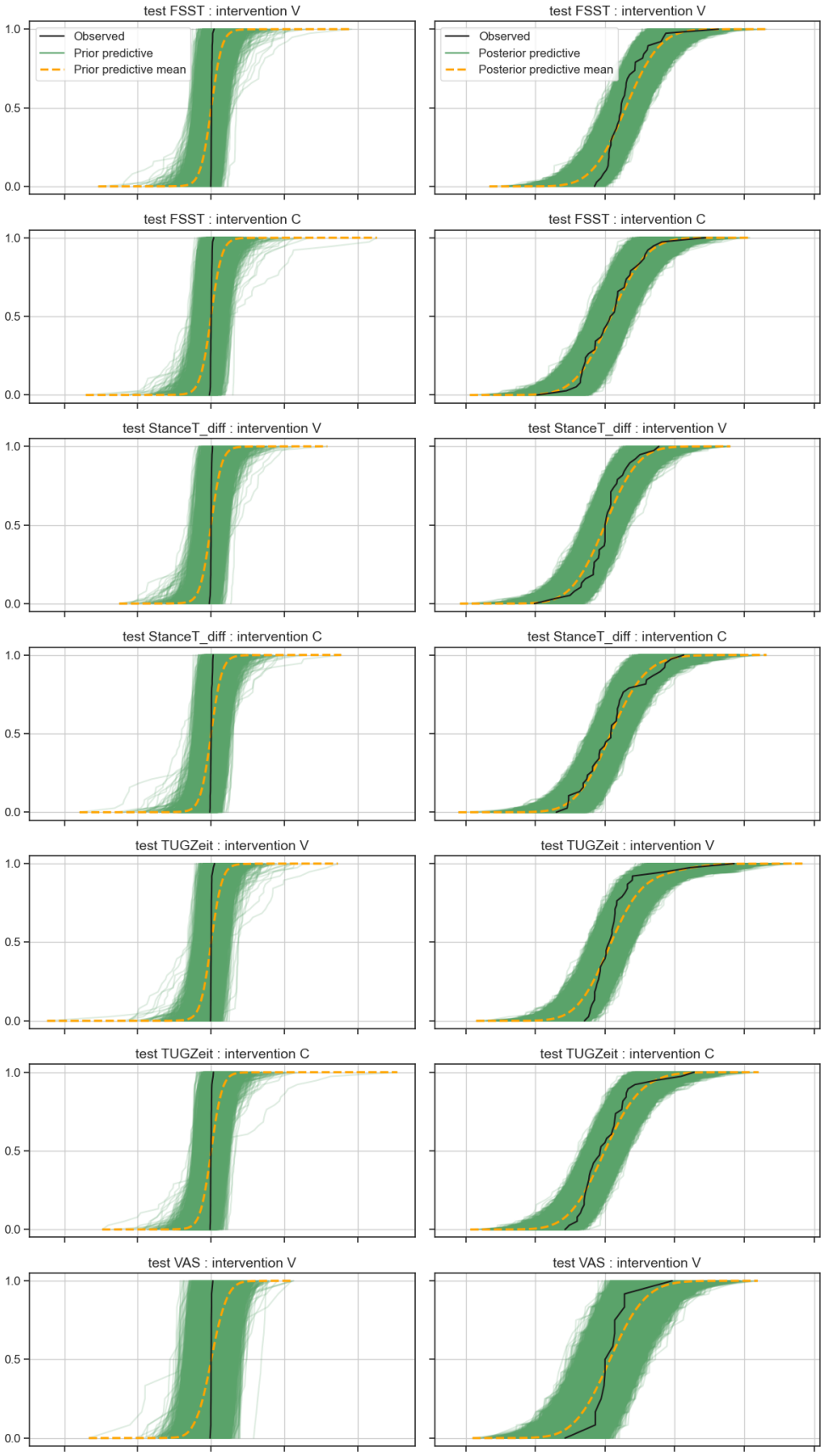

Figure 13: Prior (left) and posterior (right) predictive checks for all levels of test for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention (continued in Figure 14).

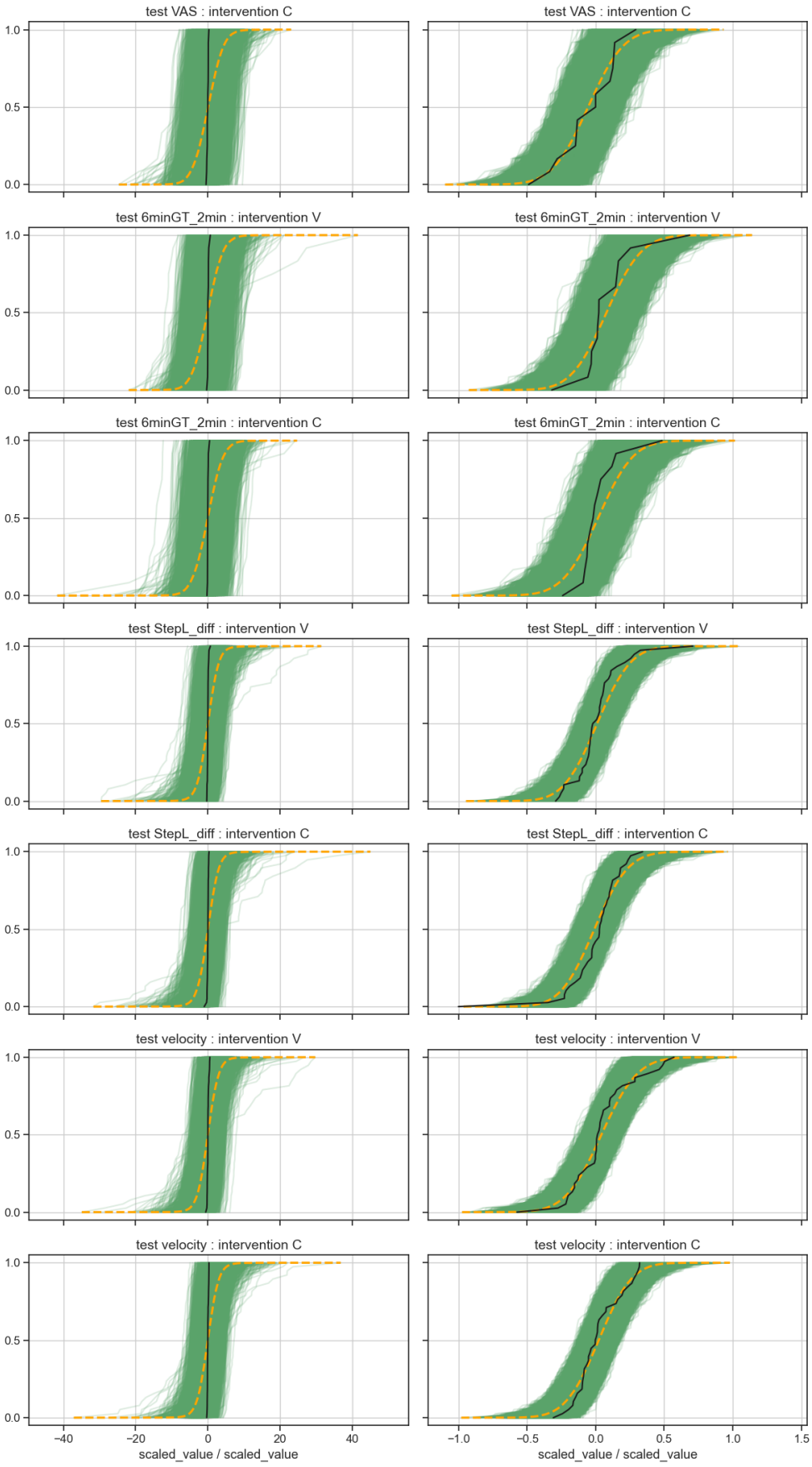

Figure 14: Continued Figures 13. Prior (left) and posterior (right) predictive checks for all levels of test for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention.

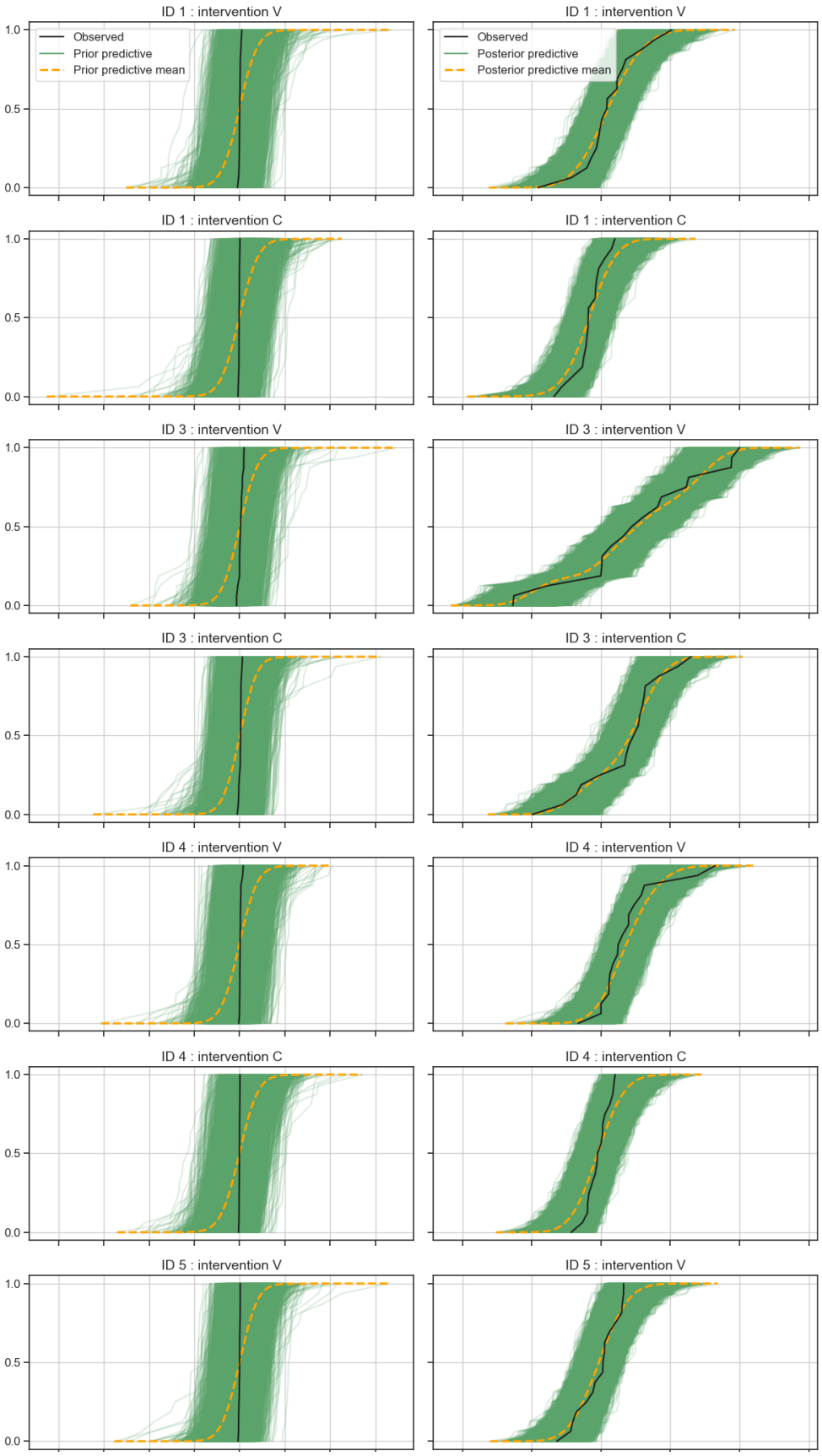

Figure 15: Prior (left) and posterior (right) predictive checks for all levels of id for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention (continued in Figure 16).

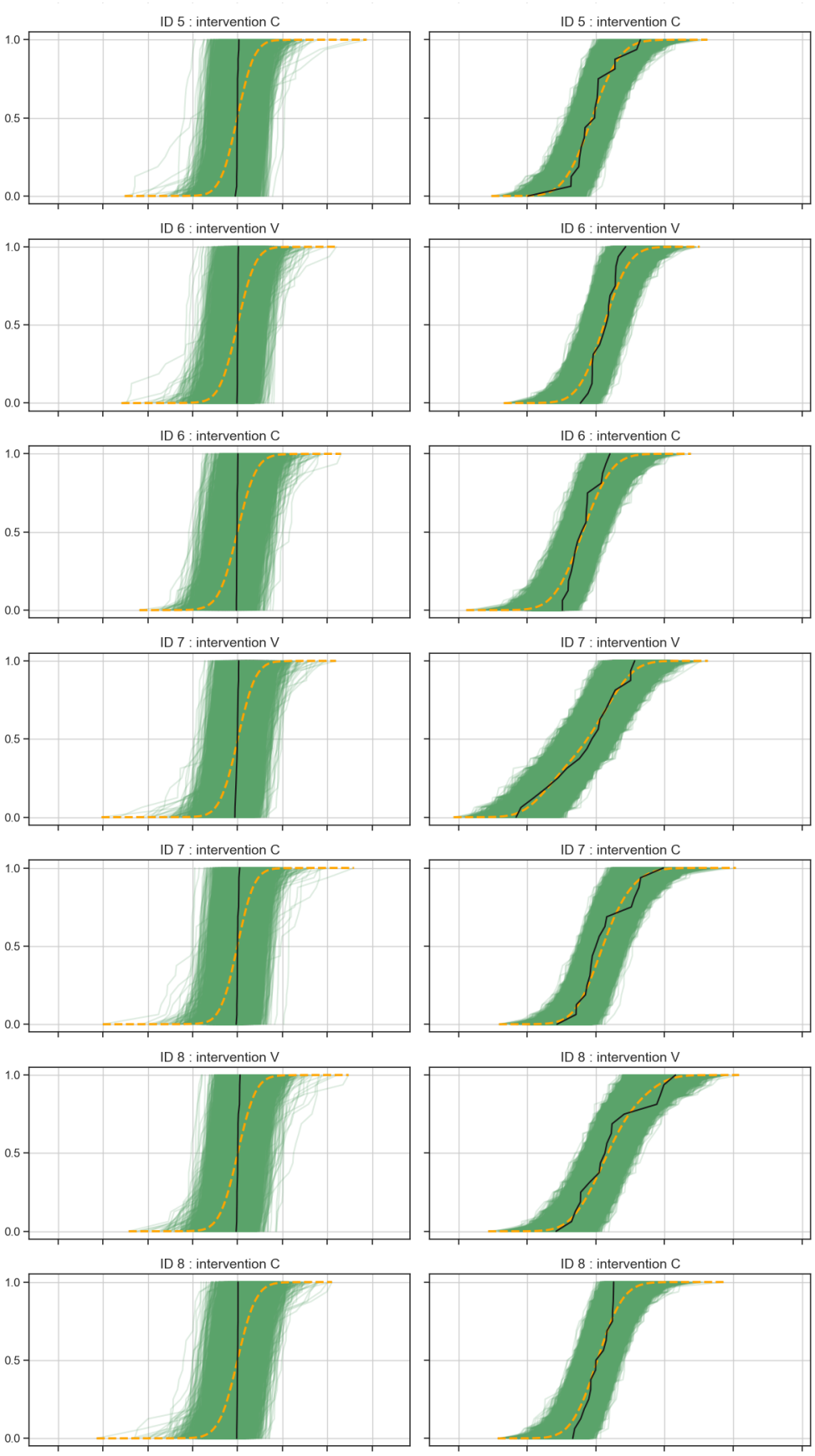

Figure 16: Prior (left) and posterior (right) predictive checks for all levels of id for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention (continued in Figure 17).

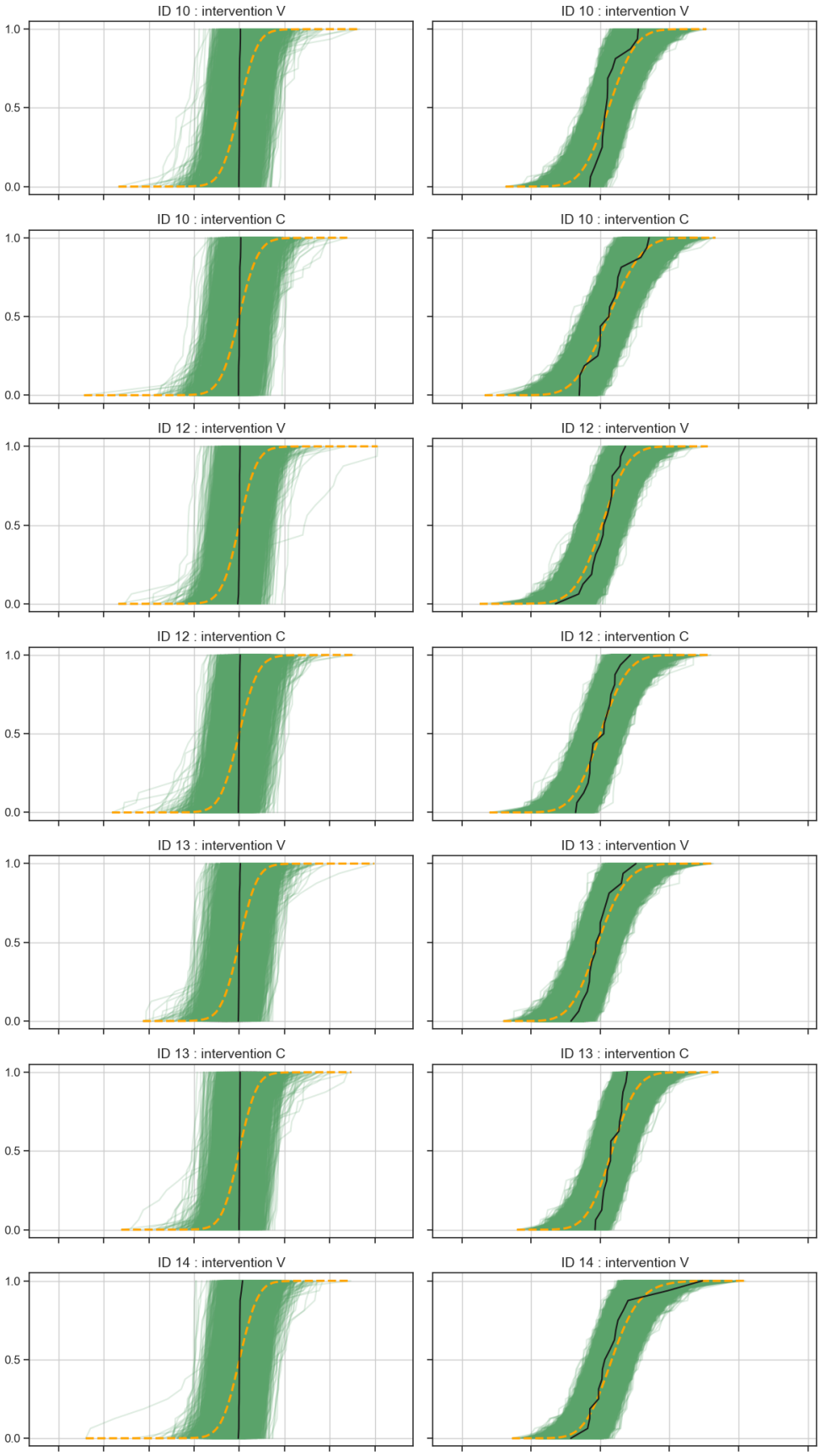

Figure 17: Continued Figures 15. Prior (left) and posterior (right) predictive checks for all levels of ID for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention (continued in Figure 18).

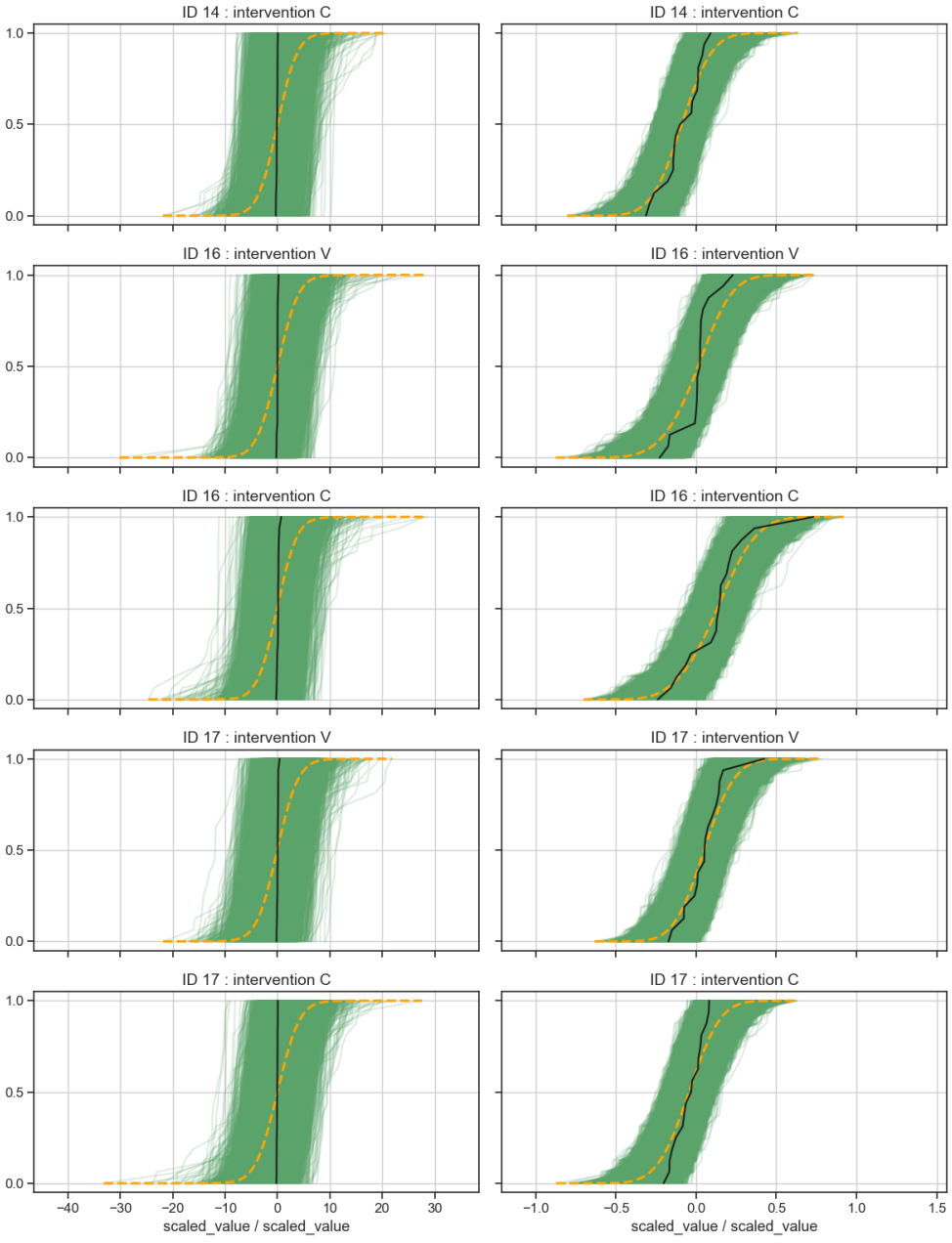

Figure 18: Continued Figures 15. Prior (left) and posterior (right) predictive checks for all levels of ID for and intervention levels as empirical cumulative distribution functions. V ... vibration intervention, C ... control intervention.

1.2 Report details of the computation

1.2.A Software

I used Python (3.11.0) with bambi (0.9.1), arviz (0.13.0), pymc (4.3.0), pandas (1.5.1), numpy (1.23.4), formulae (0.3.4), and scipy (1.9.3) for the analysis. The fitting of the Bambi models used 2100 tune draws and 4000 draws for the global model, 11 850 tune draws and 20 000 draws for the test-level model, and 6000 tune draws and 10 000 draws on eight chains with random seed 1234. The parameter $\mathrm{target\_accept}$ was set to 0.98 for all models.

1.2.B MCMC chain convergence

The Markov chain Monte Carlo effective sample size and convergence statistics for the global, test-level, and ID-level models are shown in Table 2 and Figures 19 and 20.

Table 2: Effective sample size (ESS) and Markov chain Monte Carlo convergence ($\hat{\mathrm{r}}$).

	ESS				$\hat{\mathrm{r}}$			
model	mean	sd	min	max	mean	sd	min	max
global	59 239	8 735	10 144	75 809	1	0	1	1
tests	124 465	19 427	39 003	146 662	1	0	1	1
IDs	95 650	14 523	19 002	132 367	1	0	1	1

ESS rounded to integers and $\hat{\mathrm{r}}$ to three decimals.

1.2.C MCMC chain resolution

Figure 19 shows the Markov chain Monte Carlo chain resolution (effective samples) and convergence statistic $\hat{\mathrm{r}}$ together with the $\Delta\,\mathrm{scaled}_{\mathrm{score}}$ differences between periods.

1.3 Describe the posterior distribution

1.3.A Posterior predictive check

See Section 1.1.E and the right column of Figures 12 to 18.

1.3.B Summarize posterior of variables

Figure 20 shows then mean, interquartile range, and mean of all estimated posterior variables together with their effective sample size and MCMC convergence statistic.

Table 3 shows the results of the global fixed effects and test-level multiple comparisons based on the 95% highest density intervall (HDI) and Bayes factor (BF). Table 4 shows the results of the global fixed effects and test-level multiple comparisons based on the 95% highest density intervall (HDI) and Bayes factor (BF).

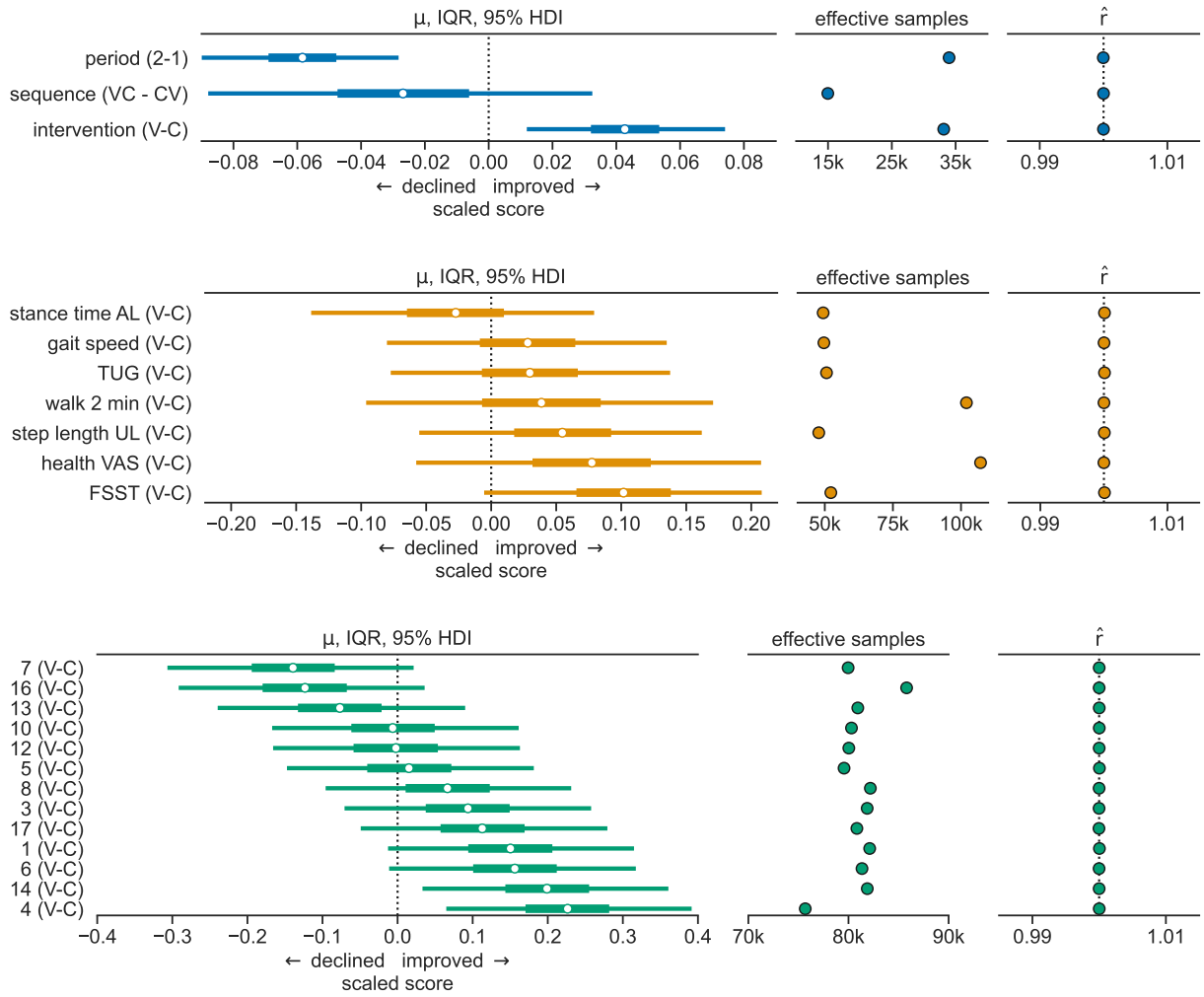

Figure 19: Mean (μ), interquartile range (IQR), and 95% highest density interval (HDI) of $\Delta\,\mathrm{scaled}_{\mathrm{score}}$ differences between period 1 and 2, $\mathrm{sequence}$ CV and VC, and $\mathrm{intervention}$ C (control) and V (vibration). The top panel (blue) represents the global model, middle (orange) the test-level model, and bottom (green) the ID-level model. Markov chain Monte Carlo convergence statistic ... $\hat{\mathrm{r}}$, affected leg ... AL, timed up and go test ... TUG, unaffected leg ... UL, visual analog scale ... VAS, four square step test ... FSST

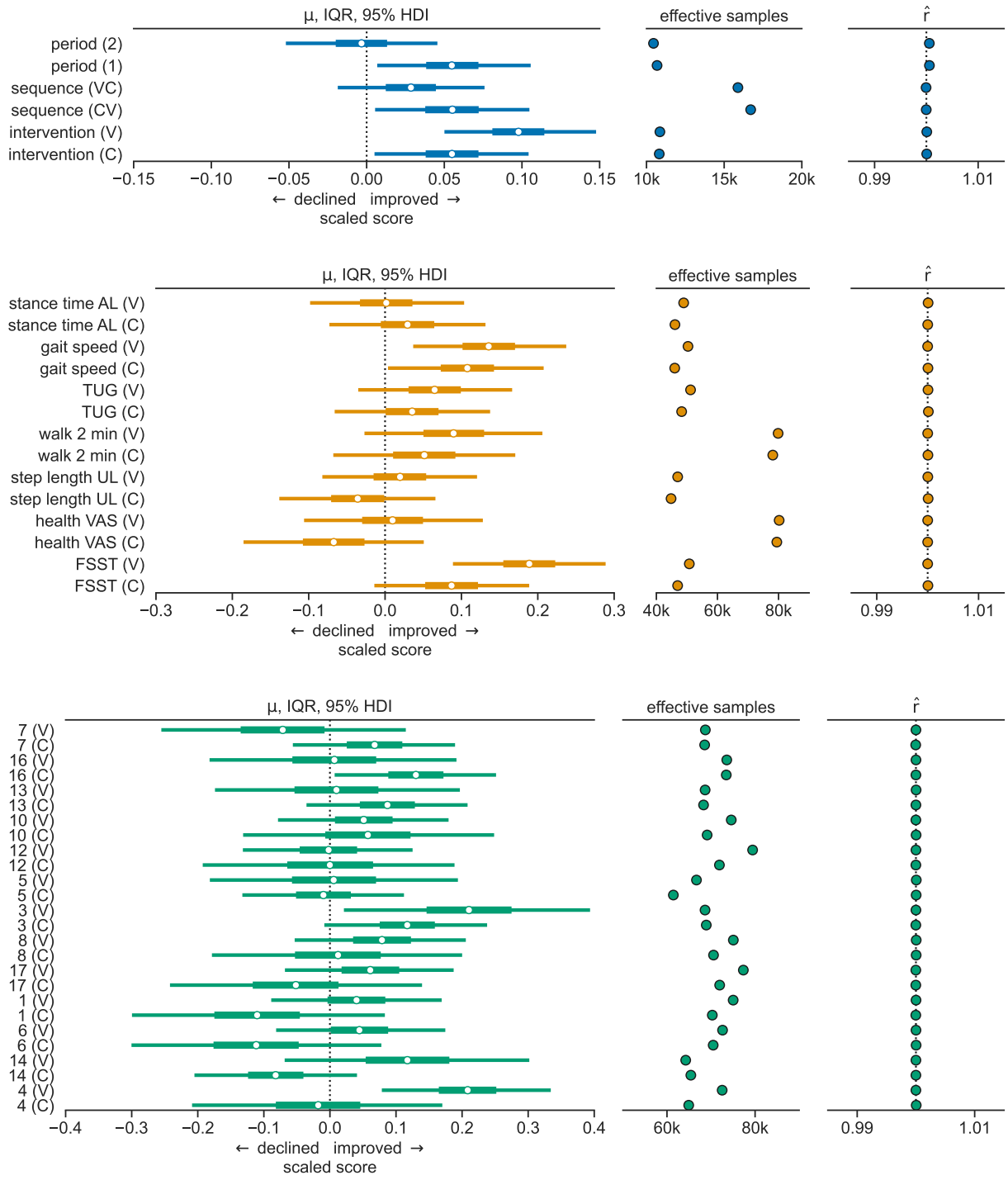

Figure 20: Mean (μ), interquartile range (IQR), and 95% highest density interval (HDI) of $\Delta\,\mathrm{scaled}_{\mathrm{score}}$ estimates for period 1 and 2, $\mathrm{sequence}$ CV and VC, and $\mathrm{intervention}$ C (control) and V (vibration). The top panel (blue) represents the global model, middle (orange) the test-level model, and bottom (green) the ID-level model. Markov chain Monte Carlo convergence statistic ... $\hat{\mathrm{r}}$, affected leg ... AL, timed up and go test ... TUG, unaffected leg ... UL, visual analog scale ... VAS, four square step test ... FSST

Table 3: Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a global outcome or a clinical test is different between the cross-over sequences CV and VC, between the time periods 1 and 2, and between interventions vibration (V) and control (C) under H1. Sequence encodes the global carry-over effect, period the time order effect, and intervention the vibration effect. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	sequence VC_Δ-CV_Δ	period 2_Δ-1_Δ	intervention V_Δ-C_Δ	TUG V_Δ-C_Δ	walk 2 min V_Δ-C_Δ	health VAS V_Δ-C_Δ	FSST V_Δ-C_Δ	stance time AL V_Δ-C_Δ	step length UL V_Δ-C_Δ	gait speed V_Δ-C_Δ
BF_{01}	0.32	0.0	0.05	0.73	0.66	0.18	0.02	0.72	0.31	0.74
p (posteriors equal) [%]	24.3	0.4	5.1	42.1	39.8	15.6	1.9	42.0	23.5	42.5
p (large negative effect) [%]	0.0	0.0	0.0	0.0	0.03	0.0	0.0	0.0	0.02	0.0
p (medium negative effect) [%]	0.0	8.76	0.0	0.0	0.36	0.05	0.0	0.09	1.48	0.12
p (small negative effect) [%]	21.36	87.62	0.0	0.09	2.45	0.57	0.0	3.42	18.15	4.58
p (negative effect < small) [%]	21.36	96.38	0.0	0.09	2.84	0.62	0.0	3.51	19.66	4.7
effect size in ROPE [%]	39.02	0.22	2.69	3.18	7.76	2.75	0.02	22.72	34.03	25.33
effect size lower HDI 0.95%	-0.32	-0.56	0.1	0.07	-0.22	0.02	0.44	-0.23	-0.45	-0.26
effect size mean	-0.12	-0.37	0.28	0.51	0.55	0.8	0.89	0.21	-0.01	0.18
effect size upper HDI 0.95%	0.06	-0.18	0.47	0.96	1.32	1.57	1.34	0.65	0.44	0.63
p (positive effect > small) [%]	0.04	0.0	81.01	91.53	81.27	93.39	99.87	51.84	18.06	46.54
p (small positive effect) [%]	0.04	0.0	79.87	39.04	26.45	16.12	4.24	41.73	16.81	38.39
p (medium positive effect) [%]	0.0	0.0	1.15	42.0	28.77	27.43	29.95	9.61	1.23	7.81
p (large positive effect) [%]	0.0	0.0	0.0	10.49	26.04	49.84	65.68	0.49	0.02	0.34
p (posteriors differ) [%]	75.7	99.6	94.9	57.9	60.2	84.4	98.1	58.0	76.5	57.5
BF_{10}	3.12	226.67	18.61	1.38	1.51	5.41	51.85	1.38	3.26	1.35
prior p (reject difference) [%]	1.66	0.02	0.28	3.68	3.36	0.96	0.1	3.66	1.59	3.75
prior p (accept difference) [%]	85.91	7.73	50.52	93.24	92.63	77.85	26.82	93.21	85.34	93.36
model	global	global	global	tests	tests	tests	tests	tests	tests	tests
	$\neq_{\mathrm{moderate}}$	$\neq_{\mathrm{strong}}$ $-_{\mathrm{S}}$	$\neq_{\mathrm{strong}}$			$\neq_{\mathrm{moderate}}$	$\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$		$\neq_{\mathrm{moderate}}$	

TUG timed up & go test, VAS visual analog scale, FSST four square step test, and Δ indicates a difference of either $\mathrm{post} - \mathrm{pre}$ vibration or control test times where period 1 was earlier in time than period 2, or $\mathrm{left} - \mathrm{right}$ in stance time difference and step length difference.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm\mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

Table 4: Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a participant is different between the interventions vibration (V) and control (C) under H1. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	ID 7 V_Δ-C_Δ	ID 16 V_Δ-C_Δ	ID 13 V_Δ-C_Δ	ID 10 V_Δ-C_Δ	ID 12 V_Δ-C_Δ	ID 5 V_Δ-C_Δ	ID 3 V_Δ-C_Δ	ID 8 V_Δ-C_Δ	ID 17 V_Δ-C_Δ	ID 1 V_Δ-C_Δ	ID 6 V_Δ-C_Δ	ID 14 V_Δ-C_Δ	ID 4 V_Δ-C_Δ
BF_{01}	0.03	0.06	0.32	1.0	1.0	0.94	0.2	0.45	0.11	0.02	0.02	0.0	0.0
p (posteriors equal) [%]	2.9	5.6	24.0	49.9	50.1	48.4	16.9	31.0	9.8	2.1	1.6	0.1	0.0
p (large negative effect) [%]	89.66	74.31	48.44	0.93	0.59	0.42	0.0	0.0	0.0	0.0	0.0	0.0	0.0
p (medium negative effect) [%]	8.64	19.25	31.53	6.21	4.52	3.67	0.04	0.01	0.0	0.0	0.0	0.0	0.0
p (small negative effect) [%]	1.58	5.56	15.79	20.69	17.23	15.38	0.63	0.2	0.13	0.0	0.0	0.0	0.0
p (negative effect < small) [%]	99.88	99.13	95.76	27.83	22.34	19.47	0.67	0.22	0.14	0.0	0.0	0.0	0.0
effect size in ROPE [%]	0.04	0.31	1.79	22.78	22.61	22.36	4.1	2.07	1.51	0.06	0.03	0.0	0.0
effect size lower HDI 0.95%	-1.92	-1.7	-1.46	-0.68	-0.6	-0.58	-0.03	0.09	0.14	0.53	0.59	0.85	1.1
effect size mean	-1.24	-1.03	-0.79	0.0	0.06	0.09	0.65	0.77	0.81	1.2	1.28	1.53	1.8
effect size upper HDI 0.95%	-0.56	-0.34	-0.11	0.67	0.75	0.75	1.32	1.43	1.49	1.89	1.96	2.23	2.48
p (positive effect > small) [%]	0.0	0.02	0.2	28.44	34.39	37.75	90.36	95.06	96.3	99.82	99.92	100.0	100.0
p (small positive effect) [%]	0.0	0.02	0.19	20.87	24.28	26.02	23.71	17.11	14.54	2.0	1.18	0.18	0.02
p (medium positive effect) [%]	0.0	0.0	0.01	6.44	8.48	9.82	33.66	31.79	30.14	10.16	7.08	1.68	0.3
p (large positive effect) [%]	0.0	0.0	0.0	1.12	1.63	1.91	32.99	46.17	51.61	87.66	91.66	98.14	99.68
p (posteriors differ) [%]	97.1	94.4	76.0	50.1	49.9	51.6	83.1	69.0	90.2	97.9	98.4	99.9	100.0
BF_{10}	33.68	16.77	3.17	1.0	1.0	1.06	4.91	2.22	9.21	45.54	61.15	1230.88	4928.43
prior p (reject difference) [%]	0.16	0.31	1.64	4.98	5.02	4.71	1.06	2.31	0.57	0.12	0.09	0.03	0.46
prior p (accept difference) [%]	36.07	53.12	85.72	94.98	95.02	94.69	79.48	89.52	67.36	29.44	23.71	1.52	0.38
model	IDs $\neq_{\mathrm{strong}}$ $-_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $-_{\mathrm{L}}$	IDs $\neq_{\mathrm{moderate}}$ $-_{\mathrm{M}}$	IDs	IDs	IDs	IDs $\neq_{\mathrm{moderate}}$	IDs	IDs $\neq_{\mathrm{moderate}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$

Δ indicates a difference of between control and vibration.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm \mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

1.3.C BF and posterior model probabilities

Tables 3 and 4 list several probabilities of the posterior distributions together with Bayes factors (BF) and prior probabilities. Figures 21 to 24 show the posterior distributions and several prior probabilities for rejecting and accepting distribution differences in detail.

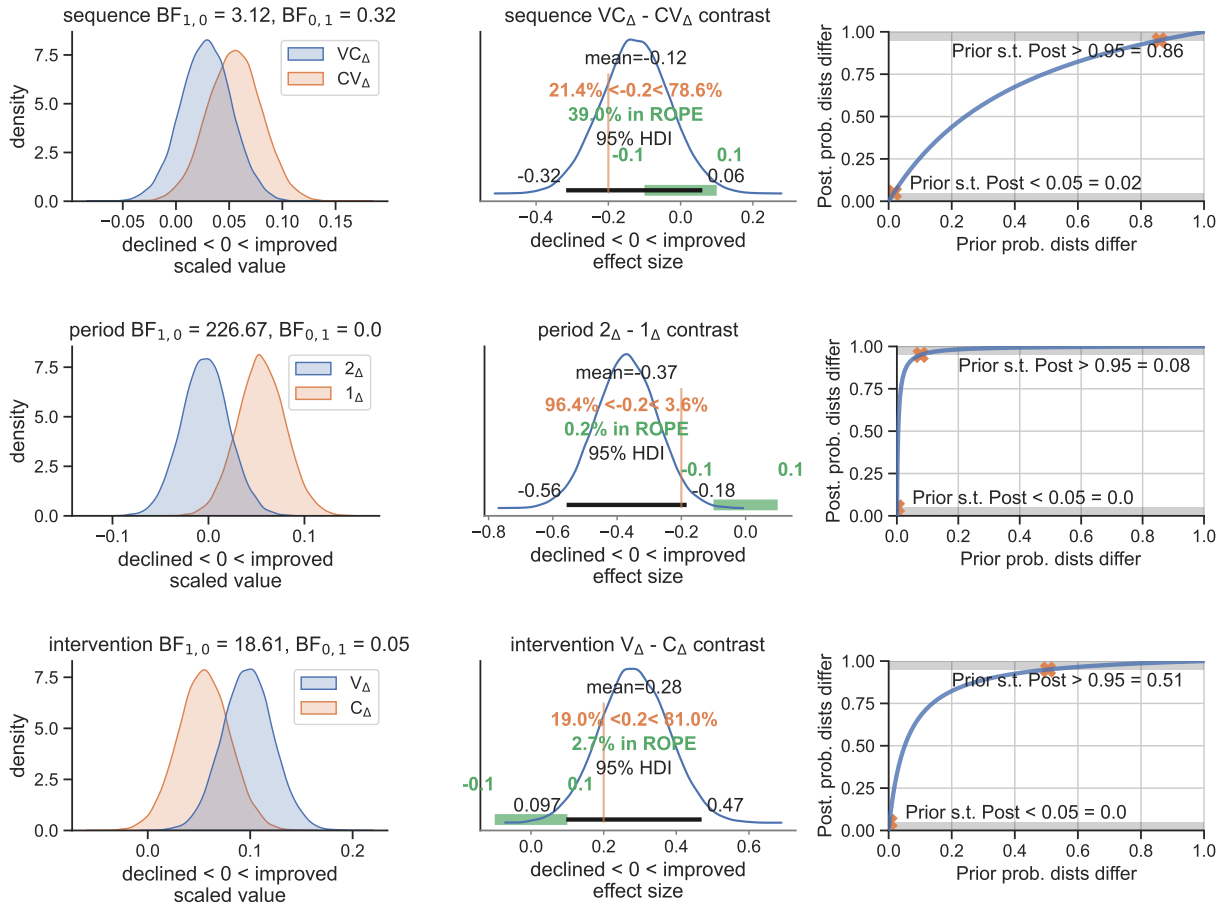

Figure 21: Left: Posterior distributions of the levels of sequence, period, and intervention ($\Delta\,\text{scaled}_{\text{score}}$); Middle: Hedges' g effect size and probabilities of the difference between levels; Right: Required prior probability to accept (posterior probability >0.95) or reject (posterior probability <0.05) the belief that the distributions of the two levels are different.

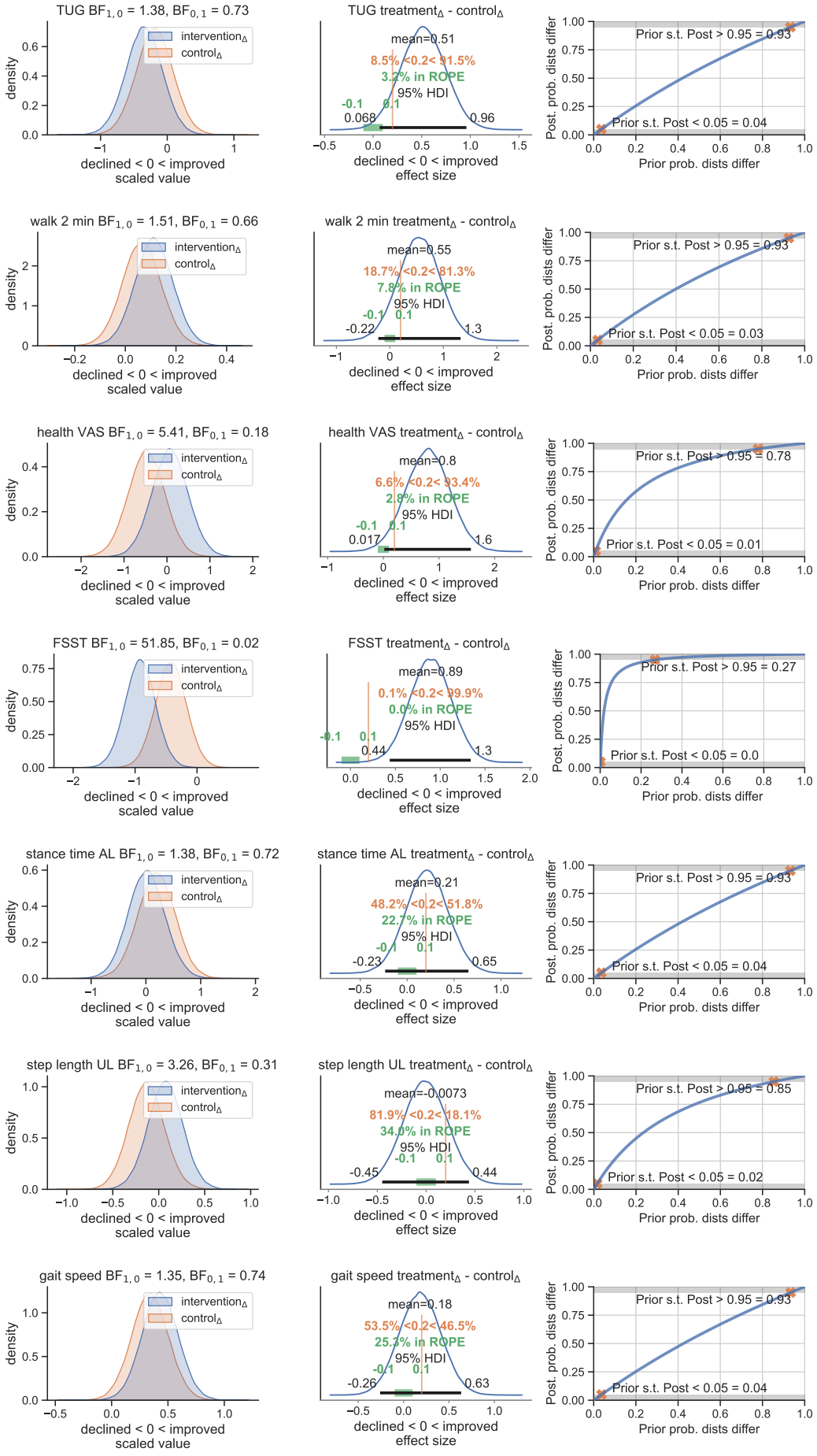

Figure 22: Left: Posterior distributions of the levels of test (Δ scaled$_{\text{score}}$); Middle: Hedges' g effect size and probabilities of the difference between levels; Right: Required prior probability to accept (posterior probability >0.95) or reject (posterior probability <0.05) the belief that the distributions of the two levels are different.

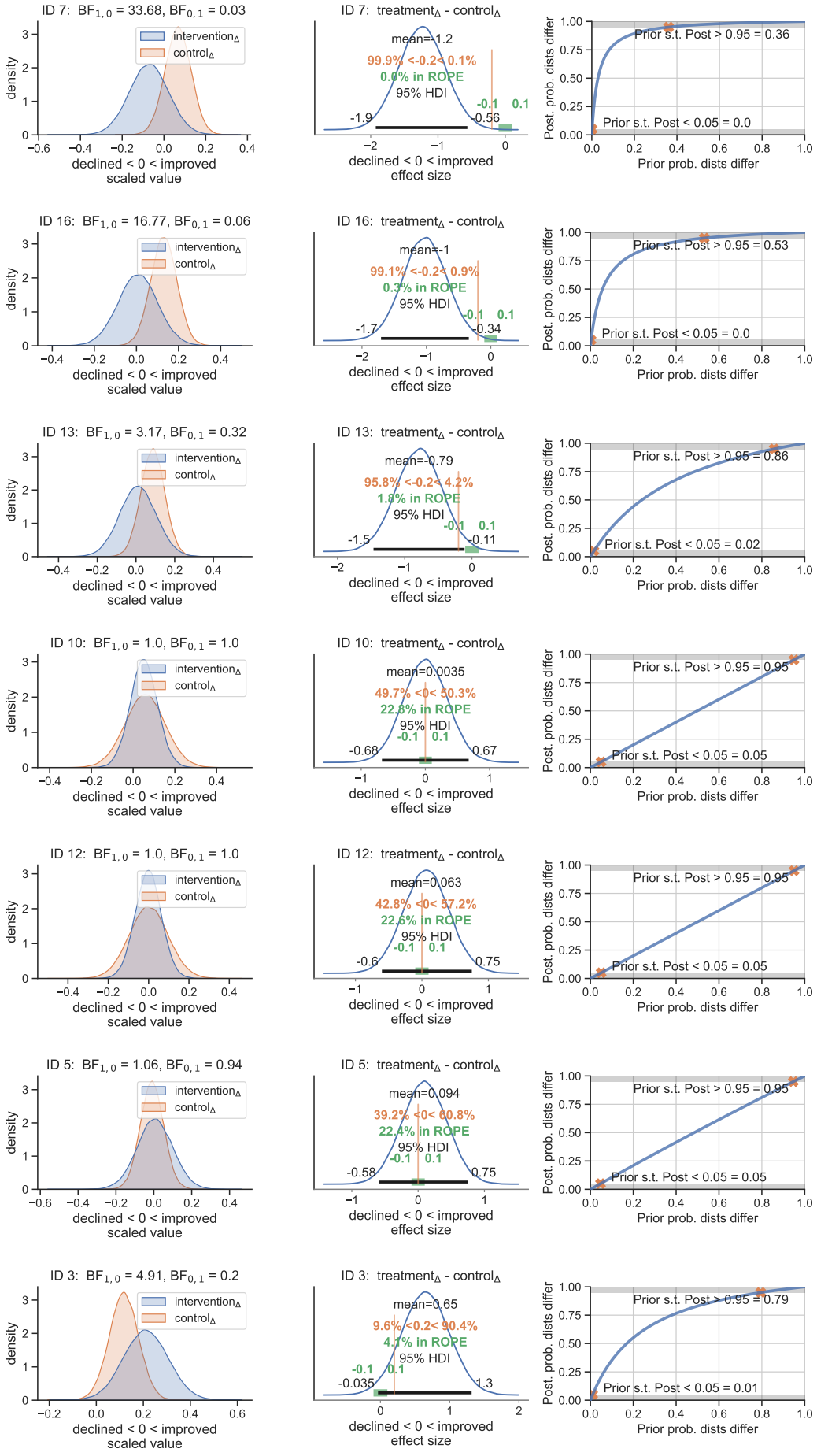

Figure 23: Left: Posterior distributions of the levels of ID ($\Delta\,\mathrm{scaled}_{\mathrm{score}}$); Middle: Hedges' g effect size and probabilities of the difference between levels; Right: Required prior probability to accept (posterior probability >0.95) or reject (posterior probability <0.05) the belief that the distributions of the two levels are different. Continued in Figure 24.

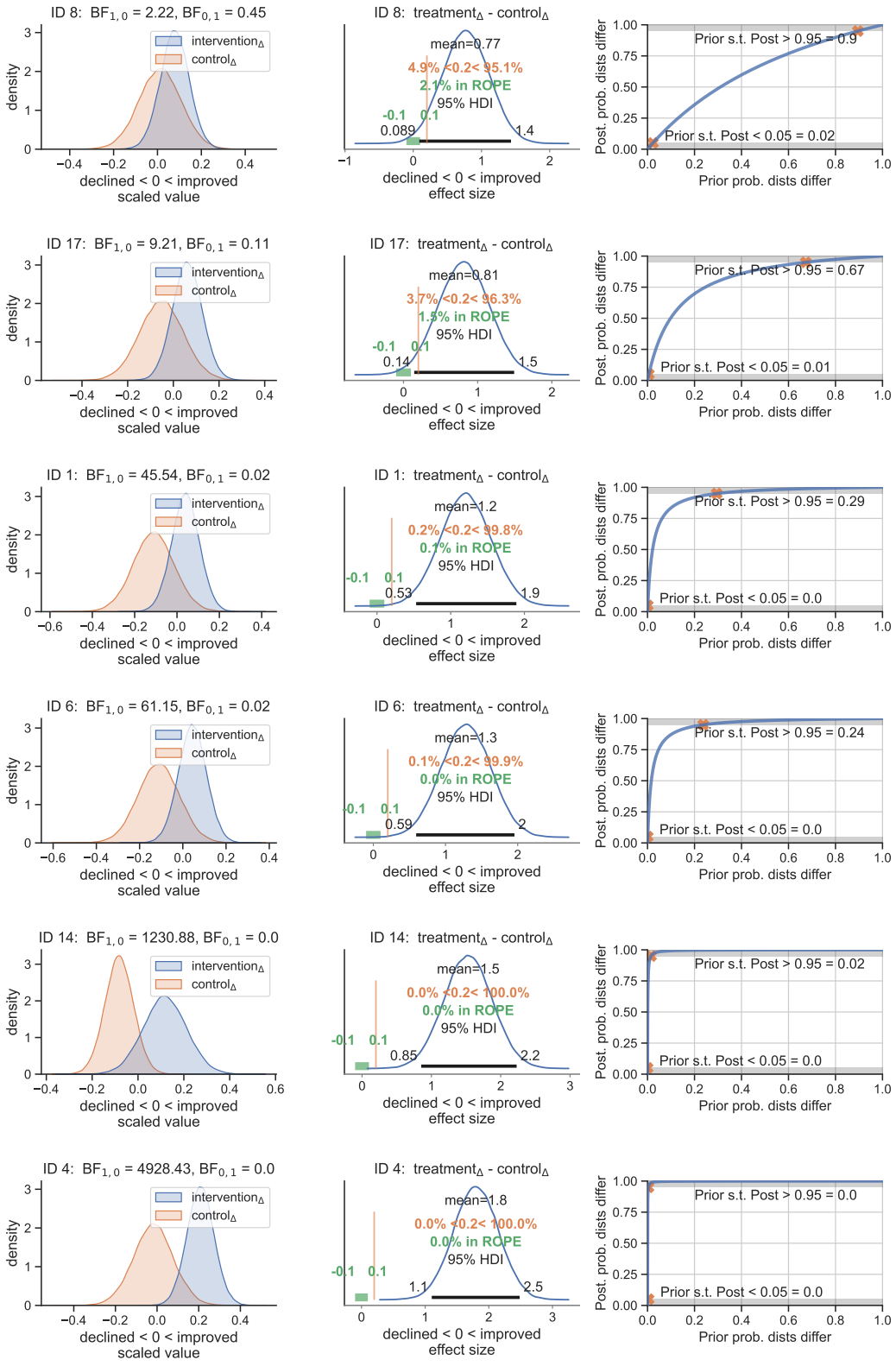

Figure 24: Left: Posterior distributions of the levels of ID ($\Delta\,\text{scaled}_{\text{score}}$); Middle: Hedges' g effect size and probabilities of the difference between levels; Right: Required prior probability to accept (posterior probability >0.95) or reject (posterior probability <0.05) the belief that the distributions of the two levels are different. Continued from Figure 23

1.4 Report decisions (if any) and their criteria

1.4.A Why decisions?

We decide the effectiveness of the study design and the focal vibration feedback intervention concerning effect size and Bayes factor.

1.4.B Loss function

Not applicable.

1.4.C ROPE limits

ROPE limits were set to Hedge's g effect size $\in [-0.1, 0.1]$, which represents an effect size magnitude of 0.2 which is the accepted threshold for small effects [1].

1.4.D BF, decision threshold and model probabilities

Tables 3 and 4 and Figures 21 to 24 provide the detailed description of the posterior and all derived decision thresholds.

A discernible effect is acknowledged if

- the 95% HDI of global, test-level, and effect sizes remains outside the ROPE $\in [-0.1, 0.1]$ or falls completely inside [6], and
- if the Bayes factor ≥ 3 indicates similarity or difference of the compared levels at a prior probability < 0.55 [1].
- Binary categorization of responders and non-responders at the ID-level model was based on the probability such that the individual global effect > 0.2 ($15\,000$ random samples from the ID-level posterior distribution were used to estimate the ability of tests and the related thresholds to predict the classes at M0 via receiver-operator characteristics and true skill statistic (Youden-index)) [7, 8].

Based on those decision rules, we are 95% sure that our decisions were informed by distributions that were different (or equal) and that the effect was extreme enough such that the majority of the probability mass lay above a small effect (or within the ROPE). An additional consideration is given when more than 80% of the probability mass of the effect size lies above 0.2, which may be interpreted as a measure of statistical power, usually accepted if greater than 0.8.

To check the custom implementation of multiple comparisons and effect size calculation, I compared the results of the global model (Table 1, Figure 2) with established R packages for frequentist GLMMs using the same data. Table 5 shows the effect sizes and p-values calculated in R (version 4.2.2, lmerTest v3.1-3 ::lmer, performance v0.10.1, emmeans v1.8.2, set.seed(1234)). These results are similar to the implementations of the Bayesian GLMM and would lead to the same conclusions (Table 3, Figure 21). The test-level and ID-level models did not converge using lmerTest::lmer.

1.4.E Estimated values too

Tables 3 and 4 and Figures 21 to 24 show all estimted values.

Table 5: Effect sizes from estimates marginal means for the difference between vibration period effect and control period effect calculated in R using a frequentist generalized linear mixed effects model (lmerTest::lmer).

				95% CI		
	effect size	SE	df	lower	upper	p-value
sequence ($VC_{\triangle}$-$CV_{\triangle}$)						
lmer	-0.163	0.182	90.4	-0.525	0.199	0.3744
period ($2_{\triangle}$-$1_{\triangle}$)						
lmer	-0.352	0.0964	356	-0.542	-0.163	0.0003
intervention ($V_{\triangle}$-$C_{\triangle}$)						
lmer	0.258	0.0961	356	0.0695	0.447	0.0073

CI ... confidence interval, SE ... standard error, df ... degrees of freedom. Table 3 describes further abbreviations and classifications.

1.5 Report sensitivity analysis

1.5.C For default priors

Figure 25 compares the parametrized prior distributions used with the default settings and the custom-prior-scenarios using neutral (centered around zero) normal, Student's T, and Cauchy priors. Figures 26 and 27 compare the effect sizes derived from the different priors and show that there is no discernible difference between them. Comparing Tables 3 and 4 to Tables 6 to 11 shows that despite differences in posterior distributions that the final decisions would not change. Please note that the Student's T and Cauchy priors induced problems in the estimation of the posterior (low estimated sample size and MCMC convergence problems). While the Student's T prior had problems with precision (very wide posterior) and caused complete Bayes factor insensitivity, the ROPE+HDI testing was still robust and led to similar compared to the default prior.

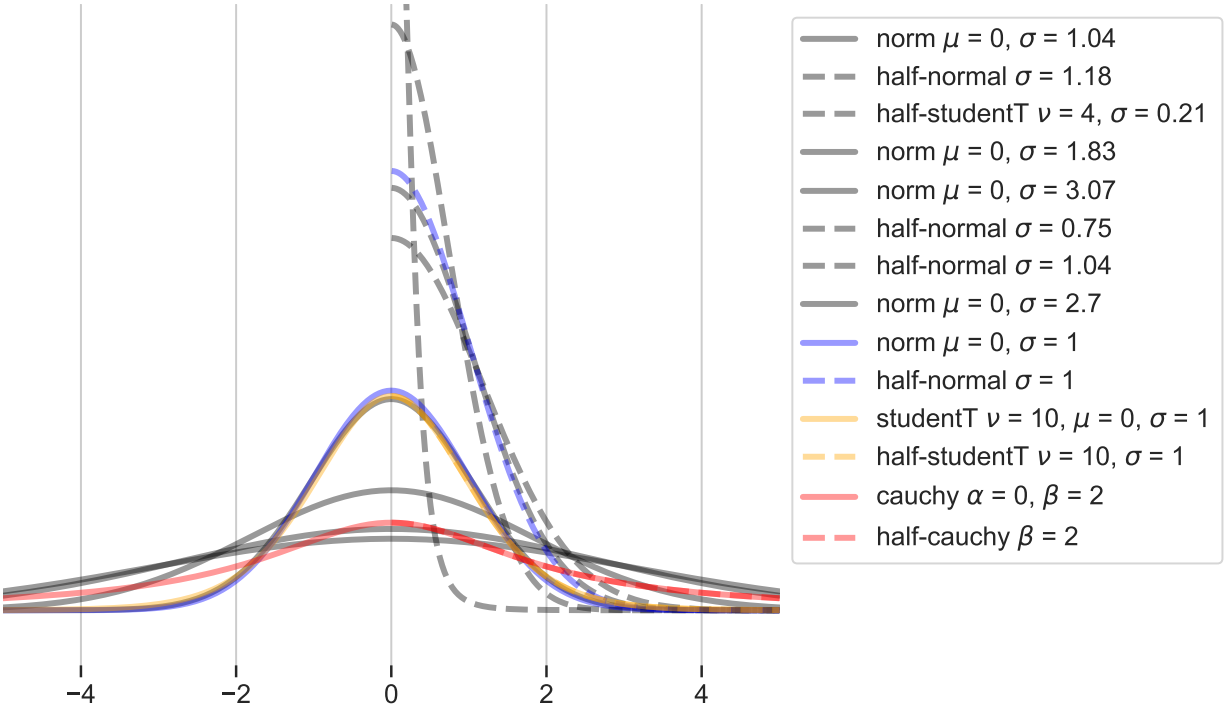

Figure 25: Plot of all parametrized prior distributions: Black ... default, other colors ... custom priors (see legend).

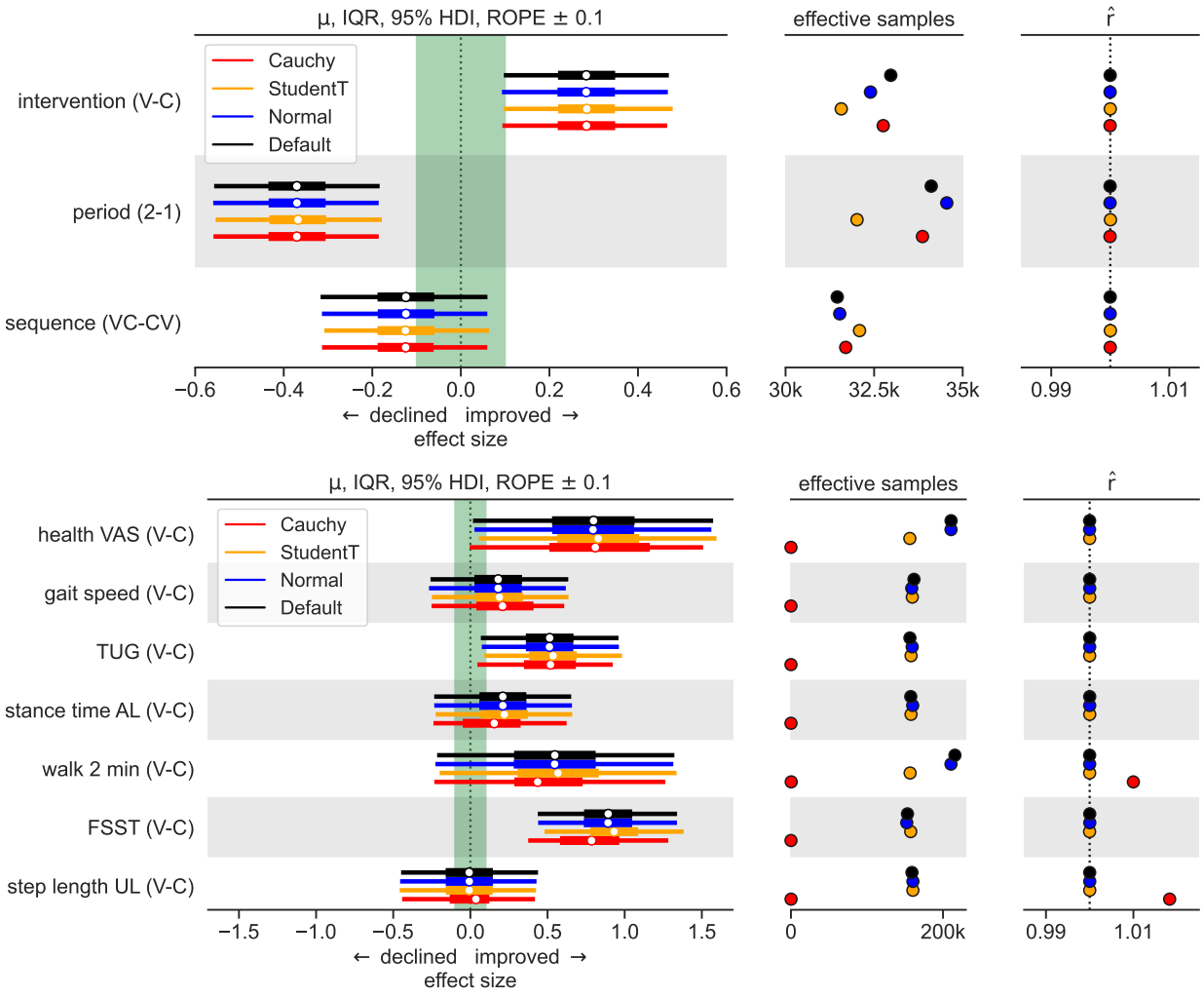

Figure 26: Global and test-level effect sizes for vibration (V) - control (C) comparison for different priors (see Figure 25). Markov chain Monte Carlo convergence statistic ... $\hat{r}$, affected leg ... AL, timed up and go test ... TUG, unaffected leg ... UL, visual analog scale ... VAS, four square step test ... FSST

Table 6: **Custom prior: normal distribution ($\mu = 0, \sigma = 1$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a global outcome or a clinical test is different between the cross-over sequences CV and VC, between the time periods 1 and 2, and between interventions vibration (V) and control (C) under H1. Sequence encodes the global carry-over effect, period the time order effect, and intervention the vibration effect. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	sequence VC_Δ-CV_Δ	period 2_Δ-1_Δ	intervention V_Δ-C_Δ	TUG V_Δ-C_Δ	walk 2 min V_Δ-C_Δ	health VAS V_Δ-C_Δ	FSST V_Δ-C_Δ	stance time AL V_Δ-C_Δ	step length UL V_Δ-C_Δ	gait speed V_Δ-C_Δ
BF_{01}	0.31	0.0	0.06	0.72	0.66	0.2	0.02	0.76	0.33	0.75
p (posteriors equal) [%]	23.6	0.4	5.3	41.8	39.9	16.3	1.9	43.1	24.6	43.0
p (large negative effect) [%]	0.0	0.0	0.0	0.0	0.03	0.0	0.0	0.0	0.02	0.0
p (medium negative effect) [%]	0.0	8.73	0.0	0.0	0.35	0.05	0.0	0.09	1.42	0.14
p (small negative effect) [%]	21.33	87.57	0.0	0.08	2.54	0.51	0.0	3.49	18.2	4.62
p (negative effect < small) [%]	21.33	96.3	0.0	0.08	2.92	0.56	0.0	3.58	19.64	4.77
effect size in ROPE [%]	38.87	0.28	2.85	3.02	7.81	2.7	0.03	22.72	34.15	25.47
effect size lower HDI 0.95%	-0.31	-0.56	0.09	0.07	-0.23	0.03	0.44	-0.23	-0.46	-0.27
effect size mean	-0.12	-0.37	0.28	0.51	0.55	0.8	0.89	0.21	-0.01	0.18
effect size upper HDI 0.95%	0.06	-0.19	0.47	0.96	1.32	1.56	1.34	0.66	0.43	0.62
p (positive effect > small) [%]	0.03	0.0	80.79	91.72	81.05	93.52	99.87	51.87	17.99	46.47
p (small positive effect) [%]	0.03	0.0	79.7	39.6	26.43	16.1	4.27	41.71	16.74	38.64
p (medium positive effect) [%]	0.0	0.0	1.09	41.66	28.58	27.99	30.14	9.68	1.23	7.56
p (large positive effect) [%]	0.0	0.0	0.0	10.46	26.03	49.43	65.46	0.49	0.02	0.27
p (posteriors differ) [%]	76.4	99.6	94.7	58.2	60.1	83.7	98.1	56.9	75.4	57.0
BF_{10}	3.24	232.62	17.91	1.39	1.51	5.14	51.73	1.32	3.06	1.33
prior p (reject difference) [%]	1.6	0.02	0.29	3.65	3.38	1.01	0.1	3.83	1.69	3.82
prior p (accept difference) [%]	85.41	7.55	51.48	93.18	92.66	78.7	26.86	93.5	86.14	93.47
model	global $\neq_{\mathrm{moderate}}$	global $\neq_{\mathrm{strong}}$ $-_{\mathrm{S}}$	global $\neq_{\mathrm{strong}}$	tests	tests	tests $\neq_{\mathrm{moderate}}$	tests $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	tests	tests $\neq_{\mathrm{moderate}}$	tests

TUG timed up & go test, VAS visual analog scale, FSST four square step test, and Δ indicates a difference of either $\mathrm{post} - \mathrm{pre}$ vibration or control test times where period 1 was earlier in time than period 2, or $\mathrm{left} - \mathrm{right}$ in stance time difference and step length difference.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm\mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

Table 7: **Custom prior: StudentT distribution ($\mu = 0, \nu = 10$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a global outcome or a clinical test is different between the cross-over sequences CV and VC, between the time periods 1 and 2, and between interventions vibration (V) and control (C) under H1. Sequence encodes the global carry-over effect, period the time order effect, and intervention the vibration effect. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	sequence VC_Δ-CV_Δ	period 2_Δ-1_Δ	intervention V_Δ-C_Δ	TUG V_Δ-C_Δ	walk 2 min V_Δ-C_Δ	health VAS V_Δ-C_Δ	FSST V_Δ-C_Δ	stance time AL V_Δ-C_Δ	step length UL V_Δ-C_Δ	gait speed V_Δ-C_Δ
BF_{01}	0.94	0.84	0.91	0.98	0.98	0.94	0.9	1.0	1.0	1.0
p (posteriors equal) [%]	48.4	45.7	47.6	49.4	49.4	48.4	47.4	50.0	49.9	49.9
p (large negative effect) [%]	0.0	0.0	0.0	0.0	0.03	0.0	0.0	0.0	0.02	0.0
p (medium negative effect) [%]	0.0	8.52	0.0	0.0	0.31	0.04	0.0	0.07	1.42	0.12
p (small negative effect) [%]	21.12	87.57	0.0	0.06	2.15	0.41	0.0	3.15	18.23	4.18
p (negative effect < small) [%]	21.12	96.08	0.0	0.06	2.49	0.46	0.0	3.21	19.67	4.3
effect size in ROPE [%]	38.52	0.22	2.88	2.52	7.17	2.26	0.02	22.0	34.33	24.55
effect size lower HDI 0.95%	-0.31	-0.55	0.1	0.09	-0.2	0.05	0.48	-0.22	-0.46	-0.25
effect size mean	-0.12	-0.37	0.28	0.54	0.57	0.83	0.93	0.22	-0.01	0.19
effect size upper HDI 0.95%	0.06	-0.18	0.48	0.98	1.34	1.6	1.38	0.66	0.43	0.64
p (positive effect > small) [%]	0.03	0.0	80.94	93.01	82.76	94.53	99.93	53.5	17.83	47.87
p (small positive effect) [%]	0.03	0.0	79.59	36.89	25.8	14.66	2.98	42.73	16.61	39.5
p (medium positive effect) [%]	0.0	0.0	1.35	43.86	29.25	26.95	25.32	10.26	1.21	8.02
p (large positive effect) [%]	0.0	0.0	0.0	12.25	27.71	52.92	71.62	0.51	0.01	0.35
p (posteriors differ) [%]	51.6	54.3	52.4	50.6	50.6	51.6	52.6	50.0	50.1	50.1
BF_{10}	1.06	1.19	1.1	1.02	1.02	1.07	1.11	1.0	1.0	1.0
prior p (reject difference) [%]	4.71	4.24	4.56	4.89	4.88	4.7	4.53	5.0	4.98	4.98
prior p (accept difference) [%]	94.69	94.12	94.52	94.89	94.88	94.68	94.49	95.0	94.98	94.98
model	global	global $-_{\mathrm{S}}$	global	tests	tests	tests	tests $+_{\mathrm{L}}$	tests	tests	tests

TUG timed up & go test, VAS visual analog scale, FSST four square step test, and Δ indicates a difference of either $\mathrm{post} - \mathrm{pre}$ vibration or control test times where period 1 was earlier in time than period 2, or $\mathrm{left} - \mathrm{right}$ in stance time difference and step length difference.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm \mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

Table 8: **Custom prior: Cauchy distribution ($\alpha = 0, \beta = 2$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a global outcome or a clinical test is different between the cross-over sequences CV and VC, between the time periods 1 and 2, and between interventions vibration (V) and control (C) under H1. Sequence encodes the global carry-over effect, period the time order effect, and intervention the vibration effect. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	sequence VC_Δ-CV_Δ	period 2_Δ-1_Δ	intervention V_Δ-C_Δ	TUG V_Δ-C_Δ	walk 2 min V_Δ-C_Δ	health VAS V_Δ-C_Δ	FSST V_Δ-C_Δ	stance time AL V_Δ-C_Δ	step length UL V_Δ-C_Δ	gait speed V_Δ-C_Δ
BF_{01}	0.31	0.0	0.06	0.05	0.72	0.17	0.0	0.05	0.0	0.94
p (posteriors equal) [%]	23.5	0.4	5.3	4.5	42.0	14.7	0.0	4.5	0.0	48.5
p (large negative effect) [%]	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.0	0.02	0.0
p (medium negative effect) [%]	0.0	8.74	0.0	0.0	0.41	0.08	0.0	0.09	1.22	0.12
p (small negative effect) [%]	21.38	87.52	0.0	0.14	2.69	0.75	0.0	3.53	15.72	4.5
p (negative effect < small) [%]	21.38	96.27	0.0	0.14	3.14	0.83	0.0	3.62	16.96	4.61
effect size in ROPE [%]	38.69	0.26	2.67	3.92	7.77	3.23	0.11	33.6	29.86	23.04
effect size lower HDI 0.95%	-0.31	-0.56	0.09	0.05	-0.23	-0.01	0.37	-0.24	-0.44	-0.25
effect size mean	-0.12	-0.37	0.28	0.5	0.49	0.81	0.8	0.16	0.01	0.2
effect size upper HDI 0.95%	0.06	-0.19	0.47	0.92	1.26	1.51	1.28	0.62	0.42	0.61
p (positive effect > small) [%]	0.02	0.0	80.79	89.98	81.02	92.33	99.59	43.15	15.95	51.23
p (small positive effect) [%]	0.02	0.0	79.66	37.07	36.72	16.46	7.15	35.32	14.82	45.04
p (medium positive effect) [%]	0.0	0.0	1.13	45.85	24.5	25.06	44.46	7.49	1.11	5.98
p (large positive effect) [%]	0.0	0.0	0.0	7.05	19.8	50.82	47.98	0.34	0.02	0.21
p (posteriors differ) [%]	76.5	99.6	94.7	95.5	58.0	85.3	100.0	95.5	100.0	51.5
BF_{10}	3.26	225.9	18.04	21.04	1.38	5.81	227459.77	21.17	5303.54	1.06
prior p (reject difference) [%]	1.59	0.02	0.29	0.25	3.66	0.9	nan	0.25	0.54	4.72
prior p (accept difference) [%]	85.34	7.76	51.3	47.45	93.21	76.59	44.72	47.3	0.36	94.71
model	global $\neq_{\mathrm{moderate}}$	global $\neq_{\mathrm{strong}}$ $-_{\mathrm{S}}$	global $\neq_{\mathrm{strong}}$	tests $\neq_{\mathrm{strong}}$	tests	tests $\neq_{\mathrm{moderate}}$	tests $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	tests $\neq_{\mathrm{strong}}$	tests $\neq_{\mathrm{strong}}$	tests

TUG timed up & go test, VAS visual analog scale, FSST four square step test, and Δ indicates a difference of either $\mathrm{post} - \mathrm{pre}$ vibration or control test times where period 1 was earlier in time than period 2, or $\mathrm{left} - \mathrm{right}$ in stance time difference and step length difference.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm\mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

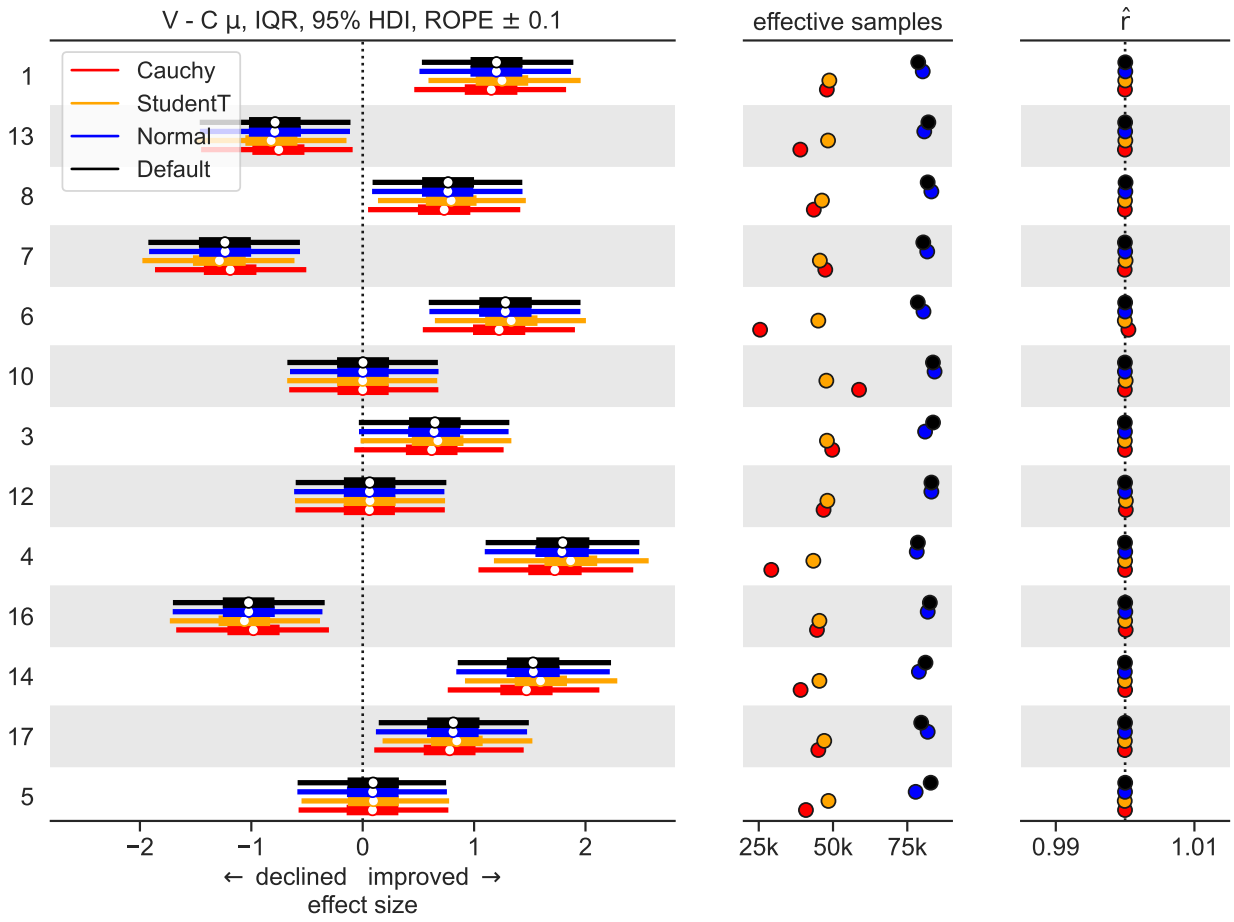

Figure 27: ID-level effect sizes for vibration (V) - control (C) comparison for different priors (see Figure 25). Markov chain Monte Carlo convergence statistic ... $\hat{r}$, affected leg ... AL, timed up and go test ... TUG, unaffected leg ... UL, visual analog scale ... VAS, four square step test ... FSST

Table 9: **Custom prior: normal distribution ($\mu = 0, \sigma = 1$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a participant is different between the interventions vibration (V) and control (C) under H1. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	ID 7 V_Δ-C_Δ	ID 16 V_Δ-C_Δ	ID 13 V_Δ-C_Δ	ID 10 V_Δ-C_Δ	ID 12 V_Δ-C_Δ	ID 5 V_Δ-C_Δ	ID 3 V_Δ-C_Δ	ID 8 V_Δ-C_Δ	ID 17 V_Δ-C_Δ	ID 1 V_Δ-C_Δ	ID 6 V_Δ-C_Δ	ID 14 V_Δ-C_Δ	ID 4 V_Δ-C_Δ
BF_{01}	0.03	0.06	0.33	0.99	1.0	0.96	0.21	0.47	0.12	0.02	0.02	0.0	0.0
p (posteriors equal) [%]	2.8	5.6	25.0	49.8	50.0	49.0	17.1	31.9	10.7	2.0	1.5	0.1	0.0
p (large negative effect) [%]	89.66	74.07	48.8	0.93	0.64	0.44	0.0	0.0	0.0	0.0	0.0	0.0	0.0
p (medium negative effect) [%]	8.7	19.58	30.96	6.17	4.42	3.82	0.05	0.01	0.01	0.0	0.0	0.0	0.0
p (small negative effect) [%]	1.5	5.58	15.83	20.67	17.16	15.64	0.62	0.26	0.16	0.0	0.0	0.0	0.0
p (negative effect < small) [%]	99.86	99.23	95.58	27.78	22.22	19.9	0.68	0.27	0.17	0.0	0.0	0.0	0.0
effect size in ROPE [%]	0.05	0.29	1.81	22.92	22.78	22.33	4.22	2.12	1.55	0.08	0.03	0.0	0.0
effect size lower HDI 0.95%	-1.92	-1.71	-1.46	-0.65	-0.62	-0.59	-0.03	0.08	0.12	0.51	0.6	0.84	1.1
effect size mean	-1.24	-1.02	-0.79	0.0	0.06	0.09	0.64	0.76	0.81	1.2	1.28	1.53	1.79
effect size upper HDI 0.95%	-0.56	-0.36	-0.11	0.68	0.73	0.76	1.31	1.43	1.48	1.87	1.95	2.22	2.48
p (positive effect > small) [%]	0.0	0.02	0.24	28.2	34.32	37.38	90.02	94.89	96.11	99.79	99.9	99.99	100.0
p (small positive effect) [%]	0.0	0.02	0.23	20.95	24.24	25.84	24.14	17.18	14.52	2.02	1.14	0.14	0.01
p (medium positive effect) [%]	0.0	0.0	0.01	6.28	8.52	9.62	33.56	31.87	30.36	10.25	7.18	1.69	0.26
p (large positive effect) [%]	0.0	0.0	0.0	0.97	1.57	1.92	32.33	45.84	51.24	87.52	91.58	98.16	99.73
p (posteriors differ) [%]	97.2	94.4	75.0	50.2	50.0	51.0	82.9	68.1	89.3	98.0	98.5	99.9	100.0
BF_{10}	34.27	16.9	3.0	1.01	1.0	1.04	4.83	2.13	8.33	48.48	63.73	1575.05	6506.1
prior p (reject difference) [%]	0.15	0.31	1.73	4.97	4.99	4.81	1.08	2.41	0.63	0.11	0.08	0.05	0.81
prior p (accept difference) [%]	35.67	52.92	86.37	94.97	94.99	94.8	79.73	89.9	69.52	28.16	22.97	1.19	0.29
model	IDs $\neq_{\mathrm{strong}}$ $-_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $-_{\mathrm{L}}$	IDs $-_{\mathrm{M}}$	IDs	IDs	IDs	IDs $\neq_{\mathrm{moderate}}$	IDs	IDs $\neq_{\mathrm{moderate}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$	IDs $\neq_{\mathrm{strong}}$ $+_{\mathrm{L}}$

Δ indicates a difference of between control and vibration.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm \mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

Table 10: **Custom prior: StudentT distribution ($\mu = 0, \nu = 10$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a participant is different between the interventions vibration (V) and control (C) under H1. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	ID 7 V_Δ-C_Δ	ID 16 V_Δ-C_Δ	ID 13 V_Δ-C_Δ	ID 10 V_Δ-C_Δ	ID 12 V_Δ-C_Δ	ID 5 V_Δ-C_Δ	ID 3 V_Δ-C_Δ	ID 8 V_Δ-C_Δ	ID 17 V_Δ-C_Δ	ID 1 V_Δ-C_Δ	ID 6 V_Δ-C_Δ	ID 14 V_Δ-C_Δ	ID 4 V_Δ-C_Δ
BF_{01}	0.89	0.89	0.96	1.0	1.0	1.0	0.92	0.99	0.94	0.92	0.92	0.75	0.79
p (posteriors equal) [%]	47.2	47.1	49.1	50.0	50.0	50.0	47.9	49.9	48.4	47.9	47.9	42.9	44.1
p (large negative effect) [%]	91.93	77.72	52.55	0.96	0.58	0.39	0.0	0.0	0.0	0.0	0.0	0.0	0.0
p (medium negative effect) [%]	6.94	17.21	29.58	6.37	4.36	3.49	0.03	0.01	0.0	0.0	0.0	0.0	0.0
p (small negative effect) [%]	1.06	4.51	14.13	20.6	17.36	15.17	0.55	0.21	0.13	0.0	0.0	0.0	0.0
p (negative effect < small) [%]	99.93	99.44	96.26	27.94	22.31	19.05	0.59	0.22	0.13	0.0	0.0	0.0	0.0
effect size in ROPE [%]	0.02	0.18	1.59	22.78	22.22	22.19	3.58	1.54	1.19	0.04	0.02	0.0	0.0
effect size lower HDI 0.95%	-1.98	-1.73	-1.5	-0.68	-0.61	-0.55	-0.02	0.14	0.18	0.59	0.65	0.92	1.18
effect size mean	-1.29	-1.06	-0.82	0.0	0.06	0.1	0.67	0.79	0.85	1.25	1.33	1.6	1.87
effect size upper HDI 0.95%	-0.61	-0.38	-0.15	0.67	0.74	0.78	1.33	1.46	1.52	1.96	2.0	2.29	2.57
p (positive effect > small) [%]	0.0	0.01	0.17	28.23	34.76	37.99	91.41	95.97	97.06	99.9	99.96	99.99	100.0
p (small positive effect) [%]	0.0	0.01	0.17	20.7	24.64	26.45	22.03	15.07	12.72	1.35	0.7	0.1	0.0
p (medium positive effect) [%]	0.0	0.0	0.01	6.45	8.47	9.67	33.84	31.55	29.11	8.44	5.55	1.08	0.13
p (large positive effect) [%]	0.0	0.0	0.0	1.08	1.66	1.86	35.55	49.35	55.24	90.11	93.71	98.81	99.86
p (posteriors differ) [%]	52.8	52.9	50.9	50.0	50.0	50.0	52.1	50.1	51.6	52.1	52.1	57.1	55.9
BF_{10}	1.12	1.12	1.04	1.0	1.0	1.0	1.09	1.01	1.06	1.09	1.09	1.33	1.27
prior p (reject difference) [%]	4.49	4.47	4.83	5.0	5.0	5.0	4.62	4.97	4.71	4.62	4.62	3.8	3.99
prior p (accept difference) [%]	94.43	94.41	94.83	95.0	95.0	95.0	94.59	94.97	94.69	94.59	94.59	93.45	93.75
model	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs
	$-_{\mathrm{L}}$	$-_{\mathrm{L}}$	$-_{\mathrm{L}}$					$+_{\mathrm{M}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$

Δ indicates a difference of between control and vibration.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm \mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

Table 11: **Custom prior: Cauchy distribution ($\alpha = 0, \beta = 2$)** Effect size (Hedges' g) and Bayes factor testing if the $\mathrm{post} - \mathrm{pre}$ difference within a participant is different between the interventions vibration (V) and control (C) under H1. The probability that the effect is greater than 0.2 can be interpreted similarly to statistical power. Bayesian multiple comparisons do (usually) not require adjustments [5].

	ID 7 V_Δ-C_Δ	ID 16 V_Δ-C_Δ	ID 13 V_Δ-C_Δ	ID 10 V_Δ-C_Δ	ID 12 V_Δ-C_Δ	ID 5 V_Δ-C_Δ	ID 3 V_Δ-C_Δ	ID 8 V_Δ-C_Δ	ID 17 V_Δ-C_Δ	ID 1 V_Δ-C_Δ	ID 6 V_Δ-C_Δ	ID 14 V_Δ-C_Δ	ID 4 V_Δ-C_Δ
BF_{01}	1.0	0.01	0.01	1.0	0.99	1.0	0.32	1.0	0.01	0.0	0.0	0.0	0.0
p (posteriors equal) [%]	49.9	1.2	0.8	50.0	49.7	49.9	24.0	49.9	0.7	0.0	0.0	0.0	0.0
p (large negative effect) [%]	86.67	69.9	44.79	0.97	0.57	0.45	0.0	0.0	0.0	0.0	0.0	0.0	0.0
p (medium negative effect) [%]	10.89	21.64	32.27	6.07	4.44	3.86	0.04	0.02	0.01	0.0	0.0	0.0	0.0
p (small negative effect) [%]	2.25	7.16	17.68	20.6	17.37	15.59	0.75	0.34	0.19	0.01	0.0	0.0	0.0
p (negative effect < small) [%]	99.81	98.7	94.73	27.64	22.37	19.9	0.8	0.36	0.21	0.01	0.0	0.0	0.0
effect size in ROPE [%]	0.07	0.54	2.28	23.05	22.34	22.22	4.74	2.74	1.88	0.11	0.07	0.01	0.0
effect size lower HDI 0.95%	-1.86	-1.67	-1.45	-0.66	-0.6	-0.58	-0.08	0.05	0.1	0.46	0.54	0.76	1.04
effect size mean	-1.19	-0.98	-0.76	0.0	0.06	0.09	0.62	0.73	0.78	1.15	1.22	1.47	1.73
effect size upper HDI 0.95%	-0.51	-0.3	-0.09	0.68	0.74	0.77	1.27	1.41	1.45	1.83	1.91	2.12	2.43
p (positive effect > small) [%]	0.0	0.03	0.25	28.28	34.11	37.15	88.89	93.65	95.34	99.73	99.83	99.99	100.0
p (small positive effect) [%]	0.0	0.03	0.24	20.96	24.1	25.5	25.24	19.05	16.26	2.77	1.68	0.29	0.02
p (medium positive effect) [%]	0.0	0.0	0.01	6.33	8.46	9.76	33.65	32.38	31.36	12.52	9.3	2.51	0.42
p (large positive effect) [%]	0.0	0.0	0.0	0.99	1.56	1.89	30.0	42.23	47.73	84.45	88.85	97.19	99.55
p (posteriors differ) [%]	50.1	98.8	99.2	50.0	50.3	50.1	76.0	50.1	99.3	100.0	100.0	100.0	100.0
BF_{10}	1.0	80.39	118.76	1.0	1.01	1.0	3.16	1.0	141.39	42321.63	1275227.82	844603.26	12119651.75
prior p (reject difference) [%]	4.99	0.07	0.04	4.99	4.95	4.98	1.64	4.99	0.04	34.74	nan	nan	nan
prior p (accept difference) [%]	94.99	19.12	13.79	94.99	94.95	94.98	85.72	94.99	11.85	0.67	nan	nan	nan
model	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs	IDs
		$\neq_{\mathrm{strong}}$	$\neq_{\mathrm{strong}}$				$\neq_{\mathrm{moderate}}$		$\neq_{\mathrm{strong}}$	$\neq_{\mathrm{strong}}$	$\neq_{\mathrm{strong}}$	$\neq_{\mathrm{strong}}$	$\neq_{\mathrm{strong}}$
	$-_{\mathrm{L}}$	$-_{\mathrm{L}}$								$+_{\mathrm{L}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$	$+_{\mathrm{L}}$

Δ indicates a difference of between control and vibration.

Bayes factor (BF) from uninformed prior odds = 1 reflecting posterior distribution equality under H0: BF $\in [1, 3)$ weak, BF $\in [3, 10)$ moderate, BF $\in [10, \mathrm{inf})$ strong. BF_{10} reflects evidence for H1, while BF_{01} reflects evidence for H0.

Posterior probability p of effect size Hedge's g in defined regions: ($g \in (\pm \mathrm{inf}, \pm 0.8]$ large (L), $g \in (\pm 0.8, \pm 0.5]$ medium (M), ($g \in \pm 0.5, \pm 0.2]$ small (S), $g \in [-0.1, 0.1]$ region of practical equivalence (ROPE)) and the lower, mean, and upper 95% highest density interval (HDI).

Symbols encode BF evidence for equal posterior distributions with $=$ and different posterior distributions with $\neq$, and effect size ROPE+HDI testing with $+$ for a discernible positive effect (the lower HDI bound is above the upper ROPE limit) and with $-$ for a discernible negative effect (the upper HDI bound is below the lower ROPE limit).

1.5.D BFs and model probabilites

Please see Section 1.4.D

1.5.E Decisions

Decisions would not change. Please see Section 1.5.C for more details.

1.6 Make it reproducible

1.6.A Software and installation

Please see Section 1.2.A.

1.6.B Software version details

Please see Section 1.2.A.

1.6.C Script and data

- The related jupyter notebook and data are available via gitfront https://gitfront.io/r/haripen/6ZQ4pngRwEWa/Safety-Efficacy-vibrotactile-feedback-patients-lower-limb-amputation/.
- The custom functions used are also available via gitfront https://gitfront.io/r/haripen/VwUFNvUqe1Pa/bambi-helper-funs/blob/bambi_helper_funs.py and are loaded when using the jupyter notebook.

1.6.D Readable for humans

The jupyter notebook and packages should be sufficiently annotated such that an intermediate Python user can understand the steps.

1.6.E All auxiliary files

No other files than the data and the custom functions are required.

1.6.F Runs as posted

The jupyter notebook was uploaded in a single run while loading data and packages from remote repositories.

1.6.G MCMC chains for time-intensive runs

The analysis with the default prior should run in a reasonable time on modern laptops (i.e., MacOS 12.6.2, 2021 MacBook Pro 14", chip Apple M1 Pro, unified memory 32 GB.).

1.6.H Reproducible MCMC

All pseudo-random number generators were set to 1234.

2 Calculation of ICC, MDC, and SEM

To check the reliability of the outcome measures that were repeated three times in each session we calculated the intraclass correlation coefficient (ICC), the minimally detectable change (MDC), and the standard error of measurement (SEM). We used a frequentist mixed linear model with formula $\mathrm{score} \sim \mathrm{times} + \mathrm{tries} + (1|\mathrm{ID})$ to calculate the covariance and error variance from tests repeated three times (Python: 3.11.0, statsmodels (mixedlm): 0.13.5, numpy: 1.23.5, scipy: 1.9.3). The intraclass correlation coefficient ($\mathrm{ICC}_{\mathrm{data}}$) results from dividing the random effects covariance matrix (Ψ) by the sum of Ψ and the error variance (σ^2_{err})

$$\mathrm{ICC}_{\mathrm{data}} = \frac{\Psi}{\Psi + \sigma^2_{\mathrm{err}}}. \tag{3}$$

The standard error of measurment ($\mathrm{SEM}_{\mathrm{data}}$) is calculated from the square root of the linear mixed effects models' estimated σ^2_{err} for the tests repeated three times [9, 10]

$$\mathrm{SEM}_{\mathrm{data}} = \sqrt{\sigma^2_{\mathrm{err}}}. \tag{4}$$

For VAS and 2 min walking time $\mathrm{SEM}_{\mathrm{lit}}$ may be calculated from literature providing the intraclass correlation coefficient $\mathrm{ICC}_{\mathrm{lit}}$, mean of the first trial, standard deviation of the first trial (SD), and the sample size using

$$\mathrm{SEM}_{\mathrm{lit}} = \mathrm{SD} \cdot \sqrt{1 - \mathrm{ICC}_{\mathrm{lit}}}. \tag{5}$$

Finally, the minimal detectable change (MDC) was calculated using

$$\mathrm{MDC} = \mathrm{t} \cdot \mathrm{SEM} \cdot \sqrt{2}, \tag{6}$$

where t is the 95% confidence interval two-sided Student t-distribution value adjusted for the sample size [11]. Table 12 shows the reliability coefficients of the tests that were repeated three times at each measurement session of this study.

Table 12: Reliability and measurement errors of tests repeated three times at each testing session [9, 10, 12].

	$\mathrm{ICC}_{3,1}$	MDC	MDC [%]	SEM	SEM [%]
TUG [s]	0.91	2.86	40.37	0.94	13.21
FSST [s]	0.92	2.59	38.39	0.85	12.57
stance time affected side [s]	0.52	0.24	36.9	0.08	12.08
step length unaffected side [cm]	0.81	13.62	20.1	4.46	6.58
gait speed [km/h]	0.73	1.47	31.8	0.48	10.41

$\mathrm{ICC}_{3,1}$: 2-way random intraclass correlation coefficient, MDC: minimal detectable change, SEM: standard error of measurement, TUG: timed up & go test, FSST: four square step test.

References

[1] J K Kruschke. Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5(10):1282–1291, 2021. doi: 10.1038/s41562-021-01177-7.

[2] D McNeish. On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling*, 23(5): 750–773, 2016. doi: 10.1080/10705511.2016.1186549.

[3] T Capretto, C Piho, R Kumar, J Westfall, T Yarkoni, and O A Martin. Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103(15), 2022. doi: 10.18637/jss.v103.i15.

[4] C Y Lim and J In. Considerations for crossover design in clinical study. *Korean Journal of Anesthesiology*, 74(4): 293–299, 2021. doi: 10.4097/KJA.21165.

[5] A Gelman, J Hill, and M Yajima. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012. doi: 10.1080/19345747.2011.618213.

[6] J K Kruschke. Rejecting or Accepting Parameter Values in Bayesian Estimation. 1(2):270–280, 2018. doi: 10.1177/2515245918771304.

[7] D P Todd. Sun-spots. *Science*, ns-4(93):453–453, nov 1884. ISSN 0036-8075. doi: 10.1126/science.ns-4.93.453.

[8] W J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[9] J P Weir. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1):231–240, 2005. doi: 10.1519/15184.1.

[10] M Geiger, A Supiot, D Pradon, M C Do, R Zory, and N Roche. Minimal detectable change of kinematic and spatiotemporal parameters in patients with chronic stroke across three sessions of gait analysis. *Human Movement Science*, 64(April 2018):101–107, 2019. doi: 10.1016/j.humov.2019.01.011.

[11] H Beckerman, M E Roebroeck, G J Lankhorst, J G Becher, P D Bezemer, and A L M Verbeek. Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, 10(7):571–578, 2001. doi: 10.1023/A:1013138911638.

[12] P E Shrout and J L Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2): 420–428, 1979. doi: 10.1037/0033-2909.86.2.420.