Dual task turning in place: a reliable, valid and responsive outcome measure of freezing of gait

Nicholas D'Cruz, PhD,[1] Jana Seuthe, PhD,[2,3] Clara de Somer, MSc,[1] Femke Hulzinga, MSc,[1] Pieter Ginis, PhD,[1] Christian Schlenstedt, PhD,[2,3] Alice Nieuwboer, PhD[1]

[1] *KU Leuven, Department of Rehabilitation Sciences, Neurorehabilitation Research Group, B-3000 Leuven, Belgium*

[2] *Christian-Albrechts-University (CAU) Kiel, University Hospital Schleswig-Holstein, Department of Neurology, Kiel, Germany*

[3] *MSH Medical School Hamburg, Section Exercise and Neurosciences, Hamburg, Germany*

Corresponding Author: Nicholas D'Cruz, Department of Rehabilitation Sciences, KU Leuven, Tervuursevest 101 Bus 1501, B-3001 Leuven, Belgium. Ph: +3216376003 Email: nicholas.dcruz@kuleuven.be

Word count: Abstract: 248 Manuscript: 3849

Figures: 2 Tables: 3 Supplementary tables: 5 Supplementary figures: 3

Running Title: Objective measurement of FOG

Keywords: Gait disorders, Dual tasking, Outcome measures, Split-belt treadmill, Interrater reliability

Competing interests: The authors declare that there are no potential conflicts of interest in regard to this work.

Funding sources: Research grant from the Jacques and Gloria Gossweiler Foundation

Abstract

Background: Freezing of gait (FOG) is a complex symptom in Parkinson's disease (PD) which is both elusive to elicit and varied in its presentation. These complexities present a challenge to measuring FOG in a sensitive and reliable way, precluding therapeutic advancement.

Objective: We investigated the reliability, validity and responsiveness of manual video-annotations of the turning in place task and compared it to the sensor-based FOG ratio.

Methods: Forty-five optimally medicated people with PD and FOG performed rapid alternating 360° turns without and with an auditory stroop dual task, thrice over two consecutive days. Tasks were recorded with video and inertial sensors placed on the lower back and shins. Interrater reliability between three raters, criterion validity with self-reported FOG, and responsiveness to single-session split-belt treadmill (SBT) training were investigated and contrasted with the sensor-based FOG ratio.

Results: Visual ratings showed excellent agreement between raters for the percent time frozen (%TF) (ICC = 0.99), the median duration of a FOG episode (ICC = 0.90) and the number of FOG episodes (ICC = 0.86). Dual tasking improved the sensitivity and validity of visual FOG ratings resulting in increased FOG detection, criterion validity with self-reported FOG ratings, and responsiveness to a short SBT intervention. The sensor-based FOG ratio on the other hand, showed complex FOG presentation-contingent relationships with visual and self-reported FOG ratings, and limited responsiveness to SBT training.

Conclusions: Manual video-annotations of FOG during dual task turning in place generate reliable, valid, and sensitive outcomes for investigating therapeutic effects on FOG.

Introduction

Freezing of gait (FOG) is a disabling symptom for persons with Parkinson's disease (PwPD). FOG is defined as the "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk."[1] Due to these unpredictable interruptions to movement, PwPD who experience FOG are at a higher risk of falling[2], and its unwanted consequences[3]. In fact, a staggering 61% of falls in PD are directly attributable to FOG[4]. Rehabilitation in conjunction with medical therapies[5,6] aimed at reducing FOG severity or delaying its onset are therefore urgently needed. A novel and relevant training for FOG involves split-belt treadmill (SBT) whereby gait perturbations are imposed by driving the belts at different speeds. Repeated exposure to such perturbations, even within one session, has been found to reduce gait asymmetry, enhance adaptation to asymmetry and improve turning speed[7–9], but it is unclear whether it also reduces FOG.

In order to test intervention effects on FOG, reliable, valid and sensitive measures of FOG severity are required. Quantification of FOG severity is frequently performed with subjective rating scales such as the Freezing of Gait Questionnaire and the New Freezing of Gait Questionnaire (NFOGQ)[10,11]. Although these measures have shown adequate validity to detect the presence of FOG, recent work suggests that retest reliability of subjective measures is poor,[12] rendering them less suitable for measuring FOG severity. Given the sensitivity of subjective measures to recall and expectation bias, objective measures of FOG severity are increasingly being sought after[13,14].

Objective quantification of FOG uses FOG provoking tasks[15], often under stress[16], to elicit freezing episodes which can be quantified using manual or automated methods. Manually labelling FOG episodes from video recordings of Timed Up and Go tasks to quantify the percentage of the task time with FOG (percent time frozen) is considered the gold-standard for measuring FOG severity[17]. However, poor sensitivity to FOG and limited response to both

medication and training[18] have raised questions about its utility as an outcome for clinical trials. Subsequent work has shown that full and fast turning in place is more sensitive in eliciting FOG[19,20], and this sensitivity can be further improved by dual tasking[21]. So far, determining criteria for annotating FOG episodes while turning in place and their validation as an outcome of FOG severity has not been undertaken. FOG annotation is challenging during alternating fast 360° turns, as these do not show the same stepping patterns as straight-line walking tasks. This is important, as reliability of labelling FOG episodes will have an impact on the measurement error and subsequent usefulness of the FOG metric as an outcome for intervention.

More recently, an automated sensor-based metric has been proposed to detect FOG during turning in place, using automated algorithms based on temporal decomposition of the leg movement signal[20]. The FOG ratio has been used to quantify FOG severity in several observational and interventional studies[22–24] and has undergone initial validitation[22]. However, the FOG ratio may be affected by turning fragmentation and less affected by periods of complete absence of movement[25], thus reducing its FOG sensitivity and thereby, its responsiveness.

In this study, we developed standardized criteria for visually rating FOG episodes from 360° turning in place videos and investigated the interrater reliability when these were applied to an optimally medicated representative sample of PwPD and FOG. As a secondary objective, we investigated criterion validity through associations with a validated subjective rating scale – the NFOGQ, as well as automated sensor-based ratings of FOG severity. Finally, we compared the responsiveness of the visual-rated and sensor-based metrics to the effects of a split-belt treadmill intervention. We expected that visual-rated outcomes would show validity with both subjective and sensor-based FOG measures, but owing to higher FOG-sensitivity, would be more responsive than sensor-based outcomes of FOG severity.

Methods

2.1 Participants

Included participants were a subset of a larger multi-center study to investigate the short-term effects of split-belt treadmill training on gait, turning and adaptation (ClinicalTrials.gov NCT03725215). Forty-five people with Parkinson's disease and freezing of gait (PD+FOG) were included in this study. Eligibility criteria included PD diagnosis based on the UK Brain Bank Criteria, presence of FOG based on self-reported answer to the question "did you experience freezing of gait in the past month", and the ability to walk unassisted for at least 5 minutes. Exclusion criteria included marked cognitive impairment (MMSE $\leq$ 24), cardiovascular risk for exercise, musculoskeletal disorders affecting gait, or recent changes in Parkinson's medication or deep brain stimulation settings (< 1 month). Ethical approval was obtained from the respective Institutional Review Boards and all participants provided written informed consent prior to enrollment in the study. Measurements and training were performed in the optimally medicated state and repeated measurements were standardized in relation to medication intake.

2.2 Study design

The study utilized a randomized parallel design with one control and three active arms. FOG provoking tests were performed at three moments over two days – pre and post intervention on day one, and once the following day (retention). In addition, clinical questionnaires were administered once to characterize participants' cognitive, balance and motor disease severity. Subjective FOG severity was characterized using the NFOGQ, which enquires about the severity and impact of FOG in the past month[11].

2.3 FOG provoking task

Based on previous work[22], the turning in place task for 60 seconds was performed to provoke FOG. Participants were instructed to turn in place as quickly and as safely possible, alternating direction after each full turn, and to take steps rather than pivot on one leg. The task was performed without and with a cognitive dual task (in that order), namely the auditory stroop task, delivered through a wireless headset. Inertial measurement units (Opals, APDM, Portland, USA) were placed on the shins and lower back of the participants to capture objective turning and FOG metrics. One video camera providing a single-angle (diagonal to starting position) neck-to-foot view of the participant captured the trial for subsequent rating.

2.4 Visual FOG rating

Three raters (CDS, JS and ND) annotated the video recordings with the ELAN toolbox (version 5.8, Max Planck Institute for Psycholinguistics, The Netherlands), based on recent recommendations[26]. Raters were blinded to the time point, and whether it was single or dual task (by muting sound). Criteria and definitions for all labels were developed in two iterations. In the first iteration, using established criteria, 40 turning in place trials were rated between two pairs of raters (20 trials for rater A and B, and 20 trials for rater B and C) and interrater reliability was evaluated. Despite fair to excellent interrater reliability (ICC > 0.93, 95% CI between 0.78 and 0.98 for both rater pairs), limits of agreement were large (18.1% and 37.7% for the two rater pairs), therefore the decision was made to revise the criteria and re-annotate the videos. Chief sources of variation in ratings were attributed to difficulty in labelling the start and end of the task when it began or ended with akinetic FOG, as well as varied interpretation of the "ineffective step" to label the start and end of the FOG episodes. These criteria were therefore changed between iterations (Table 1). For the second iteration, 20 trials with the largest interrater differences in the first iteration on the percent time frozen were rated by all raters, and interrater reliability was reassessed. The remaining trials were randomly distributed among raters and outcomes were calculated. Percent time frozen (%TF) was the

primary outcome, with number of FOG episodes and median duration of a FOG episode as secondary outcomes. To minimize potential overfitting of criteria development to the 40 videos earlier labelled, a random sample of 10 trials was re-evaluated after 15 months.

2.5 Sensor-based FOG rating

The FOG ratio as described by Mancini et al. (2017) was calculated from the power spectral density (PSD) of the anterior-posterior acceleration signal in the shin sensors (sampled at 128Hz). Using a four-second Hanning window, the PSD was calculated with the Welch method. The ratio of the square of the power within the freezing band (3 – 8 Hz) to the square of the power within the movement band (0.5 to 3 Hz), averaged over the trial and over the two legs, gave the FOG ratio. Further, the FOG ratio calculated from the mediolateral (task movement direction) acceleration signal was also obtained. Objective measures of turning performance were calculated from the lower back sensor, including mean and peak turning speed (yaw angular velocity) and mediolateral jerkiness (measure of turning fluidity)[21].

2.6 Training intervention

The training comprised of one session of 30 minutes of walking on either a split-belt treadmill with both belts going at the same speed (tied-belt mode – TBT) or with each belt going at a different speed (split-belt mode – SBT). Three SBT conditions were used to compare effectiveness of the various modes, and participants were randomly assigned to receive TBT or any one of the SBT conditions. Two SBT conditions differed in the extent of speed reduction on the slow belt (25% or 50% slower) and one condition switched between the two speeds.

Statistical analysis

3.1 Interrater reliability analysis

Interrater reliability was assessed using the intra-class correlation coefficient (ICC) with a two-way random effects analysis (random trials, random raters) for the absolute agreement between raters for a single measurement (ICC(2,1)). Wilcoxon signed-rank tests, Bland-Altman plots and limits of agreement were investigated for systematic bias between pairs of raters.

3.2 Criterion validity of visual FOG rating

To investigate criterion validity of the visual ratings, we performed correlation analyses with the NFOG-Q and FOG ratio at pre-training. Distribution of the FOG outcomes were assessed with histograms and found to be highly skewed to the right, so spearman rank correlation was performed. Non-parametric bootstrapping (1000 resamples, unrestricted random sampling) was used to estimate confidence intervals.

3.3 Responsiveness to treatment

To investigate if the FOG metrics were responsive to treatment, we tested whether the visual-rated or sensor-based metrics demonstrated treatment effects. Constrained longitudinal data analysis implemented in a linear mixed model framework was applied to investigate changes within and between TBT and any SBT condition from pre-training to retention[9]. Data were transformed with an inverse hyperbolic sine function to reduce skewness while allowing inclusion of zero scores. Normality of model residuals were visually assessed with histograms and QQ plots.

3.4 Exploratory analysis – Sensor-based turning metrics related to FOG over time

To investigate sensor-based turning metrics related to visual-rated FOG, we performed non-parametric repeated measures correlation between visual-rated FOG and various turning metrics. 1000 random permutations were used to obtain null distribution of the resulting Z score and calculation of p-values[27].

Results

4.1 Participant characteristics and missing data

Forty-five PD+FOG, Hoehn & Yahr stages I - IV were included in this study. Participants varied in their ages (mean: 68.6 years, range: 48 – 86), disease duration (mean: 12.8 years, range: 1 – 38) and self-reported freezing severity (NFOGQ mean: 16.3, range: 6 - 29) (Supplementary Table 1). No significant differences were found for any of the cognitive, balance, disease-related or training-intensity measures between the four training groups (Supplementary Table 2). Due to technical difficulties with video capture, seven participants did not have recordings available, hence visual ratings are reported on the 38 remaining participants.

4.2 Interrater reliability of visual FOG rating

Twenty videos from with varying amounts and presentations of FOG were rated to assess the inter-rater reliability (mean and ranges of rater means – number of episodes: 4.55 (2 – 13); median duration of an episode in seconds: 6.01 (0.4 - 27); %TF: 40.5 (1.35 - 90.1)). Intraclass correlation coefficient for *absolute agreement* of %TF for a single rater (ICC (2,1)) using a 2-way mixed effects model was 0.993 (95% CI: 0.986 – 0.997). Reliability was slightly lower for other FOG outcomes, with the number of FOG episodes showing the lowest reliability. Importantly, the standard error of measurement was less than 3% for the %TF (Table 2).

Bland-Altman plots revealed no systematic error across FOG severity, although scores tended to be more spread out around medium severity (35 – 60 %TF) trials. Only number of episodes showed a statistically significant difference between raters, mainly due to multiple short FOG episodes being pooled into longer episodes (Rater A vs Rater C - Wilcoxon signed rank test Z = -2.167, p = 0.030) (Figure 1). Re-evaluation of the annotations resulted in similarly excellent

reliability with smaller limits of agreement and lower measurement error (Supplementary Table 3 and Supplementary Figure 1).

4.3 FOG episode characteristics in all the annotated videos

Three hundred and sixty-four freezing episodes were annotated (ST: 159 episodes from 25 participants, DT: 205 episodes from 27 participants) (for detailed freezing characteristics see Supplementary Table 4). Freezing most often occurred while initiating turns and within the first 120° arc (ST: 58.5%, DT: 72% of the time) and the outer foot was more frequently unable to initiate a step (ST: 68%, DT: 65% of the time). Notably, dual tasking led to a higher number of freezing episodes (Wilcoxon signed rank test $Z = -2.295$, $p = 0.022$), longer episode duration ($Z = -2.639$, $p = 0.008$) and higher %TF ($Z = -2.476$, $p = 0.013$) when pooled across time points within participants showing freezing at any one measurement ($N = 29$).

4.4 Criterion validity of visual-rated FOG

Significant associations were found between the NFOGQ total score and the pre-training DT FOG duration (rho = 0.44, p = 0.007), DT %TF (rho = 0.47, p = 0.004) and a trend for DT number of episodes (rho = 0.30, p = 0.074). NFGOQ sub-scores for turning FOG frequency were associated with ST and DT number of episodes, and sub-scores for turning FOG duration were associated with the ST and DT %TF, and DT number of episodes and duration. Sensor-based AP and ML FOG ratio were only associated with DT number of episodes (Table 3).

4.5 Responsiveness to intervention

Only DT visual-rated FOG outcomes showed any indication of differences within (ps: 0.047 – 0.088) or between intervention groups (ps: 0.026 – 0.067) from pre-training to retention, with two SBT groups showing significant reductions in number and duration of FOG episodes and %TF compared to TBT (effect sizes d from t: SB75 = -0.99 – -1.12; SBCR = -0.79 – -0.88)

(Figure 2). Figure 2 also illustrates differing responsiveness between AP and ML FOG ratio, where the ML ratio showed a more similar response to the %TF.

4.6 Exploratory analysis – visual FOG ratings and sensor-based turning in place metrics

Repeated measures associations showed that turning jerkiness was significantly associated with number of episodes in both single and dual task, but not with duration of episodes. FOG ratio in both AP and ML directions was significantly associated with number and duration of FOG episodes and the %TF during single task, but not dual task (Supplementary Table 5).

Discussion

This study investigated the reliability and validity of visual ratings of making 360° turns in comparison to validated subjective and previously proposed objective automated FOG rating methods. 360° turns were sensitive in provoking FOG when ON-medication, replicating earlier work in the context of walking 180° turns[21,28] in OFF. We found that application of the proposed criteria for visual rating FOG resulted in higher interrater reliability compared to previous studies (ICCs were 0.73[17], 0.86[18] and 0.9[29] respectively), and low measurement error, making the percent time frozen during turning in place a very promising outcome measure. Importantly, the dual task visual-rated metrics showed criterion validity through small but significant associations with the validated NFOGQ and were responsive to the immediate effects of intervention. This was not verified in relation to the FOG ratio particularly in the AP direction, as this metric did not show consistent associations with visual-rated FOG, nor responsiveness to intervention.

Dual task turning in place as a sensitive and valid measure of FOG

Like previous work[19], 65% of self-professed freezers displayed clinically observable FOG while performing alternating turns in place for one minute. In this respect, turning has shown remarkable consistency[29,30] and remains the most reliable trigger of FOG, possibly due to its

demands on coupling postural control and movement[31] in the absence of external visual strategies. Dual tasking not only improved this sensitivity (71%) as previously seen with walking turns (180°)[28], but also revealed patterns of freezing that were more consistent with the self-reported FOG severity, both in the number but particularly in the duration of FOG episodes. Although large angle turns are less commonly encountered in daily life[32], the 360° turning task presents a greater motor challenge, revealing the extent of motor automaticity deficits[21,28]. In addition, the auditory stroop task presents an ongoing attentional demand with response inhibition and set-switching components[33], loading strongly on prefrontal control circuits, thereby limiting the ability to compensate for deficits in motor automaticity with cognitive strategies. Functional near-infrared spectroscopy while turning in place showed increased prefrontal activity without a dual task and decreased prefrontal activity with a dual task in freezers[34], lending support to these observations. Dual task 360° turns therefore likely reveal the "true" degree of FOG severity, making it not only a sensitive outcome measure, but an ecologically valid one.

Sensor-based FOG metrics show mixed relationship to FOG

Consistent with earlier work[20], we found that the FOG ratios are sensitive to the occurrence of freezing episodes, however results were inconsistent across tasks. Contrasting relationships between the FOG ratio and visual ratings in single and dual task turning may be explained by relative contributions of the number of FOG episodes and their duration to the percentage time frozen. Interestingly, out of the two directions, the ML FOG ratio showed greater responsiveness to the split-belt intervention (recall Figure 2), which may reflect a greater FOG specificity of the ML FOG ratio during the 360° turning task, also supported by stronger repeated measures correlations (see Supplementary Table 5). In contrast to previous work[22,23], relationships between both FOG ratios and the NFOGQ were absent, with scatter plots suggesting an inverted – U relationship, apart from a few outlying points (see Supplementary

Figure 2). A similar distribution can be observed in a larger sample of freezers[23] (Peterson et al., 2020 Supplementary Figure 1), corroborating our findings and suggesting that the FOG ratio is only useful in the early disease stages when episodes are more frequent and shorter in duration. However, once episodes become fewer and longer in duration, the FOG ratio inadequately represents FOG severity. Hence, we advocate caution regarding the use of the FOG ratio in its present form as a measure of FOG severity.

Split-belt treadmill as a tool to reduce FOG severity

Although not a self-evident choice for testing the responsiveness of the FOG outcomes, split-belt treadmill training was found to improve turning performance in the same cohort[9] and therefore we expected to see similar effects on the freezing while turning. In line with this hypothesis, we showed that the same SBT arms that showed the largest improvements on turning speed (SB75 and SBCR), also improved FOG severity during the DT 360° turning task. Mechanisms for the reduction of FOG are likely improved amplitude generation in the leg that walked on the fast belt. Previously, circular treadmill walking[35,36] as well as cued treadmill walking to improve amplitude[37] have reported robust effects on freezing, likely via similar mechanisms. Interestingly, improving spatiotemporal control of one limb may be sufficient to reduce FOG[38].

Challenges and future directions

Rating FOG severity is particularly challenging due to the broad definition of "ineffective stepping" that presents heterogeneously across participants[39]. Previously, comparisons to "relatively normal" steps[29,30] were used to define effectiveness. However, during turning in place, the stepping pattern is altered, making differentiation of effective small stepping patterns from FOG-related shuffling particularly challenging. Furthermore, conflicting viewpoints exist as to whether hesitations and shuffling or festination without a complete motor arrest should

be included within the calculation of the percentage time frozen[40]. Treating these as equivalent to complete and prolonged motor blocks, would likely lead to an overestimation of FOG severity and burden. Thus, while we do not directly compare these approaches, here we provide evidence that using a high-specificity approach provides reliable and valid FOG outcomes. Future work may apply a weighting approach to integrate the various FOG-spectrum presentations, similar to those used in clinical scoring tools[22,41,42]. Additionally, testing these methods and criteria in a larger sample of raters from multiple institutions may provide more robust reliability estimates and lead to further refinement and validation of this approach.

Another limitation of this work is that we compared our clinical ratings to a summary measure of the FOG ratio, rather than segment FOG episodes using automated algorithms[43]. An attempt to segment FOG episodes based on thresholding of the FOG ratio (>2.5) resulted in FOG severity values that were unrelated to the visual-rated values (Supplementary Figure 3). Recent work has shown moderate levels of reliability between automated FOG annotation models and clinician ratings based on a similar two or three-sensor setup[44–46]. Critically, however, none of these FOG segmentation algorithms have been validated for the turning in place task, and most studies only included very severe freezers, off medication. We have highlighted some of the challenges in manual annotation of FOG episodes – which are the basis for training the models, as well as the inconsistent relationships between the FOG ratio and FOG severity in the context of both FOG progression over the disease course (NFOGQ representing trait-FOG) and actual task performance (manual annotations representing state-FOG), therefore these issues do not appear to be trivial.

Finally, a third of the self-professed freezers did not freeze during the 360° turning task, highlighting the scope for improving the sensitivity of FOG provoking tasks. Specifically, studies validating these tasks should include milder freezers while ON medication to evaluate the true sensitivity. Addition of anxiety-provoking components to these tasks through virtual

reality environments[47], may further overload the compensatory resources in early freezers. Alternatively, longitudinal studies using non-episodic markers of FOG (such as turning jerkiness[32] or the ML FOG ratio) may serve to establish them as valid surrogates of FOG severity and eliminate the need for eliciting FOG episodes completely.

Manual rating of FOG severity – barriers and possibilities

A recent systematic review of exercise effects on FOG[6] found that only one[18] out of the fifty included studies annotated FOG episodes and used the percentage time frozen as an outcome measure. Besides the challenges of eliciting FOG in the lab, the time and expertise to annotate FOG is likely the biggest barrier to widespread adoption of these methods. Here, we used an open-source software (https://archive.mpi.nl/tla/elan) having published methods[26] to perform the annotations. Further, we developed standardized criteria to rate FOG trials which would reduce the need for prior expertise. We believe that in doing so, the barriers to manually annotating FOG would be lowered, resulting in more valid estimates of therapeutic effects on FOG.

In summary, we showed that FOG severity during turning in place can be reliably rated using our developed criteria. The resulting visual FOG ratings while dual tasking were both valid and responsive to a split-belt treadmill intervention over a short timescale. Sensor-based ratings were less favorable, showing a complex relationship to FOG that is contingent on specific presentations of FOG. This work provides a robust clinical outcome to test potential therapeutic interventions aimed at reducing the burden of FOG in PD.

Acknowledgements

We would like to thank all participants for their motivated and generous engagement in this study. We are grateful to Demi Zoetewei and Prof. Colleen Canning for input on the annotation criteria.

Author's roles

1) Research project: A. Conception, B. Organization, C. Execution; 2) Statistical Analysis: A. Design, B. Execution, C. Review and Critique; 3) Manuscript: A. Writing of the first draft, B. Review and Critique.

Nicholas D'Cruz: 1B, 1C, 2A, 2B, 3A

Jana Seuthe: 1B, 1C, 2C, 3B

Clara de Somer: 1B, 1C, 2C, 3B

Femke Hulzinga: 1B, 1C, 2C, 3B

Pieter Ginis: 1B, 1C, 2C, 3B

Christian Schlenstedt: 1A, 1B, 1C, 2C, 3B

Alice Nieuwboer: 1A, 1B, 2C, 3B

Funding

Support for this study was provided through a grant from the Jacques and Gloria Gossweiler Foundation. The funding agency had no influence on the study design and implementation, or in the interpretation of its results.

Competing Interests

The authors declare no relevant financial disclosures or conflicts of interest in regard to this work.

Financial Disclosures

N.D., C.D.S., F.H., P.G., and A.N. are employed by the KU Leuven. J.S. and C.S. are employed by the CAU Kiel. F.H. receives a doctoral fellowship from the Flanders Research Funds

(FWO). P.G. received research funding from the Internal Funds of the KU Leuven. C.S. received research grants from the European Commission and the Jacques and Gloria Gossweiler Foundation. A.N. received research grants from Flanders Research Funds (FWO); Jacques and Gloria Gossweiler Foundation; King Baudouin Foundation; European Commission; KU Leuven Internal Funds and MJ Fox Foundation.

References

1. Nutt JG, Bloem BR, Giladi N, Hallett M, Horak FB, Nieuwboer A. Freezing of gait: Moving forward on a mysterious clinical phenomenon. *The Lancet Neurology*. 2011;10(8):734-744. doi:10.1016/S1474-4422(11)70143-0

2. Paul SS, Canning CG, Sherrington C, Lord SR, Close JCT, Fung VSC. Three simple clinical tests to accurately predict falls in people with Parkinson's disease. *Movement Disorders*. 2013;28(5):655-662. doi:10.1002/mds.25404

3. Bloem BR, Grimbergen YAM, Cramer M, Willemsen M, Zwinderman AH. Prospective assessment of falls in Parkinson's disease. *Journal of Neurology*. 2001;248(11):950-958. doi:10.1007/s004150170047

4. Pelicioni PHS, Menant JC, Latt MD, Lord SR. Falls in Parkinson's Disease Subtypes: Risk Factors, Locations and Circumstances. *International Journal of Environmental Research and Public Health*. 2019;16(12):2216. doi:10.3390/ijerph16122216

5. Cosentino C, Baccini M, Putzolu M, Ristori D, Avanzino L, Pelosin E. Effectiveness of Physiotherapy on Freezing of Gait in Parkinson's Disease: A Systematic Review and Meta-Analyses. *Movement Disorders*. 2020;35(4):523-536. doi:10.1002/mds.27936

6. Gilat M, Ginis P, Zoetewei D, et al. Effectiveness of exercise and training-based interventions on Freezing of Gait in Parkinson's disease: A systematic review with meta-analysis. *npj Parkinson's Disease*. Published online 2021.

7. Seuthe J, D'Cruz N, Ginis P, et al. The Effect of One Session Split-Belt Treadmill Training on Gait Adaptation in People With Parkinson's Disease and Freezing of Gait. *Neurorehabilitation and Neural Repair*. 2020;34(10):954-963. doi:10.1177/1545968320953144

8. Fasano A, Schlenstedt C, Herzog J, et al. Split-belt locomotion in Parkinson's disease links asymmetry, dyscoordination and sequence effect. 2016;48:6-12. Accessed December 13, 2018. https://www.sciencedirect.com/science/article/pii/S0966636216300376?via%3Dihub

9. D'Cruz N, Seuthe J, Ginis P, Hulzinga F, Schlenstedt C, Nieuwboer A. Short-Term Effects of Single-Session Split-Belt Treadmill Training on Dual-Task Performance in Parkinson's Disease and Healthy Elderly. *Frontiers in Neurology*. 2020;11. doi:10.3389/fneur.2020.560084

10. Giladi N, Tal J, Azulay T, et al. Validation of the Freezing of Gait Questionnaire in Patients with Parkinson's Disease. *Movement Disorders*. 2009;24(5):655-661. doi:10.1002/mds.21745

11. Nieuwboer A, Rochester L, Herman T, et al. Reliability of the new freezing of gait questionnaire: Agreement between patients with Parkinson's disease and their carers. *Gait and Posture*. 2009;30(4):459-463. doi:10.1016/j.gaitpost.2009.07.108

12. Hulzinga F, Nieuwboer A, Dijkstra BW, et al. The New Freezing of Gait Questionnaire: Unsuitable as an Outcome in Clinical Trials? *Movement Disorders Clinical Practice*. 2020;7(2):199-205. doi:10.1002/mdc3.12893

13. Mancini M, Bloem BR, Horak FB, Lewis SJG, Nieuwboer A, Nonnekes J. Clinical and methodological challenges for assessing freezing of gait: Future perspectives. *Movement Disorders*. 2019;34(6):783-790. doi:10.1002/mds.27709

14. Delval A, Tard C, Rambour M, Defebvre L, Moreau C. Characterization and quantification of freezing of gait in Parkinson's disease: Can detection algorithms replace clinical expert opinion? *Neurophysiologie Clinique*. 2015;45(4-5):305-313. doi:10.1016/j.neucli.2015.09.009

15. Ziegler K, Schroeteler F, Ceballos-Baumann AO, Fietzek UM. A New Rating Instrument to Assess Festination and Freezing Gait in Parkinsonian Patients. *Movement Disorders*. 2010;25(8):1012-1018. doi:10.1002/mds.22993

16. Nieuwboer A, Giladi N. Characterizing freezing of gait in Parkinson's disease: Models of an episodic phenomenon. *Movement Disorders*. 2013;28(11):1509-1519. doi:10.1002/mds.25683

17. Morris TR, Cho C, Dilda V, et al. A comparison of clinical and objective measures of freezing of gait in Parkinson's disease. *Parkinsonism & Related Disorders*. 2012;18(5):572-577. doi:10.1016/j.parkreldis.2012.03.001

18. Walton CC, Mowszowski L, Gilat M, et al. Cognitive training for freezing of gait in Parkinson's disease: a randomized controlled trial. *npj Parkinson's Disease*. 2018;4(1):15. doi:10.1038/s41531-018-0052-6

19. Snijders AH, Haaxma CA, Hagen YJ, Munneke M, Bloem BR. Freezer or non-freezer: Clinical assessment of freezing of gait. *Parkinsonism and Related Disorders*. 2012;18(2):149-154. doi:10.1016/j.parkreldis.2011.09.006

20. Zach H, Janssen AM, Snijders AH, et al. Identifying freezing of gait in Parkinson's disease during freezing provoking tasks using waist-mounted accelerometry. Published online 2015. doi:10.1016/j.parkreldis.2015.09.051

21. Bertoli M, Croce U della, Cereatti A, Mancini M. Objective measures to investigate turning impairments and freezing of gait in people with Parkinson's disease. *Gait and Posture*. Published online 2019. doi:10.1016/j.gaitpost.2019.09.001

22. Mancini M, Smulders K, Cohen RG, Horak FB, Giladi N, Nutt JG. The clinical significance of freezing while turning in Parkinson's disease. *Neuroscience*. 2017;343(December):222-228. doi:10.1016/j.neuroscience.2016.11.045

23. Peterson DS, Van Liew C, Stuart S, Carlson-Kuhta P, Horak FB, Mancini M. Relating Parkinson freezing and balance domains: A structural equation modeling approach. *Parkinsonism and Related Disorders*. 2020;79:73-78. doi:10.1016/j.parkreldis.2020.08.027

24. Silva-Batista C, de Lima-Pardini AC, Nucci MP, et al. A Randomized, Controlled Trial of Exercise for Parkinsonian Individuals With Freezing of Gait. *Movement disorders : official journal of the Movement Disorder Society*. 2020;(5):1-12. doi:10.1002/mds.28128

25. Cockx H, Nonnekes J, Bastiaan R, Radboudumc B, van Wezel R, Wang Y. Dealing with the Heterogeneous Presentations of Freezing of Gait: How Reliable are the Freezing Index and Heart Rate for Freezing Detection? Published online July 30, 2021. doi:10.21203/rs.3.rs-735366/v1

26. Gilat M. How to Annotate Freezing of Gait from Video: A Standardized Method Using Open-Source Software. *Journal of Parkinson's Disease*. 2019;9(4):821-824. doi:10.3233/JPD-191700

27. Mohr DL, Marcon RA. Testing for a 'within-subjects' association in repeated measures data. *Journal of Nonparametric Statistics*. 2005;17(3):347-363. doi:10.1080/10485250500038694

28. Spildooren J, Vercruysse S, Desloovere K, Vandenberghe W, Kerckhofs E, Nieuwboer A. Freezing of gait in Parkinson's disease: The impact of dual-tasking and turning. *Movement Disorders*. 2010;25(15):2563-2570. doi:10.1002/mds.23327

29. Shine JM, Moore ST, Bolitho SJ, et al. Assessing the utility of Freezing of Gait Questionnaires in Parkinson's Disease. *Parkinsonism and Related Disorders*. 2012;18(1):25-29. doi:10.1016/j.parkreldis.2011.08.002

30. Schaafsma JD, Balash Y, Gurevich T, Bartels AL, Hausdorff JM, Giladi N. Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease. *European Journal of Neurology*. 2003;10(4):391-398. doi:10.1046/j.1468-1331.2003.00611.x

31. Earhart GM. Dynamic control of posture across locomotor tasks. *Movement Disorders*. Published online 2013. doi:10.1002/mds.25592

32. Mancini M, Weiss A, Herman T, Hausdorff JM. Turn around freezing: Community-living turning behavior in people with Parkinson's disease. *Frontiers in Neurology*. 2018;9(JAN). doi:10.3389/fneur.2018.00018

33. McFadyen BJ, Hegeman J, Duysens J. Dual task effects for asymmetric stepping on a split-belt treadmill. *Gait and Posture*. 2009;30(3):340-344. doi:10.1016/j.gaitpost.2009.06.004

34. Belluscio V, Stuart S, Bergamini E, Vannozzi G, Mancini M. The Association between Prefrontal Cortex Activity and Turning Behavior in People with and without Freezing of Gait. *Neuroscience*. 2019;416:168-176. doi:10.1016/j.neuroscience.2019.07.024

35. Hong M, Earhart GM. Rotating treadmill training reduces freezing in Parkinson disease: Preliminary observations. *Parkinsonism & Related Disorders*. 2008;14(4):359-363. doi:10.1016/j.parkreldis.2007.07.003

36. Cheng FY, Yang YR, Wu YR, Cheng SJ, Wang RY. Effects of curved-walking training on curved-walking performance and freezing of gait in individuals with Parkinson's disease: A randomized controlled trial. *Parkinsonism and Related Disorders*. Published online 2017. doi:10.1016/j.parkreldis.2017.06.021

37. Frazzitta G, Pezzoli G, Bertotti G, Maestri R. Asymmetry and freezing of gait in parkinsonian patients. *Journal of Neurology*. 2013;260(1):71-76. doi:10.1007/s00415-012-6585-4

38. Spildooren J, Vercruysse S, Meyns P, et al. Turning and unilateral cueing in Parkinson's disease patients with and without freezing of gait. *Neuroscience*. 2012;207:298-306. doi:10.1016/j.neuroscience.2012.01.024

39. Giladi N, Nieuwboer A. Understanding and treating freezing of gait in Parkinsonism, proposed working definition, and setting the stage. *Movement Disorders*. 2008;23(SUPPL. 2):423-425. doi:10.1002/mds.21927

40. Fietzek UM, Zwosta J, Schroeteler FE, Ziegler K, Ceballos-Baumann AO. Levodopa changes the severity of freezing in Parkinson's disease. *Parkinsonism and Related Disorders*. 2013;19(10):894-896. doi:10.1016/j.parkreldis.2013.04.004

41. Fling BW, Cohen RG, Mancini M, Nutt JG, Fair DA, Horak FB. Asymmetric pedunculopontine network connectivity in parkinsonian patients with freezing of gait. *Brain*. 2013;136(8):2405-2418. doi:10.1093/brain/awt172

42. Zoetewei D, Herman T, Brozgol M, et al. *Protocol for the DeFOG Trial: A Randomized Controlled Trial on the Effects of Smartphone-Based, on-Demand Cueing for Freezing of Gait in Parkinson's Disease*.; 2021. doi:10.1016/j.conctc.2021.100817

43. Moore ST, Yungher DA, Morris TR, et al. Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. *Journal of NeuroEngineering and Rehabilitation*. 2013;10(1):19. doi:10.1186/1743-0003-10-19

44. Mancini M, Shah V V., Stuart S, et al. Measuring freezing of gait during daily-life: an open-source, wearable sensors approach. *Journal of NeuroEngineering and Rehabilitation*. 2021;18(1):1. doi:10.1186/s12984-020-00774-3

45. Reches T, Dagan M, Herman T, et al. Using Wearable Sensors and Machine Learning to Automatically Detect Freezing of Gait during a FOG-Provoking Test. Published online 2020. doi:10.3390/s20164474

46. O'Day J, Syrkin-Nikolau J, Anidi C, Kidzinski L, Delp S, Bronte-Stewart H. The turning and barrier course reveals gait parameters for detecting freezing of gait and measuring the efficacy of deep brain stimulation. Barbieri FA, ed. *PLOS ONE*. 2020;15(4):e0231984. doi:10.1371/journal.pone.0231984

47. Economou K, Quek D, MacDougall H, Lewis SJG, Ehgoetz Martens KA. Heart rate changes prior to freezing of gait episodes are related to anxiety. *Journal of Parkinson's Disease*. 2021;11(1):271-282. doi:10.3233/JPD-202146

Figure Legends:

Figure 1. Bland Altman plots for the three visual-rated outcomes. Dots represent the difference in scores for each rater pair on the Y axis, plotted against the mean score from the three raters on the X axis. No systematic error across severity was seen, and no significant differences

between raters was seen, apart from for number of FOG episodes between raters A and C (Wilcoxon signed ranks p = 0.03). LOA - limits of agreement

Figure 2. LOESS curves fit to model predicted values for DT %TF and AP and ML FOG ratios in the four intervention groups. Only %TF showed significant differences from pre-training to retention, although ML FOG ratio also partially captured this pattern. Shaded regions depict 95% confidence intervals. LOESS – Locally weighted scatterplot smoothing, DT – dual task, AP – anterior-posterior, ML – mediolateral, %TF – percent time frozen

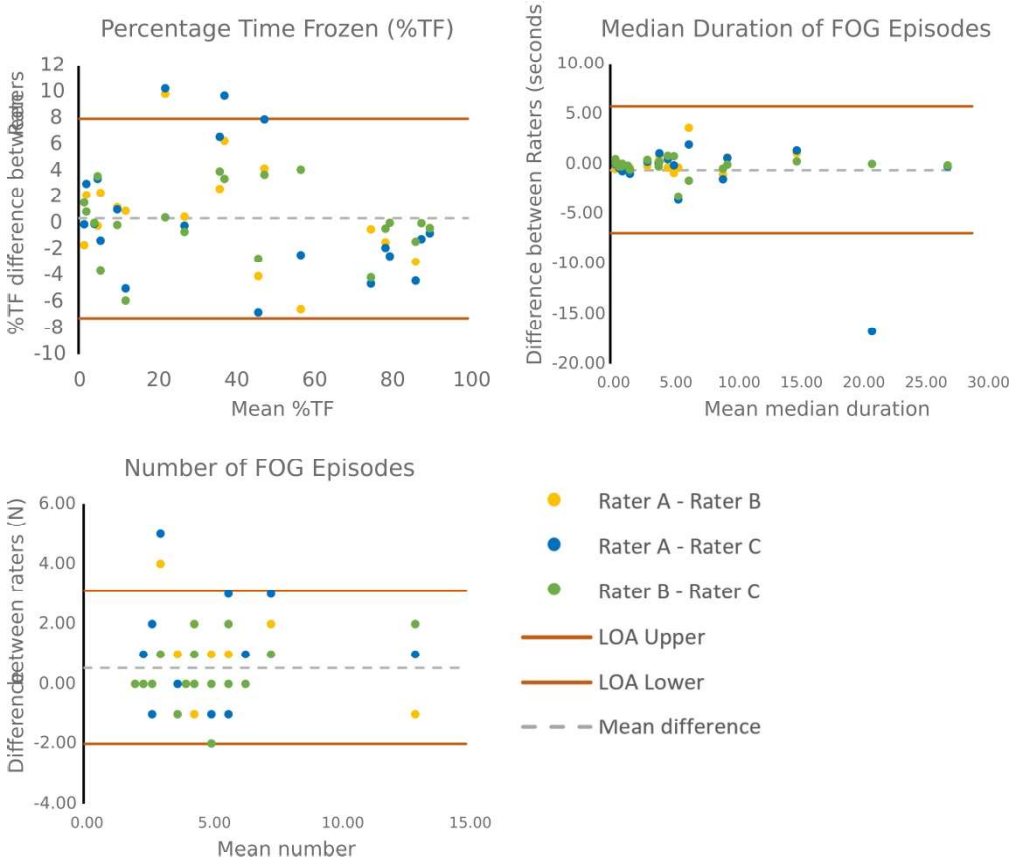


Figure 1. Bland Altman plots for the three visual-rated outcomes. Dots represent the difference in scores for each rater pair on the Y axis, plotted against the mean score from the three raters on the X axis. No systematic error across severity was seen, and no significant differences between raters was seen, apart from for number of FOG episodes between raters A and C (Wilcoxon signed ranks p = 0.03). LOA - limits of agreement

728x623mm (600 x 600 DPI)

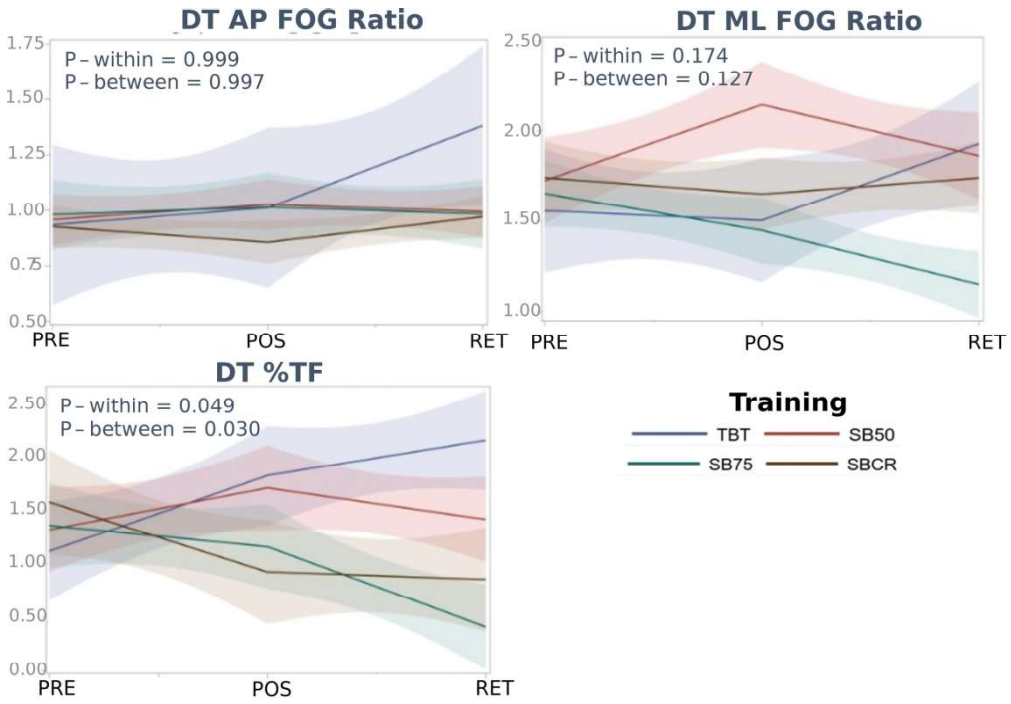


Figure 2. LOESS curves fit to model predicted values for DT %TF and AP and ML FOG ratios in the four intervention groups. Only %TF showed significant differences from pre-training to retention, although ML FOG ratio also partially captured this pattern. Shaded regions depict 95% confidence intervals. LOESS – Locally weighted scatterplot smoothing, DT – dual task, AP – anterior-posterior, ML – mediolateral, %TF – percent time frozen

213x148mm (300 x 300 DPI)

Table 1

Original and Final Criteria for Visual-Rating of FOG

Label	Tiers	Criteria - original	Criteria - final
Start and end of the trial	Trial	From the first intention to move to the last completed full turn	From the first intention to move until 60 seconds later or until the end of the video, whichever is earlier
Start and end of a turn & direction	Turn Right Turn Left	From the first intention to move in one direction to the end of the last step (heel strike) in the same direction; OR if two whole turns are completed in the same direction – the end of the first turn is when the start position is crossed	From the first intention to move in one direction to the end of the last step (heel strike) in the same direction; OR if two whole turns are completed in the same direction – the end of the first turn is when the start position is crossed. For turns that undershoot or overshoot within 90° of the starting position, these are still considered as full turns
Start and end of a FOG episode		From intention to move or the heel/toe-off of the first ineffective step to the toe-off of the first of two effective steps	From the attempted initiation of movement of the unsuccessful step to the initiation of the first of two successful voluntary steps (one on each side)
Type of FOG	Trembling Akinetic	Trembling when accompanied by high frequency movements preceding or during the FOG episode; akinetic when these movements are not visible	Trembling when accompanied by high frequency movements preceding or during the FOG episode; akinetic when these movements are not visible
Position of FOG	Transition First 120 Mid 120 End 120	Transition when FOG occurs at the change of direction (last effective step of previous turn to first effective step of next turn). Use clock position (6 – 10 – 2 – 6 when front facing camera or 12 – 4 – 8 – 12 when rear facing camera) as reference for the 120-degree arcs	Transition when FOG occurs at the change of direction (last effective step of previous turn to first effective step of next turn). Use clock position (6 – 10 – 2 – 6 when front facing camera or 12 – 4 – 8 – 12 when rear facing camera) as reference for the 120-degree arcs
FOG leg	Outer Inner	Leg that first shows sequential reductions in excursion	Leg that fails to initiate movement at the beginning of the FOG episode

Note: Annotated labels included trial duration, direction and duration of each turn, FOG episode duration, type (trembling or akinetic), position (transition or any of three 120° arcs), and leg (inner or outer). Criteria and definitions for all labels were developed in two iterations. In the first iteration, two sets of 20 turning in place trials were rated by two pairs of raters (set 1 – rater A & B, set 2 – rater B & C) using the original criteria. Of these, 20 trials with the largest interrater variability on the percent time frozen were selected (by ND) for the second and final rating. Three of these trials were first discussed between the raters to refine criteria and align operationalization (final criteria). Interrater reliability between the three raters was calculated on these 20 trials. Freezing is defined as lack of voluntary stepping despite the *intention* to move, which is determined by the movements of the upper body, arms, and opposite leg. In case the participant moves the freezing leg due to turning momentum (usually trailing behind the hip), freezing is labelled so long as the freezing leg does not initiate a *voluntary* step. This might be seen as dragging of the leg without foot clearance, or an unexpected lack of progression that is followed by a corrective balance or stepping response. Festination or shuffling *without movement arrests* is not classified as freezing as per this definition and is not included in the rating. Video playback speed was set at 80% to capture shorter episodes and volume was muted to avoid bias for dual tasking. Trunk rotation was used as reference for labelling turn completion, and intention to move was determined from rotation of the shoulders and trunk, from arm movements, or from foot, knee of hip lifting.

Table 2

Interrater Reliability for Visual-Rated FOG Outcomes

Outcome	ICC	LCL	UCL	P	α	Mean Δ	LOA	SEM
% Time Frozen	0.993	0.986	0.997	<0.001	0.998	0.35	±7.59	2.68
Number of Episodes	0.859	0.723	0.937	<0.001	0.955	0.53	±2.55	0.97
Duration of Episodes(s)	0.908	0.820	0.959	<0.001	0.967	-0.64	±6.34	2.27

Note: ICC – Intraclass Correlation Coefficient, LCL – 95% lower confidence limit, UCL – 95% upper confidence limit, α – Cronbach's alpha, Δ – mean difference between the three pairs of raters, LOA – limits of agreement, SEM – standard error of measurement

Table 3

Criterion validity of pre-training visual FOG metrics

Outcome	%TF			FOG number			FOG duration		
	rho	p	CI	rho	p	CI	rho	p	CI
ST									
NFOGQ total	0.22	0.209	(-0.15 - 0.53)	0.23	0.177	(-0.14 - 0.54)	0.16	0.346	(-0.18 - 0.50)
Q3 turn FOG frequency	0.29	0.086	(-0.02 - 0.56)	**0.34**	**0.044**	**(0.02 – 0.58)**	0.24	0.153	(-0.07 - 0.52)
Q4 turn FOG duration	0.14	0.469	(-0.23 - 0.50)	0.12	0.510	(-0.25 - 0.48)	0.10	0.602	(-0.28 - 0.46)
FOG Ratio AP	0.12	0.505	(-0.26 - 0.47)	0.12	0.487	(-0.24 - 0.49)	0.10	0.560	(-0.25 - 0.46)
FOG Ratio ML	0.31	0.074	(-0.05 - 0.62)	0.37	0.029	(-0.01 - 0.62)	0.25	0.141	(-0.15 - 0.58)
DT									
NFOGQ	**0.47**	**0.004**	**(0.14 - 0.73)**	0.30	0.074	(-0.07 - 0.61)	**0.44**	**0.007**	**(0.11 - 0.71)**
Q3 turn FOG frequency	**0.34**	**0.043**	**(0.01 - 0.63)**	0.32	0.059	(-0.04 - 0.6)	0.27	0.107	(-0.08 - 0.57)
Q4 turn FOG duration	**0.52**	**0.002**	**(0.23 - 0.76)**	**0.39**	**0.032**	**(0.03 - 0.68)**	**0.57**	**0.001**	**(0.26 - 0.77)**
FOG Ratio AP	0.25	0.145	(-0.11 - 0.59)	**0.41**	**0.014**	**(0.07 - 0.66)**	0.23	0.179	(-0.16 - 0.58)
FOG Ratio ML	0.34	0.045	(-0.02 - 0.64)	**0.57**	**<0.001**	**(0.23 - 0.77)**	0.25	0.149	(-0.10 - 0.59)

Note: Spearman rank correlation coefficients, p-values and bootstrap 95% percentile confidence intervals (CI) are presented. ST – single task, DT – dual task, AP – anterior-posterior, ML – mediolateral, %TF – percent time frozen

Supplementary Material

Supplementary Table 1

Demographic and clinical profile of the included sample of people with PD

Measure	Mean (Range)
Demographics, cognition & balance	
Age (Years)	68.62 (48 - 86)
Gender (%F)	26.7
MMSE (/30)	28.38 (24 - 30)
MOCA (/30)	24.52 (17 - 30)
FAB (/18)	15.77 (11 - 18)
Mini-BEST (/28)	20.42 (4 - 28)
FES-I (/64)	28.5 (16 - 54)
Fallers (%)	52.3
Disease characteristics	
MDS-UPDRS III (/132)	35.88 (6 - 81)
Disease Duration (Years)	12.89 (1 - 38)
Hoehn & Yahr (% I/II/III/IV)	2.2/40/44.4/13.3
LEDD (mg)	818.6 (175 - 1698.5)
NFOGQ (/30)	16.33 (6 - 29)

Note: Means and ranges or percentages are reported. MMSE – Mini Mental Status Examination, MOCA – Montreal Cognitive Assessment, FAB – Frontal Assessment Battery, FES-I – International version of the Falls Efficacy Scale, MDS-UPDRS III – Motor subscale of the Movement Disorders Society sponsored revision of the Unified Parkinson's Disease Rating Scale, LEDD – Daily Levodopa Equivalent Dose, NFOGQ – New Freezing of Gait Questionnaire

Supplementary Table 2

Demographic, clinical and training information for the four groups included in the study

Demographics, Cognition and Balance						
N	10	12	11	12		
Age	67 (53.5 - 77.5)	66.5 (57.75 - 75)	71 (65 - 78)	71.5 (63.25 - 78.5)	NS	0.680
Sex (% Female)	10	33.3	27.3	33.3	NS	0.523
MMSE	28.5 (26.5 - 30)	28 (28 - 29)	28 (28 - 29)	29 (28 - 30)	NS	0.551
MOCA	26.5 (20.5 - 27)	24 (22 - 29)	24 (20 - 26)	25.5 (23.25 - 28)	NS	0.611
FAB	16 (15 - 17.25)	16 (13 - 18)	16 (15 - 17)	16 (15.25 - 17.75)	NS	0.938
FES-I	22 (20.75 - 42.75)	23 (17.25 - 34.75)	28 (20 - 36)	26.5 (22.25 - 33)	NS	0.764
Retrospective falls (N)	1 (0 - 9.5)	2 (0 - 24)	1 (0 - 5)	0 (0 - 2)	NS	0.413
Mini-BEST	23 (17.75 - 25)	20.5 (17 - 24.75)	18 (17 - 22)	24 (17 - 26)	NS	0.619
Disease-specific Scales						
Disease Duration (years)	9.75 (5 - 17.75)	13.5 (7.75 - 16.75)	14 (8.5 - 15)	11.5 (6 - 15.5)	NS	0.882
H&Y % I/II/III/IV	10/40/40/10	0/41.7/41.7/16.7	0/27.3/54.5/18.2	0/50/41.7/8.3	NS	0.857
LEDD	773 (576.25 - 1189.25)	814 (516.5 - 933.75)	810 (704 - 958.75)	805 (543.75 - 994.37)	NS	0.994
MDS-UPDRS Part III	32.5 (22 - 54.75)	43 (23 - 53)	35 (25 - 41)	33.5 (26 - 43.75)	NS	0.827
NFOGQ	16 (9 - 21.5)	15 (13.5 - 23.5)	15 (10 - 20)	17.5 (13.25 - 18.75)	NS	0.833
Training Engagement and Intensity						
Training Velocity (m/s)	0.98 (0.77 - 1.30)	1.09 (1.04 - 1.23)	0.95 (0.80 - 1.21)	1.12 (0.86 - 1.27)	NS	0.870
Training Duration (min)	30 (24.75 - 30)	30 (29.25 - 30)	30 (25.1 - 30)	30 (27.75 - 30)	NS	0.979
Training Reduced (% Yes)	30	18.2	27.3	27.3	NS	0.888
Handrail use (% Yes)	20	16.7	36.4	16.7	NS	0.655
Borg During	13 (11.75 - 13.5)	12.5 (10.25 - 14.5)	13 (12.75 - 14.25)	13 (11 - 15)	NS	0.702
Borg Post	14 (13 - 15.5)	13 (12.25 - 14.75)	15 (13 - 15.25)	14 (13 - 16.75)	NS	0.643
VAS Mental Pre	2 (1.15 - 3.55)	2.5 (1 - 5.8)	1.55 (0.47 - 3.12)	1.9 (0.425 - 3)	NS	0.566
VAS Mental Post	4.5 (2 - 6.25)	3.85 (1.35 - 7.4)	4.85 (2.87 - 6.3)	3.2 (2 - 5.35)	NS	0.793
VAS Physical Pre	3.1 (0.75 - 5.9)	3.1 (1.5 - 7.5)	4 (1.27 - 5.01)	1.35 (0.8 - 4.75)	NS	0.574
VAS Physical Post	4.6 (1.95 - 7.3)	4 (3.42 - 6.1)	4.9 (4 - 6.12)	3.4 (1.27 - 5.95)	NS	0.620

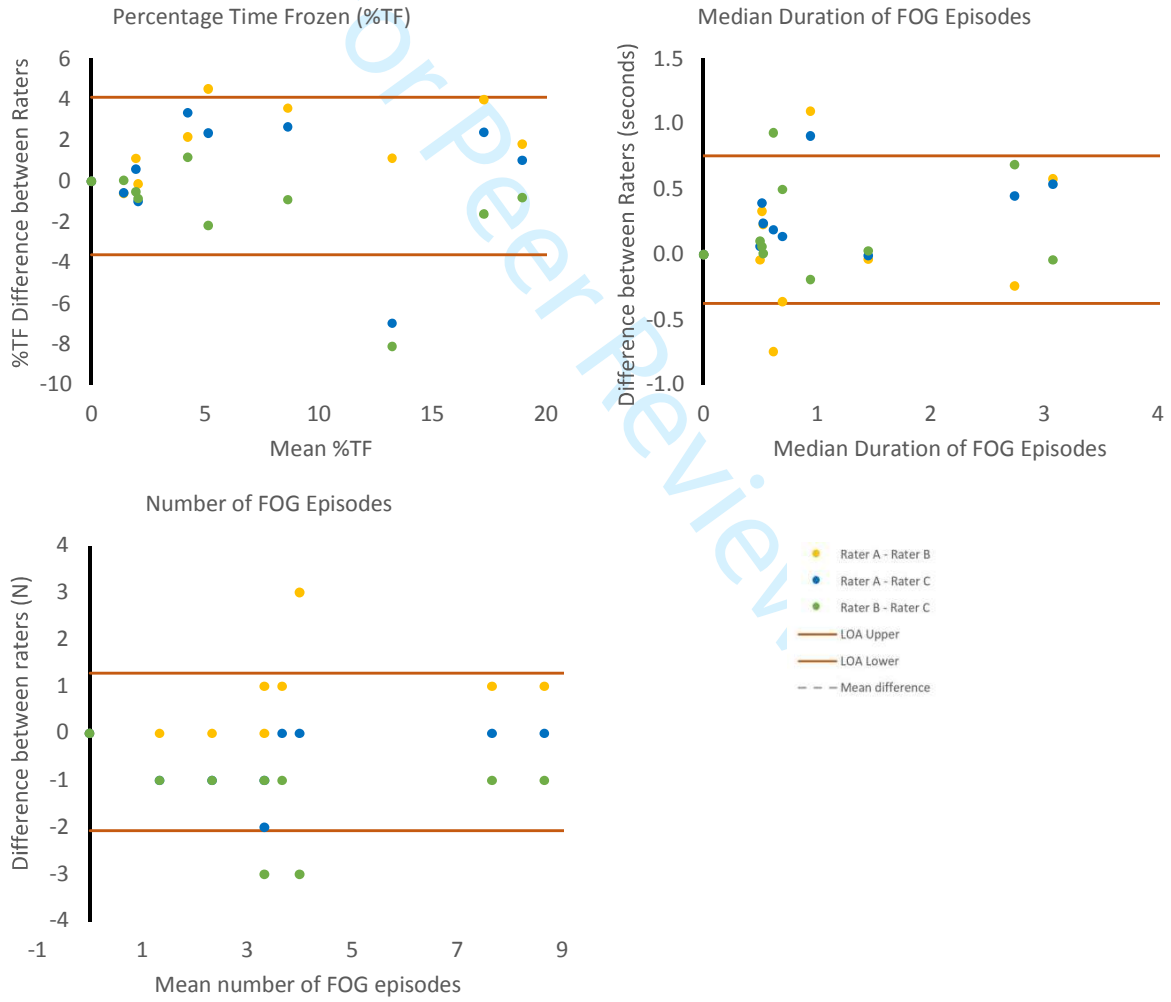
Note: Median values (Q1 – Q3) or percentages are reported. KW – Kruskal-Wallis test, MMSE – Mini Mental Status Examination, MOCA – Montreal Cognitive Assessment, FAB – Frontal Assessment Battery, FES-I – International version of the Falls Efficacy Scale, Mini-BEST – LEDD – Daily Levodopa Equivalent Dose, MDS-UPDRS part III – Motor subscale of the Movement Disorders Society sponsored revision of the Unified Parkinson's Disease Rating Scale, NFOGQ – New Freezing of Gait Questionnaire, Pre – Pre-training, During – Halfway through training (>3 blocks), Post – Post-training, VAS – Visual Acuity Scale

Supplementary Table 3

Re-evaluation of Interrater Reliability for Visual-Rated FOG Outcomes after 15 months

Outcome	ICC	LCL	UCL	P	α	Mean Δ	LOA	SEM
% Time Frozen	0.928	0.805	0.980	<0.001	0.979	0.25	±3.87	1.90
Number of Episodes	0.901	0.613	0.975	<0.001	0.982	0.4	±1.67	0.87
Duration of Episodes(s)	0.915	0.772	0.976	<0.001	0.974	0.19	±0.57	0.30

Note: ICC – Intraclass Correlation Coefficient, LCL – 95% lower confidence limit, UCL – 95% upper confidence limit, α – Cronbach's alpha, Δ – mean difference between the three pairs of raters, LOA – limits of agreement, SEM – standard error of measurement



Supplementary Figure 1 – Re-evaluation of visual annotations. Bland Altman plots for the three visual-rated outcomes for 10 additional trials rated 15 months after criteria development. Dots represent the difference in scores for each rater pair on the Y axis, plotted against the mean score from the three raters on the X axis. Interrater reliability was similar to the second iteration with smaller LOA and measurement error. LOA - limits of agreement, FOG – Freezing of gait

Supplementary Table 4

Distribution of freezing episodes among participants during ST and DT turning in place

ID	FOG in ST	FOG in DT	ST number of episodes	DT number of episodes
001	YES	YES	11	16
002	YES	YES	11	16
003	YES	YES	1	4
004	YES	YES	7	5
005	NO	YES	0	2
006	YES	YES	14	11
007	NO	YES	0	1
008	YES	YES	11	12
009	NO	NO	0	0
010	YES	YES	13	12
011	YES	NO	2	0
012	YES	YES	8	12
013	NO	NO	0	0
014	YES	YES	4	6
015	NO	NO	0	0
016	YES	YES	5	6
017	YES	YES	19	9
018	YES	YES	4	4
019	NO	NO	0	0
020	NO	NO	0	0
021	NO	YES	0	1
022	YES	YES	2	2
023	YES	YES	13	25
024	YES	YES	1	2
025	NO	NO	0	0
026	NO	NO	0	0
027	NO	YES	0	2
028	YES	YES	2	1
029	YES	NO	1	0
030	YES	YES	3	6
031	YES	YES	11	17
032	YES	YES	4	4
033	YES	YES	1	1
034	NO	NO	0	0
035	NO	NO	0	0
036	YES	YES	1	5
037	YES	YES	6	19
038	NO	NO	0	0
039	YES	YES	4	4
Total	**25**	**27**	**159**	**205**

Note: Video data was captured from 39 people with Parkinson's disease and FOG. FOG was provoked in 29 of the participants – 23 in both tasks, 25 in ST and 27 in DT. FOG – Freezing of gait, ST – single task, DT – dual task

Supplementary Table 5

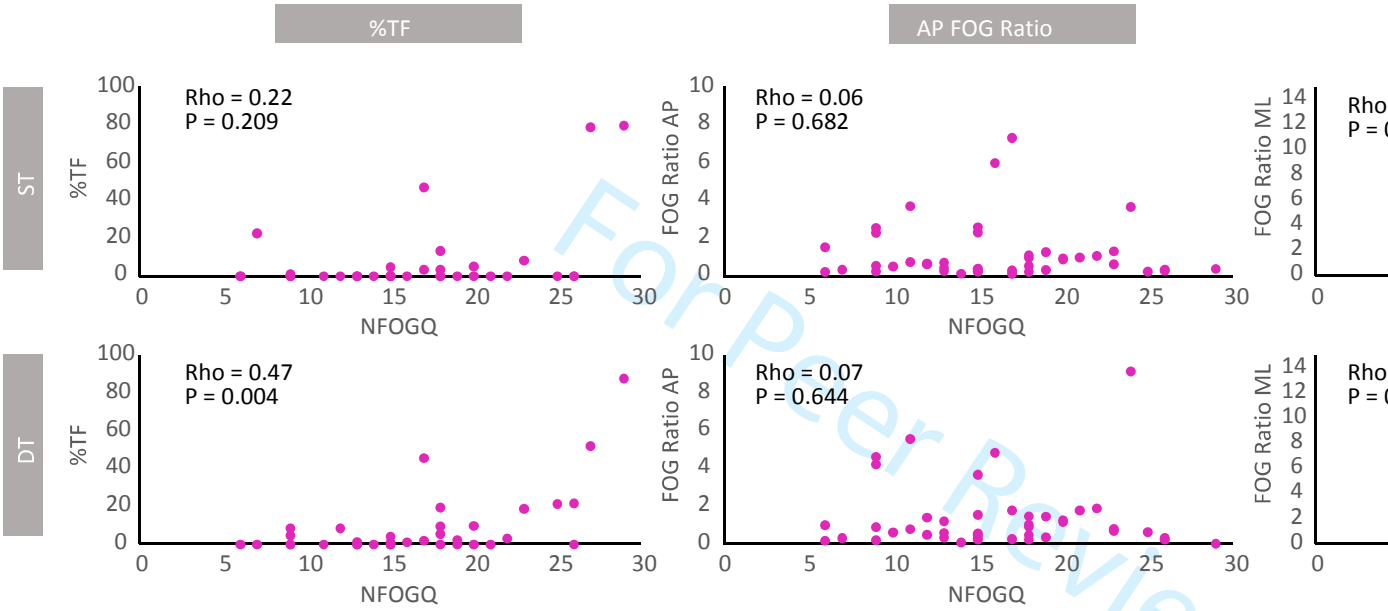
Associations between turning metrics and visual-rated FOG over time

Outcome	Task	Episode duration	Episode number	%TF
Turning				
Mean turning speed	ST	-1.51 (0.120)	-0.35 (0.723)	-1.05 (0.263)
	DT	-0.53 (0.595)	-1.69 (0.088)	-1.73 (0.067)
Peak turning speed	ST	-0.38 (0.702)	0.11 (0.913)	-0.54 (0.611)
	DT	-0.94 (0.331)	**-1.95 (0.045)**	**-2.15 (0.03)**
Mean Jerk	ST	1.75 (0.078)	**2.25 (0.023)**	**2.38 (0.019)**
	DT	-0.76 (0.425)	**2.05 (0.036)**	1.04 (0.315)
FOG Ratio AP	ST	**2.14 (0.029)**	**2.72 (0.004)**	**2.61 (0.009)**
	DT	0.90 (0.371)	1.61 (0.095)	1.35 (0.168)
FOG Ratio ML	ST	**2.14 (0.029)**	**2.88 (0.002)**	**2.92 (0.004)**
	DT	1.31 (0.210)	**1.94 (0.036)**	1.61 (0.099)

Note: Non-parametric within-subject repeated measures correlation Z score and associated p-values (in brackets) are shown for visual FOG rating with objective measures of turning. Measures with a significant p-value are shown in bold. AP - antero-posterior, ML - mediolateral

Supplementary Figure 2

Scatter plots for the pre-training single and dual task objective FOG measures and the subjective NFOGQ



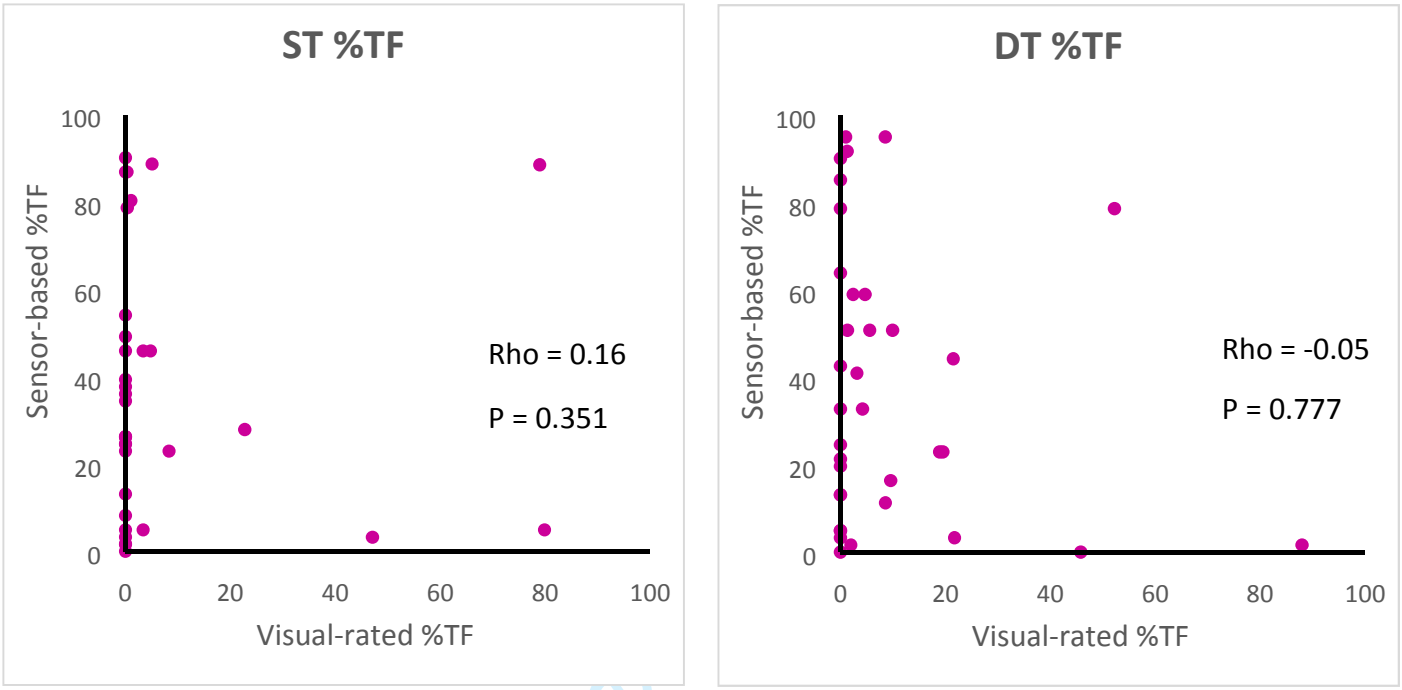
Note: Spearman rank correlations (rho) and associated p-values are presented in each panel. Dual task perc

significantly associated with the subjective rating. One outlying value (all FOG ratios > 28) was omitted fro

the analysis. ST – single task, DT – dual task, AP – anterior-posterior, ML – mediolateral, %TF – percent time

Supplementary Figure 3

Scatter plots for the sensor-based and visual-rated percent time frozen at baseline



Note: Sensor-based percent time frozen was calculated by thresholding the AP FOG ratio > 2.5 to segment periods of possible-freezing from normal turning. Scatter plots and spearman correlations revealed no relationship between these two metrics.