Movement Disorders
CLINICAL PRACTICE

RESEARCH ARTICLE

The New Freezing of Gait Questionnaire: Unsuitable as an Outcome in Clinical Trials?

Femke Hulzinga, MSc,[1] Alice Nieuwboer, PhD,[1] Bauke W. Dijkstra, MSc,[1] Martina Mancini, PhD,[2] Carolien Strouwen, PhD,[3] Bastiaan R. Bloem, MD, PhD,[4] and Pieter Ginis, PhD[1,*]

ABSTRACT: Background: Freezing of gait (FOG) is a common gait deficit in Parkinson's disease. The New Freezing of Gait Questionnaire (NFOG-Q) is a widely used and valid tool to quantify freezing of gait severity. However, its test-retest reliability and minimal detectable change remain unknown.
Objective: To determine the test-retest reliability and responsiveness of the NFOG-Q.
Methods: Two groups of freezers, involved in 2 previous rehabilitation trials, completed the NFOG-Q at 2 time points (T1 and T2), separated by a 6-week control period without active intervention. Sample 1 (N = 57) was measured in ON and sample 2 (N = 14) in OFF. We calculated various reliability statistics for the NFOG-Q scores between T1 and T2 as well as correlation coefficients with clinical descriptors to explain the variability between time points.
Results: In sample 1 the NFOG-Q showed modest reliability (intraclass correlation coefficient = 0.68 [0.52–0.80]) without differences between T1 and T2. However, a minimal detectable change of 9.95 (7.90–12.27) points emerged for the total score (range 28 points, relative minimal detectable change of 35.5%). Sample 2 showed largely similar results. We found no associations between cognitive-related or disease severity–related outcomes and variability in NFOG-Q scores.
Conclusions: We conclude that the NFOG-Q is insufficiently reliable or responsive to detect small effect sizes, as changes need to go beyond 35% to surpass measurement error. Therefore, we warrant caution in using the NFOG-Q as a primary outcome in clinical trials. These results emphasize the need for robust and objective freezing of gait outcome measures.

Freezing of gait (FOG) is a prominent and debilitating symptom of Parkinson's disease (PD). It affects up to 80% of PD patients during the course of the disease.[1–3] FOG is defined as the inability to progress forward stepping despite the intention to walk and reach a destination.[4] Furthermore, FOG is one of the most frequent causes of falls in PD, thus contributing to high fall rates ranging from 35% up to 90%.[1,5,6] FOG seriously impedes daily life functioning and overall quality of life.[7,8] So far, the treatment of FOG, including pharmacological, surgical, and rehabilitation interventions, is only partially effective.[4,9] Therefore, new and more personalized rehabilitation approaches are now being developed. To evaluate their effectiveness, valid and reliable assessment is necessary to document FOG severity and its progression.

A recent review of Mancini and colleagues[10] highlighted that FOG assessment is hampered by several factors in clinical and laboratory settings. First, the episodic and unpredictable nature of FOG increases the likelihood of missing the event during formal or "online" performance tests of gait. Second, various "testing effects" may be at play that enhance or reduce the occurrence of FOG, such as consciously attending to walking, stepping in broad and well-lit corridors, and experiencing medication effects and stress.[4,10,11] To overcome these issues, Giladi and colleagues[11] developed the original Freezing of Gait Questionnaire (FOG-Q), which was later revised

[1]KU Leuven, Department of Rehabilitation Sciences, Neurorehabilitation Research Group, Leuven, Belgium; [2]Department of Neurology, Oregon Health & Science University, Portland, Oregon, USA; [3]UHasselt, Faculty of Rehabilitation Sciences, Rehabilitation Research Center, Hasselt, Belgium; [4]Radboud University Medical Centre, Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Centre of Expertise for Parkinson & Movement Disorders, Nijmegen, the Netherlands

***Correspondence to:** Dr. Pieter Ginis, Department of Rehabilitation Sciences, Tervuursevest 101, 3001 Leuven, Belgium; E-mail: pieter.ginis@kuleuven.be
Keywords: new freezing of gait questionnaire, freezing of gait, minimal detectable change, reliability.
Relevant disclosures and conflicts of interest are listed at the end of this article.
Received 21 November 2019; revised 13 December 2019; accepted 30 December 2019.
Published online 14 January 2020 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mdc3.12893

by Nieuwboer and colleagues[12] into the New FOG-Q (NFOG-Q). Specifically, the NFOG-Q comes with an accompanying video showing the different types of FOG episodes, making it easier to explain FOG to patients and improving its recognition.[13]

The NFOG-Q is in essence a valid, self-reported questionnaire assessing the clinical aspects of freezing (frequency and duration) and its impact on quality of life when looking back during a period of 1 month.[12,14–16] The total score ranges between 0 and 28 points and requires a rating of FOG occurrence in both *on* or *off* medication states. The questionnaire is brief and easy to administer without any additional equipment. The NFOG-Q has showed a high reliability of scores between people with PD and their caregivers (intraclass correlation coefficient [ICC] = 0.78; 95% confidence interval [95% CI], 0.65–0.87) and also high internal consistency (Cronbach's α = 0.84).[12,17] Both the original FOG-Q and the NFOG-Q have been used frequently as assessment tools in clinical trials[18–21] and were endorsed by the "MDS Task Force on posture, gait and balance instruments" as "recommended" and "suggested," respectively.[17]

To interpret effect sizes in clinical studies, the reliability, and even more so the responsiveness of a primary outcome, indicate whether the effects surpass the measurement error. The most important measure of responsiveness is the minimal detectable change (MDC), expressing the smallest change that cannot be attributed to chance or measurement error.[22] Surprisingly (and despite the NFOG-Q's widespread use), its test-retest reliability and MDC have never been determined.[17,23,24] Therefore, the aim of this study was to investigate these metrics for the NFOG-Q, covering a typical time period needed for studying the effects of rehabilitation, that is, 6 weeks. For this analysis, we based ourselves on the existing data of 2 recent studies that used the NFOG-Q as a secondary outcome.[25,26]

Methods

Participants

We analyzed the datasets of 2 recent studies in which participants completed the NFOG-Q at 2 time points (T1 and T2), separated by an interval of 6 weeks without active treatment.

Study 1 included 117 participants with PD, with or without FOG, recruited in 2 centers. They participated in the Duality randomized controlled trial,[25,27] comparing 2 6-week training programs including cognitive and gait training, either combined or trained separately, aiming at improving dual-task gait velocity (clinicaltrials.gov; NTC01375413). After randomization, all of the participants underwent 2 baseline assessments with an interval of 6 weeks, serving as a control period before training started. The 57 patients who had FOG at T1, based on item 1 of the NFOG-Q score ≥ 1, were included in the current analysis. The participants were tested in the *on* state of medication, although the rating of FOG severity as part of the NFOG-Q contains both the *on* and *off* states. The inclusion criteria of study 1 were the following: diagnosis of PD according to the UK Brain Bank criteria,[28] Hoehn and Yahr stages 2 to 3 during the *on* phase of PD medication,[29] able to walk 10 minutes continuously, the presence of dual-task interference,[27] Mini Mental State Examination ≥24/30, stable medication during the past 3 months, no hearing or visual problems interfering with testing or training, and either no deep brain stimulator or stable deep brain stimulator settings during the past year. The participants were excluded if they presented severe balance deficits or medical conditions other than PD affecting gait. This study received ethical approval from the Ethical Committee UZ-KU Leuven (B322201213165) and Central Committee on Research Involving Human Subjects Region Arnhem-Nijmegen (NL39530.091.12).

Data from dataset 2 served as a verification cohort. This dataset contained 46 individuals who participated in 2 mirroring studies with a cross-over design, investigating the effect of a novel exercise intervention (Cognitive Challenging Agility Boot Camp[26]) on balance in PD patients with and without FOG. Both studies have been registered (NCT02231073 and NCT02236286), and the protocol, which was similar in both studies, has been published.[26] The participants were randomly assigned to either an exercise-first or an education-first intervention. For the purpose of this study, we excluded people who started with the exercise as the cross-over design did not include a washout period between the 2 interventions. A total of 14 freezers from the education-first group, as determined by an item 1 NFOG-Q score ≥ 1 at T1, were included. They received a chronic disease education program for 6 weeks, which did not

TABLE 1 Descriptors at first baseline assessment (T1) in both cohorts

Variable	Dataset 1, N = 57 *on*	Dataset 2, N = 14 *off*	*P* Value
Age, yrs	65.96 ± 9.29	70.29 ± 9.95	0.129
Gender, M/F[a]	43/14	11/3	0.806
Disease duration, yrs[b]	12.00 (9.00)	7.67 (6.52)	0.381
≥1 falls 6 months prior to study, yes/no[a]	35/22	9/5	0.842
MoCA, 0–30[b]	26.00 (4.00)	23.00 (15)	***0.039***
FAB, 0–18[b]	16.00 (3.00)	13.00[c]	0.510
MDS-UPDRS 3, 0–132	36.46 ± 12.03	52.86 ± 14.97	***0.001***

[a]Analyzed with chi-squared test.
[b]Analyzed with Mann-Whitney *U* test.
[c]Only available for 2 participants; no range could be calculated.
Descriptors displayed as means ± standard deviations in cases of parametric statistics or median (interquartile range) in cases of nonparametric statistics.
M, male; F, female; MoCA, Montreal Cognitive Assessment; FAB, Frontal Assessment Battery; MDS-UPDRS 3, Movement Disorders Society Unified Parkinson's Disease Rating Scale, motor part 3.
Bold and italics indicate statistically significant *P*-values.

TABLE 2 Mean scores and reliability outcomes

Dataset	T1, Mean (SD)	T2, Mean (SD)	*P* Value	ICC (95% CI)	SEM (95% CI)	MDC	%MDC
Total (0–28)							
Dataset 1, N = 57	12.96 (5.53)	12.00 (7.10)	0.16	0.68 (0.52–0.80)	3.59 (2.85–4.43)	9.95	35.5
Dataset 2, N = 14	14.14 (4.44)	14.86 (4.79)	0.53	0.60 (0.12–0.85)	2.88 (1.76–4.26)	7.99	28.5
Part 2 (0–19) FOG severity							
Dataset 1, N = 57	9.72 (4.25)	9.02 (5.31)	0.21	0.62 (0.43–0.76)	3.60(2.88–4.40)	9.98	52.5
Dataset 2, N = 14	11.14 (3.30)	11.36 (3.23)	0.81	0.50 (−0.02 to 0.81)	2.91 (1.81–4.16)	8.09	42.6
Part 3 (0–9) FOG impact on daily life							
Dataset 1, N = 57	3.25 (2.17)	2.98 (2.30)	0.33	0.58 (0.38–0.73)	1.47 (1.18–1.79)	4.07	45.2
Dataset 2, N = 14	3.00 (1.41)	3.50 (1.91)	0.21	0.65 (0.21–0.87)	1.02 (0.61–1.54)	2.83	31.4

T1, first assessment; T2, second assessment; SD, standard deviation; ICC, intraclass correlation coefficient; 95% CI, 95% confidence interval; SEM, standard error of the mean; MDC, minimal detectable change; %MDC, relative minimal detectable change; FOG, freezing of gait.

TABLE 3 Outcomes of the Mann-Whitney U test between stable and unstable freezers to nonfreezer in dataset 1

Variable	Stable Freezers, Median (IQR)	Unstable Freezers, Median (IQR)	*P* Value
MDS-UPDRS 3, 0–132	35.50 (26.75–47.25)	34.00 (32.00–34.00)	0.896
NFOG-Q total score T1, 0–28	14.00 (10.00–18.00)	5.00 (3.00–9.00)	***<0.001***
NFOG-Q part 2 score T1, 0–19	11.00 (7.75–14.00)	4.00 (2.00–7.00)	***<0.001***
NFOG-Q part 3 score T1, 0–9	3.00 (2.00–5.00)	1.00 (0.00–2.00)	***0.018***
SCOPA-cog, 0–43	26.00 (23.00–30.25)	22.00 (15.00–29.00)	0.207
PDQ-39 cognitive subscore, 0–100	37.50 (21.88–43.75)	31.25 (12.50–56.25)	0.942
FAB, 0–18	16.00 (14.00–17.00)	17.00 (10.00–18.00)	0.625
MoCA, 0–30	26.00 (24.00–28.00)	25.00 (23.00–27.00)	0.543

IQR, interquartile range reported as median (quartile 1 – quartile 3); MDS-UPDRS 3, Movement Disorders Society Unified Parkinson's Disease Rating Scale, motor part 3; NFOG-Q, New Freezing of Gait Questionnaire; SCOPA-cog, Scales for Outcomes in Parkinson's Disease–Cognition; PDQ-39, Parkinson's Disease Questionnaire–39; FAB, Frontal Assessment Battery; MoCA, Montreal Cognitive Assessment.

address FOG. The participants were tested in the *off* state of medication (at least 12 hours withdrawal of dopaminergic medication), but as stated previously referred to both states when actually rating FOG. Inclusion criteria for study 2 were the following: diagnosis of PD according to the UK Brain Bank criteria, aged between 50 to 90, stable dopaminergic medication for at least a month, and able to stand or walk for 2 minutes without an assistive device. Participants were excluded if they had comorbidities that contraindicates exercise participation; significant musculoskeletal, peripheral, or central nervous system disorders or had deep brain stimulation surgery; excessive use of alcohol or recreational drugs; and contraindications to magnetic resonance imaging scans. This study was approved by the joint Oregon Health & Science University and Veterans Affairs Portland Health Care System institutional review board ethics committees (4131 and 8979). Patients for both included studies provided written informed consent prior to study participation.

Outcomes

The measures included in this analysis were the following: (1) the NFOG-Q total and subscores part 2 (FOG severity) and part 3 (FOG impact), (2) the Movement Disorders Society–Unified Parkinson's Disease Rating Scale part 3,[30] and (3) the levodopa equivalent daily dose.[31] Furthermore, we selected the following common cognitive outcomes: (1) the Scales for Outcomes in Parkinson's Disease–Cognition,[32] (2) Montreal Cognitive Assessment,[33] (3) Frontal Assessment Battery,[34] and (4) the Parkinson's Disease Questionnaire–39 cognitive subscore.[35]

Statistics

Shapiro-Wilk tests showed that the total and subscores of the NFOG-Q were not normally distributed in the 2 cohorts. Therefore, nonparametric Wilcoxon signed-rank tests were performed to investigate the possible systematic differences in NFOG-Q scores between T1 and T2. However, because the reliability tests were based on parametric assumptions, we also performed paired *t* tests to verify if the findings were similar. Because the findings of both parametric and nonparametric tests showed similar results, we opted to only report the parametric test outcomes. The reliability of NFOG-Q scores was calculated using ICC 2-way, mixed-effects model with absolute agreement for single measures. MDC was calculated using the following formula: $MDC = SEM \times 1.96 \times \sqrt{2}$, whereby $SEM = SD_{pooled} \times \sqrt{1 - ICC}$. Relative MDC (%MDC) was calculated using the formula $\%MDC = \frac{MDC}{Maximum\ Score} \times 100$. Test-retest reliability and the MDC of the NFOG-Q were determined in datasets 1 and 2 separately, as there were clinical differences between the 2 groups. Bland and Altman plots were used to visualize test-retest disagreement in relation to freezing severity. In study 1, the NFOG-Q was not always administered by the same testers at T1 and T2. Therefore, we also tested whether this had an impact on T1 to T2 differences by comparing groups with and without variable testers with an independent *t* test. There was no significant effect of tester status ($P = 0.52$). Study 2 always had the same testers at T1 and T2.

We also performed a correlation analysis to explore whether the T1 to T2 differences were explained by clinical variables.

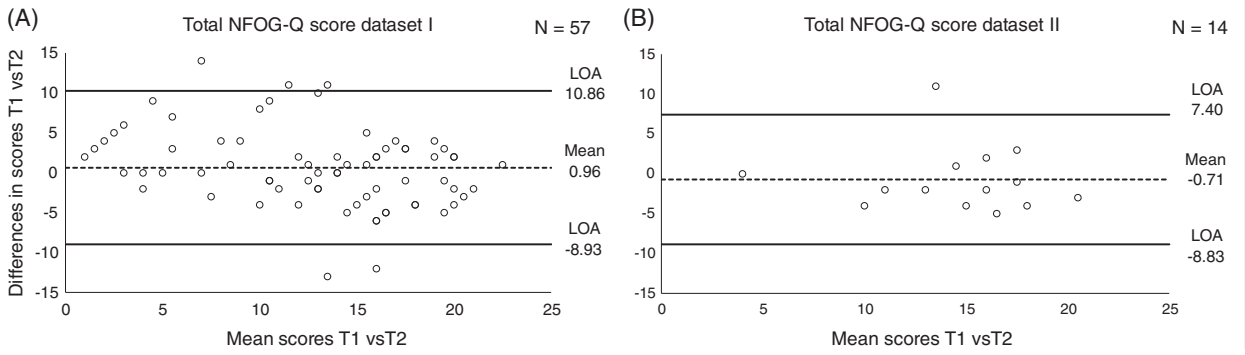

FIG. 1. Bland-Altman plots of the total NFOG-Q score of dataset 1 (**A**) and dataset 2 (**B**). LOA, 95% limits of agreement; Mean, mean T1–T2 NFOG-Q difference; NFOG-Q, New Freezing of Gait Questionnaire; T1, first assessment; T2, second assessment.

Depending on the data distribution, Spearman or Pearson correlations were employed between T1 and T2 differences, and a number of cognitive and disease-related outcome measures were measured at T1 (i.e., NFOG-Q total score at T1, Movement Disorders Society–Unified Parkinson's Disease Rating Scale part 3, Scales for Outcomes in Parkinson's Disease–Cognition, Montreal Cognitive Assessment, Frontal Assessment Battery, and levodopa equivalent daily dose).

As an exploratory analysis, a Mann-Whitney *U* test was performed in dataset 1 for participants who indicated to be freezers at T1 and nonfreezers at T2 (unstable freezers) versus those who indicated to be freezers at both T1 and T2 (stable freezers). We also analyzed whether it was possible to improve the reliability of the NFOG-Q by including subgroups only. All statistical analyses were performed in SPSS (IBM SPSS Statistics 25, Chicago, IL). *P* values <0.05 were considered statistically significant.

Results

Demographic data showed that both datasets were comparable in age, disease duration, fall history, and gender distribution (Table 1). However, the participants in dataset 2 scored significantly lower on the Montreal Cognitive Assessment and higher on Movement Disorders Society–Unified Parkinson's Disease Rating Scale part 3. In dataset 1, 6 participants changed their levodopa equivalent dose during the 6-week period (dose increase in 5 participants). After changing the medication dose of the 6 participants, 2 of them reported no FOG experience in the past 4-week period at T2. In dataset 2, none of the participants changed their levodopa intake.

Table 2 shows the test-retest reliability outcomes and MDC of both datasets. The ICC of the NFOG-Q total score was 0.68 and 0.60 for datasets 1 and 2, respectively, which corresponds with moderate reliability. Consequently, the MDC and relative MDC were 9.95 points (35.5% of the total score of 28) and 7.99 points (28.5%) for datasets 1 and 2. With respect to FOG severity (NFOG-Q part 2), the MDCs of both datasets were 9.98 points in dataset 1 and 7.86 points in dataset 2 of the total score of 19. However, the relative MDC was much higher in study 1, being around 50%. As for the impact of FOG on daily life (NFOG-Q part 3), the MDCs and the relative MDCs (>30%) from both datasets were again high, and MDC was larger in dataset 1 than in dataset 2.

Interestingly, in dataset 1 there were 7 participants who indicated to have experienced FOG at T1 in the past month, but did not report this any longer to be the case at T2 (unstable freezers). Comparing them with those who did not change their freezer status (stable freezers) revealed that they had a significantly lower total NFOG-Q scores as well as FOG severity and FOG impact subscores at T1. Yet, as shown in Table 3, they did not differ on other disease determinants. Of these 7 unstable freezers, 2 increased their medication between T1 and T2. When excluding these 7 participants from the reliability analysis, the ICCs and MDCs showed similar or even worse results.

To explore which patients were more likely to have large T1 to T2 differences, we visually analyzed the difference scores and plotted them within the limits of agreements and across the mean NFOG-Q scores using the Bland and Altman methodology (Fig. 1). When looking at these plots, it appears that the T1 to T2 differences were larger in participants with low to moderate freezing severity (scores from 1 to 17), whereas severe freezers had more stable outcomes (scores ≥18).

To explore whether the reliability of the NFOG-Q would improve by only including the more severely affected participants, new reliability analyses were performed. We included only participants with scores ≥18 points. However, the MDCs only started to decrease substantially after including scores of 20 and above. Only 6 participants in the 2 datasets had such high scores. As an extra sensitivity analysis, we tested whether excluding the participants in dataset 1 who changed their medication dosage between T1 and T2 would improve the reliability of the NFOG-Q. This did not have a substantial effect (MDC 9.29 instead of the original 9.95).

Finally, correlation analyses revealed no associations between any of the cognitive or disease-related descriptors and the T1 to T2 NFOG-Q differences in dataset 1 (rs − 0.227 to 0.120). Similar results were apparent in dataset 2 (rs − 0.461 to 0.433), all nonsignificant.

Discussion

Our findings showed that the test-retest stability of the NFOG-Q, when freezers were retested after a 6-week interval, was modest but showed no significant time differences. Accordingly, with a modest reliability, the responsiveness of the scale was poor, as expressed by very high MDC values in relation to the total scoring range of 0 to 28. These results suggest that the NFOG-Q may not be a sufficiently reliable or responsive outcome to detect small effect sizes. Furthermore, 7 of the 57 participants of study 1 were unsure or reported a change of their freezer status based on item 1 of the NFOG-Q, as they changed from being freezers to nonfreezers during the 6-week period.

The moderate reliability found between T1 and T2 may be explained by the fact that the NFOG-Q is a self-rated questionnaire, asking about the global occurrence of freezing in the past month irrespective of medication status. At T1, the patients may not have been completely aware of their freezing severity, frequency, or impact on daily life because of recall bias. In the period between T1 and T2, the patients may have become more alert to their freezing behavior, and therefore T2 scores altered. The fact that NFOG-Q scores did not differ significantly between time points as such and that no systematic decrease or increase of scores was found across the cohorts suggests that becoming more aware of freezing behavior could go in both directions, that is, more or less positive assessments of FOG. Several studies reported recall bias as a limitation in the previous freezing of gait questionnaire (FOG-Q) or other clinical scales.[36–38] As a result, one study even used the second assessment of the FOG-Q instead of the first to compensate for this issue.[39] To help the accuracy of retrospective assessment of FOG, the NFOG-Q is accompanied by exemplary videos of FOG to serve as a reference before it is completed. However, our results imply that this method may be insufficient to overcome recall bias. Interestingly, patients with more severe FOG (dataset 2, rated in *off*) rated the severity and impact of FOG with more stability than those with moderate FOG (dataset 1 rated in *on*). Overall, our findings illustrate the patients' difficulties with evaluating FOG based on self-perception.

The MDCs of both datasets were high when considering the maximal score (28 points) and subscores (FOG severity 19 points, FOG impact on daily life 9 points). The subscore regarding FOG severity showed an MDC of 9.98, which corresponds with 52.5% of the scoring range of this part of the scale. The sensitivity of the NFOG-Q, therefore, appears to be too low to detect small but potential clinically important changes in FOG. This is evident from previously reported improvements on the NFOG-Q,[19,40,41] which fall below the MDCs as derived from the current study. Hence, the present results are useful to aid in the accurate interpretation of effects sizes. Of note, the minimal clinically importance difference of the NFOG-Q has not been determined yet,[17] and therefore it is unknown what patients themselves regard as a meaningful clinical change of their freezing. Given these results, we do not recommend to use the NFOG-Q as an outcome measure for FOG in future intervention studies.

Previously, the NFOG-Q showed good reliability when comparing the ratings of PD patients and their caregivers.[12] This suggests that the NFOG-Q remains a useful screening tool to identify freezers and rate FOG severity. We did find 7 misclassifications (unstable freezers). In 2 of these cases, this may be ascribed to receiving a higher dose of levodopa. However, medication was generally kept constant during the 6-week study periods. Moreover, excluding these cases did not yield better results. Classification is based on item 1 of the NFOG-Q: "Did you experience freezing episodes over the past month?"[12,14,24,42] Interestingly, the analysis showed that significantly lower NFOG-Q scores were apparent in the unstable freezers. This implies that mild freezers are less aware of their freezing status and may therefore have greater difficulty with acknowledging that they experience FOG. Two studies used the more stringent criterion of the NFOG-Q > 3 to define a participant as a freezer.[43,44] The results of one of these studies showed that none of the participants with a NFOG-Q score < 3 experienced freezing during their FOG-provoking protocol in the *off* medication state.[43] The NFOG-Q was not always administered by the same testers on both occasions, but protocols in both studies were rigorously standardized between testers and across centers in both studies. We also statistically checked whether the same tester scores and varying tester scorers showed more variability, but this was not the case. As the NFOG-Q is a self-reported questionnaire and not a performance test, we think it is unlikely that tester effects can explain these misclassifications.

The present findings underscore the urgency to find an objective way of measuring FOG. Shine and colleagues[45] compared the NFOG-Q outcomes with actual freezing episodes elicited during a FOG-provoking protocol in the laboratory. Two independent clinicians rated the duration and frequency of the episodes captured on video. The NFOG-Q scores did not correlate with either the duration (r = 0.35, P = 0.095) or frequency (r = 0.30, P = 0.15) of the video-rated procedure. Several studies suggested that video annotations of FOG-provoking tasks could be used as the gold standard for rating freezing severity.[45–47] However, this method is very labor intensive as it should ideally be done by at least 2 independent raters in a standardized manner to allow for comparison between studies.[46,47] Automated video algorithms may enable faster and standardized FOG video annotations.[48] However, the testing effect elicited by standardized assessments in a laboratory or a home setting may preclude many FOG episodes from occurring, even when *off* medication.[49] Therefore, in the future, the field needs to move toward unobtrusive home-based assessment of FOG using digitized mobility outcomes, as was recently suggested by Mancini and colleagues.[10] Combining video annotations with wearable sensor-based data collected during standardized tests will permit the necessary validation of FOG algorithms for domestic use,

allowing data collection spanning for much longer time periods (days or potentially even weeks).[47] Whether objectively and remotely measured FOG detection during spontaneous daily mobility is suitable as a primary outcome for clinical trials (with good test-retest and MDC values) remains to be determined.

Several limitations of this study have to be considered when interpreting our findings. First, we conducted a retrospective analysis of 2 existing datasets collected during the *on* (study 1) and *off* phases (study 2). The testing in the *off* phase may explain why participants in study 2 had more severe FOG. However, the results were very similar across both studies, presumably because both cohorts were required to rate their FOG in both the *on* and *off* conditions, as is stipulated in the scoring instructions of the NFOG-Q. The participants of study 2 did not receive an active intervention between T1 and T2 but were exposed to a chronic disease education program led by a therapist. As such, the participants may have become more aware of their disease and thus of their freezing behavior, although FOG was not explicitly addressed in the education program. Because of the clinical differences between both cohorts, we did not pool the data, which affected the statistical power of the present study. Nevertheless, the fact that we reproduced our study 1 findings with those of study 2 strengthens our conclusions.

In conclusion, based on the modest test-retest reliability and high MDC values, we argue against using the NFOG-Q as primary outcome for intervention studies where small effects sizes are expected. Furthermore, our findings provide a basis for accurate interpretation of effects sizes, namely, that a change in NFOG-Q scores should surpass 35% of the total scoring range to be beyond measurement error. The NFOG-Q remains a useful screening tool for FOG because overall no significant differences between both consecutive test moments were found and it is an easy and quick way to differentiate between freezers and non-freezers. Still, assessors need to be aware of the patients' difficulties with self-perceived ratings of FOG, especially when FOG is still mild. Developing reliable measures for FOG is not trivial as clinical trials are dependent on accurate measures. Hence, we recommend robust testing of the psychometric properties of both automated FOG video annotations and wearable sensor techniques based on large datasets.

Acknowledgments

The authors would like to thank all participants who took part in the studies included in this article. In addition, we would like to thank Dr. Moran Gilat for his input on the article.

Author Roles

(1) Research Project: A. Conception, B. Organization, C. Execution; (2) Statistical Analysis: A. Design, B. Execution, C. Review and Critique; (3) Manuscript Preparation: A. Writing of the First Draft, B. Review and Critique.

F.H.: 1A, 1B, 1C, 2A, 2B, 3A
A.N.: 1A, 1C, 2A, 2C, 3B
B.W.D.: 1C, 3B
M.M.: 1C, 3B
C.S.: 1C, 3B
B.R.B.: 1C, 3B
P.G.: 1A, 1C, 2A, 2B, 2C, 3B

Disclosures

Ethical Compliance Statement: We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this work is consistent with those guidelines. The studies involved in this work were approved by the respective ethical committees (dataset 1: Ethische Commissie UZ-KU Leuven [B322201213165] and Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen [NL39530.091.12]; dataset 2: the joint Oregon Health & Science University and Veterans Affairs Portland Health Care System institutional review board ethics committees [4131 and 8979]). Patients for both included studies provided written informed consent prior to study participation.

Funding Sources and Conflicts of Interest: The original funding sources of both studies were the Jacques and Gloria Gossweiler Foundation and the Malou Malou funds of the King Baudouin Foundation for study 1, and National Institutes of Health R01AG006457 and the Department of Veterans Affairs 5I01RX001075 for study 2. The authors report no conflicts of interest.

Financial Disclosures for the Previous 12 Months: Femke Hulzinga, Bauke W. Dijkstra, Carolien Strouwen and Pieter Ginis report no financial disclosures. Alice Nieuwboer reports no conflict of interest pertaining to the article. She has received grant funding from the European Commission, the Jacques & Gloria Gossweiler Foundation, Flanders Scientific Funds, The King Baudouin Foundation, the Michael J Fox Foundation, and KU Leuven Internal funds. Martina Mancini has received grant support from the National Institutes of Health, the Medical Research Foundation of Oregon, and the Kinnie Family Foundation. Bastiaan R. Bloem currently serves on the editorial board of *Practical Neurology and Digital Biomarkers*; has received honoraria from serving on the scientific advisory boards for Abbvie, Biogen, Union Chimique Belge (UCB), and Walk with Path; has received fees for speaking at conferences from AbbVie, Zambon, Roche, GE Healthcare, and Bial; and has received research support from the Netherlands Organization for Scientific Research, the Michael J Fox Foundation, UCB, Abbvie, the Stichting Parkinson Fonds, the Hersenstichting Nederland, the Parkinson's Foundation, Verily Life Sciences, Horizon 2020, the Topsector Life Sciences and Health, and the Parkinson Vereniging.

References

1. Perez-Lloret S, Negre-Pages L, Damier P, et al. Prevalence, determinants, and effect on quality of life of freezing of gait in Parkinson disease. *JAMA Neurol* 2014;71(7):884–890.
2. Tan D, Danoudis M, McGinley J, Morris ME. Relationships between motor aspects of gait impairments and activity limitations in people with Parkinson's disease: a systematic review. *Park Relat Disord* 2012;18(2):117–124.

3. Hely MA, Reid WGJ, Adena MA, Halliday GM, Morris JGL. The Sydney multicenter study of Parkinson's disease: The inevitability of dementia at 20 years. *Mov Disord* 2008;23(6):837–844.
4. Nutt JG, Bloem BR, Giladi N, Hallett M, Horak FB, Nieuwboer A. Freezing of gait: moving forward on a mysterious clinical phenomenon. *Lancet Neurol* 2011;10(8):734–744.
5. Latt MD, Lord SR, Morris JGL, Fung VSC. Clinical and physiological assessments for elucidating falls risk in Parkinson's disease. *Mov Disord* 2009;24(9):1280–1289.
6. Allen NE, Schwarzel AK, Canning CG. Recurrent falls in Parkinson's disease: a systematic review. *Parkinsons Dis* 2013;2013:906274.
7. Walton CC, Shine JM, Hall JM, et al. The major impact of freezing of gait on quality of life in Parkinson's disease. *J Neurol* 2015;262(1):108–115.
8. Ellis T, Cavanaugh JT, Earhart GM, Ford MP, Foreman KB, Dibble LE. Which measures of physical function and motor impairment best predict quality of life in Parkinson's disease? *Park Relat Disord* 2011;17(9):693–697.
9. Schlenstedt C, Shalash A, Muthuraman M, Falk D, Witt K, Deuschl G. Effect of high-frequency subthalamic neurostimulation on gait and freezing of gait in Parkinson's disease: a systematic review and meta-analysis. *Eur J Neurol* 2017;24(1):18–26.
10. Mancini M, Bloem BR, Horak FB, Lewis SJG, Nieuwboer A, Nonnekes J. Clinical and methodological challenges for assessing freezing of gait: future perspectives. *Mov Disord* 2019;34(6):783–790.
11. Giladi N, Shabtai H, Simon ES, Biran S, Tal J, Korczyn AD. Construction of freezing of gait questionnaire for patients with Parkinsonism. *Park Relat Disord* 2000;6(3):165–170.
12. Nieuwboer A, Rochester L, Herman T, Vandenberghe W, Emil G, Thomaes T. Reliability of the new freezing of gait questionnaire: agreement between patients with Parkinson's disease and their carers. *Gait Posture* 2009;30(4):459–463.
13. Nieuwboer A, Rochester L, Herman T, Vandenberghe W, Emil GE, Thomaes T. Reliability of the new freezing of gait questionnaire: agreement between patients with Parkinson's disease and their carers. *Gait and Posture* 2009;30:459–463.
14. Snijders AH, Haaxma CA, Hagen YJ, Munneke M, Bloem BR. Freezer or non-freezer: Clinical assessment of freezing of gait. *Park Relat Disord* 2012;18(2):149–154.
15. de Souza NS, Leite MAA, Martins ACG, et al. Motor imagery effect on gait in Parkinson's disease: a systematic review. *EC Neurol* 2018;5(1):346–353.
16. Barthel C, Mallia E, Debû B, Bloem BR, Ferraye MU. The practicalities of assessing freezing of gait. *J Parkinsons Dis* 2016;6(4):667–674.
17. Bloem BR, Marinus J, Almeida Q, et al. Measurement instruments to assess posture, gait, and balance in Parkinson's disease: critique and recommendations. *Mov Disord* 2016;31(9):1342–1355.
18. Rawson KS, Creel P, Templin L, Horin AP, Duncan RP, Earhart GM. Freezing of Gait Boot Camp: feasibility, safety and preliminary efficacy of a community-based group intervention. *Neurodegener Dis Manag* 2018;8(5):307–314.
19. Volpe D, Signorini M, Marchetto A, Lynch T, Morris ME. A comparison of Irish set dancing and exercises for people with Parkinson's disease: a phase II feasibility study. *BMC Geriatr* 2013;13(1):54.
20. Myers PS, McNeely ME, Pickett KA, Duncan RP, Earhart GM. Effects of exercise on gait and motor imagery in people with Parkinson disease and freezing of gait. *Park Relat Disord* 2018;53:89–95.
21. Matsuura K, Kajikawa H, Tabei K, et al. The effectiveness of istradefylline for the treatment of gait deficits and sleepiness in patients with Parkinson's disease. *Neurosci Lett* 2018;662:158–161.
22. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther* 2006;86(5):735–743.
23. Martin T, Weatherall M, Anderson TJ, Macaskill MR. A randomized controlled feasibility trial of a specific cueing program for falls management in persons with Parkinson disease and freezing of gait. *J Neurol Phys Ther* 2015;39:179–184.
24. Nieuwboer A, Giladi N. The challenge of evaluating freezing of gait in patients with Parkinson's disease. *Br J Neurosurg* 2008;22(suppl 1), S16–S18.
25. Strouwen C, Molenaar EALM, Münks L, et al. Training dual tasks together or apart in Parkinson's disease: results from the DUALITY trial. *Mov Disord* 2017;32(8):1201–1210.
26. King LA, Peterson DS, Mancini M, et al. Do cognitive measures and brain circuitry predict outcomes of exercise in Parkinson Disease: a randomized clinical trial. *BMC Neurol* 2015;15(1):4–11.
27. Strouwen C, ALM Molenaar E, Keus SH, et al. Protocol for a in Parkinson's randomized comparison of integrated versus consecutive dual task practice disease: the DUALITY trial. *BMC Neurol* 2014;14(61):1–12.
28. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry* 1992;55(3):181–184.
29. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. *Neurology* 17(5):427–442.
30. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord* 2008;23(15):2129–2170.
31. Tomlinson CL, Stowe R, Patel S, Rick C, Gray R, Clarke CE. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov Disord* 2010;25(15):2649–2653.
32. Marinus J, Visser M, Verwey N, et al. Assessment of cognition in Parkinson's disease. *Neurology* 2003;61:1222–1228.
33. Dalrymple-Alford JC, MacAskill MR, Nakas CT, et al. The MoCA: well-suited screen for cognitive impairment in Parkinson disease. *Neurology* 2010;75(19):1717–1725.
34. Lima CF, Meireles LP, Fonseca R, Castro SL, Garrett C. The Frontal Assessment Battery (FAB) in Parkinson's disease and correlations with formal measures of executive functioning. *J Neurol* 2008;255(11):1756–1761.
35. Jenkinson C, Fitzpatrick R, Peto V, Greenhall R, Hyman N. The Parkinson's Disease Questionnaire (PDQ-39): development and validation of a Parkinson's disease summary index score. *Age Ageing* 1997;26 (5):353–357.
36. Mirelman A, Giladi N, Hausdorff JM. Body-fixed sensors for Parkinson disease. *J Am Med Assoc* 2015;314(9):873–874.
37. Tessitore A, Amboni M, Esposito F, et al. Resting-state brain connectivity in patients with Parkinson's disease and freezing of gait. *Park Relat Disord* 2012;18(6):781–787.
38. Wang M, Jiang S, Yuan Y, et al. Alterations of functional and structural connectivity of freezing of gait in Parkinson's disease. *J Neurol* 2016;263 (8):1583–1592.
39. Lo AC, Chang VC, Gianfrancesco MA, Friedman JH, Patterson TS, Benedicto DF. Reduction of freezing of gait in Parkinson's disease by repetitive robot-assisted treadmill training: a pilot study. *J Neuroeng Rehabil* 2010;7(1):51.
40. Iijima M, Orimo S, Terashi H, et al. Efficacy of istradefylline for gait disorders with freezing of gait in Parkinson's disease: a single-arm, open-label, prospective, multicenter study. *Expert Opin Pharmacother* 2019;20 (11):1405–1411.
41. Vercruysse S, Vandenberghe W, Muïks L, Nuttin B, Devos H, Nieuwboer A. Effects of deep brain stimulation of the subthalamic nucleus on freezing of gait in Parkinson's disease: a prospective controlled study. *J Neurol Neurosurg Psychiatry* 2014;85(8):872–878.
42. Mancini M, Weiss A, Herman T, Hausdorff JM. Turn around freezing: community-living turning behavior in people with Parkinson's disease. *Front Neurol* 2018;9:1–9.
43. 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;343:222–228.
44. 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.
45. Shine JM, Moore ST, Bolitho SJ, et al. Assessing the utility of Freezing of Gait Questionnaires in Parkinson's disease. *Park Relat Disord* 2012;18(1): 25–29.
46. Moore ST, Yungher DA, Morris TR, et al. Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. *J Neuroeng Rehabil* 2013;10(1):1–10.
47. Gilat M. How to annotate freezing of gait from video: a standardized method using open-source software. *J Parkinsons Dis* 2019;9:821–824.
48. Hu K, Wang Z, Wang W, et al. Graph sequence recurrent neural network for vision-based freezing of gait detection. *IEEE Trans Image Process* 2019;29:1890–1901.
49. Nieuwboer A, Dom R, De Weerdt W, Desloovere K, Fieuws S, Broens-Kaucsik E. Abnormalities of the spatiotemporal characteristics of Gait at the onset of freezing in Parkinson's disease. *Mov Disord* 2001;16 (6):1066–1075.