

Supplemental material

Appendix 1: Search strategy for the database EMBASE

1. #1 'Parkinsons disease'
2. #2 'Parkinsonism'
3. #3 'PD'
4. #4 'Shaking syndrome'
5. #5 'Movement disorder'
6. #6 'Tremor'
7. #7 'Psychometric properties'
8. #8 'Psychometric testing'
9. #9 'Clinometric properties'
10. #10 'Clinometric testing'
11. #11 'Validity'
12. #12 'Reliability'
13. #13 'Responsiveness'
14. #14 'Outcome measures'

15. #15 'Assessment tools'
16. #16 'Scale'
17. #17 'Measure'
18. #18 'Unified Parkinsons Disease Rating Scale (UPDRS)'
19. #19 'Activity-specific Balance Confidence Scale (ABC)'
20. #20 'Functional Balance Standing Scale'
21. #21 'Time Up and Go Test (TUG)'
22. #22 'Berg Balance Scale (BBS)'
23. #23 'Functional reach'
24. #24 'Functional Gait Assessment (FGA)'
25. #25 'Tinetti Balance Test'
26. #26 'Gait Difficulty Score'
27. #27 'Rating Scale for Gait Evaluation'
28. #28 'Best Evaluation Systems Test (BESTest)'
29. #29 'Gait and Balance Scale'
30. #30 'Falls Efficacy Scale'

31. #31 'Survey of Activities and Fear and Falling'
32. #32 '6-minute walk test'
33. #33 '10-minute walk test'
34. #34 'Rapid Assessment of Postural Instability'
35. #35 'Survey of Activities and Fear of Falling in the Elderly-Modified'
36. #36 'Balance'
37. #37 'Stability'
38. #38 'Equilibrium'
39. #39 'Postural control'
40. #40 'Postural balance'
41. #41 'Falls'
42. #42 'Falls incidence'
43. #43 'Falls risk'
44. #36 OR #37 OR #38 OR #39 OR #40 OR #41 OR #42 OR #43
45. #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR
#24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32 OR #33 OR #34 OR #35

46. #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13

47. #1 OR #2 OR #3 OR #4 OR #5 OR #6

48. #44 AND #45 AND #46 AND #47

Appendix 2: Summary of the included studies (participants' characteristics, outcome measure, psychometric properties tested and authors conclusion)

<i>Study Reference</i>	<i>Population</i>	<i>Outcome measure(s) studied</i>	<i>Psychometric properties tested</i>	<i>Author's conclusion</i>
Almeida (2016) ⁵	Idiopathic Parkinson's Disease (PD) Sample size: 225 Mean age: 70.7±6.6 Disease severity: H&Y stage (median) 3 (2-4) (recurrent fallers), 2.5(1-3) (non-recurrent fallers) Gender: 122M & 103F	ABC FES-I BBS FRT TUG DGI	Discriminant validity	Self-reported measures are as accurate as performance-based measures in predicting recurrent falls over 12 months in people with PD
Almeida (2017) ⁶	PD Sample size: 229 Mean age: 70.6±6.9; n=145 (non-recurrent fallers), 71.1±6.3; n=84 (recurrent fallers). Disease severity: H&Y stage(mean) 1-4(2.6±0.57). Gender: 77M & 68F (non-recurrent fallers), 45M & 39F (recurrent fallers).	BBS FES-I ABC UPDRS Modified H & Y Scale UPDRS-ADL S & E Scale 8-item PDQ FRT TUG DGI	Cut-off scores to predict falls	A 3-predictor tool comprised by history of recurrent falls in the past year, motor fluctuations and UPDRS ADL >12 points, and a 5-predictor tool comprised by these three predictors plus BBS ≤49 points and LED >700 mg/day were developed with moderate to high accuracy for predicting recurrent falls in people with PD within the next year.
Babaei-Ghazani (2017) ⁷	PD Sample size: 100 Mean age: 56.8±15.13 Disease severity:	BBS	Inter-rater reliability Internal consistency Structural Validity	Persian version of the BBS is reliable and valid for the assessment of balance in PD

	H&Y not reported Gender: 57M & 43F			
Baggio (2013) ⁸	PD Sample size (PD): 107 Mean age (PD): 62.1±11.7 Disease severity: H&Y scale mean(range) 2.21±0.79 Gender (PD): 66M & 41F	Clinical GABS UPDRS- ME Pull test FES-I BBS FOGQ H & Y S & E	Intra-rater and inter-rater reliability Internal consistency Construct validity Discriminant Validity Structural validity	The GABS is a reliable and valid instrument to assess gait and balance in patients with PD
Behrman (2002) ⁹	PD Sample size: 43 Mean age: 64.3±9.3 Disease severity: H&Y scale 3 and 4 Gender: 25M & 18F	FRT	Sensitivity Specificity Positive predictive value Negative predictive value	FRT is effective in discriminating PD patients with and without falls.
Bello-Hass (2010) ¹⁰	PD Sample size: 24 Mean age: 64.9±8.0 Disease severity: H&Y scale 1-3 Gender: 18M & 6F	TUG NUDS S & E ADL Scale ABC PDQ – Short form SSE	Internal Consistency Test-retest reliability MDC Convergent validity	ABC, S&E, PDQ-8 and SSE have moderate to excellent internal consistency and reliability. ABC is valid measure in early stages of PD. NUDS is not recommended due to poor psychometric properties.
Benka (2016) ¹¹	PD Sample size : 112 Mean age: 72.8±5.5 Disease severity: H&Y scale 1-3 Gender: 64M & 48F	Mini-BESTest	Structural validity	Findings of the Rasch analysis indicates until better tools emerge, the Mini-BESTest is the best instrument to assess balance in PD

Bergstrom (2012) ¹²	PD and chronic Stroke (1:1) Sample size: PD(n=9) Disease severity: H&Y scale 1-3 Mean age (range): 60.3(46-85)(PD) 78.4(66-90)(Stroke) Gender: 6M & 12F	Mini-BESTest BBS TUG FES	Translation and cultural adaptation Concurrent validity	Swedish version of the Mini-BESTest is suitable for the assessment of balance in PD.
Brincks (2019) ¹³	PD Sample size: 58 Mean age: 68 Disease severity: H&Y 1-4 Gender: 35M & 23F	Six-Spot Step Test TUG Mini-BESTest	Concurrent validity Divergent validity	Six-Spot Step Test is a valid measure of balance in people with PD.
Browne (2002) ¹⁴	PD (n=8), healthy adults (n=70) and older adults with a history of falls (n=9) Sample size:87 Age range: 20-90 years	Quantitative Posturography	Discriminant validity Concurrent validity	Quantitative posturography is found to be sensitive to change in the postural control system due to ageing or disease affecting balance.
Brusse (2005) ¹⁵	PD Sample size: 25 Mean age: 76±7 Disease severity: H&Y scale 1-3 Gender: 14M & 11F	BBS FFR and BFR TUG Gait speed UPDRS-ME	Inter-rater reliability Concurrent validity	BBS but not the UPDRS scores strongly correlated with comfortable and fast gait speed, TUG and BFR. BBS needs to be added to the traditional PT examination of people with PD. UPDRS is not comprehensive enough.
Candan (2019) ¹⁶	PD patient Sample size 48 Median(range): 71(47-84) Disease severity: Y&H scale median(range) 1.5(1-4) Gender: 31M & 17F	FOGQ UPDRS Modified H&Y scale BBS FES	Inter-rater reliability Intra-rater reliability Cross-cultural validation	The new Turkish version of FOG-Q could be used to quantify FOG in PD. Moreover, it could also be used in experimental research to determine the best treatment method

		TUG FTSST		for improvement of FOG in native Turkish-speaking patients with PD.
Chomiak (2015) ¹⁷	PD Sample size: 27 Mean age: 67.1±10.2 <u>Disease severity:</u> <u>Average UPDRS-III score of (16.6±7.1, n=25)</u>	SLST UPDRS-ME	Test-retest reliability Concurrent validity	Scores of below 10 seconds in SLST marks a clinically important stage of disease progression with significant worsening of postural stability in PD.
Chong (2012) ¹⁸	Idiopathic PD Sample size: 48 Disease severity: H&Y scale stage(mean), UPDRS section 3 score 1-2(1.7±0.4), 16.7±7.8; 2.5-4(1.7±0.4), 29.9±10.6)	RAPID (ADL, FOF and NOFs)	Cut-off score Sensitivity Specificity	The RAPID questionnaire is sensitive and specific for the assessment for postural instability in PD.
Claesson (2017) ¹⁹	PD Sample size: 28 Mean age: 67.6 Disease severity: H&Y scale 1-3 Gender: 11M & 17	BDL BBS UPDRS-ME TUG CTUG	Internal consistency Concurrent validity Predictive validity Convergent validity Standard error of measurement	BDL is a valid clinical assessment for balance in people with PD.
Combs (2014) ²⁰	PD Sample size: 88 Mean age: 66.4±9.3 Disease severity: H&Y range 1-4 Gender: 61M & 27F	Comfortable 10 meters walk test Fast 10 meters walk test 6MWT Mini-BESTest ABC PDQ 39-M	Test re-test reliability MDC Discriminant validity Convergent validity Responsiveness: reported but not tested	Short distance walking tests can be used as a reliable measure that is responsive to changes over time in patients with PD.
de Silva (2017) ²¹	PD Sample size: 50 Mean age: 67.4±9.0	UPDRS-ME TUG Section VI of BESTest	Floor and ceiling effect Inter-examiner reliability, Intra-examiner reliability,	The TUG-ABS is found to be reliable and valid with accuracy to identify the biomechanical

	Disease severity: H&Y scale 1-4 MMSE 24.6±3.9 UPDRS-III 13.7±4.6 Gender: 50M & 50F	TUG-ABS	Test-retest reliability, MDC Internal consistency, Construct validity Discriminant validity	characteristics and strategies used by individuals with PD.
Dibble & Lange (2006) ²²	Idiopathic PD Sample size: 45 Mean age: 69.94±11.28 Disease severity: H&Y stage(mean) 1-4(2.60±0.66) Gender: Not reported	FRT BBS DGI TUG CTUG	Sensitivity Specificity	Regardless of the performance on any individual clinical balance test, the multifactorial nature of postural instability in PD may necessitate a battery of tests to provide the most accurate identification of fall risk.
Dibble (2008) ²³	PD Sample size: 36 Mean age: 75.25±2.15 Disease Severity: H&Y scale 1-4 Gender: 28M & 8F	BBS DGI FRT TUG	Construct validity	This study provides evidence to show collective interpretation of multiple clinical tests have better clinical utility in diagnosing falls risk in people with PD
Duncan (2011) ²⁴	PD Sample size: 80 Mean age: Disease severity: H&Y scale 1-4 Gender: Not reported	Mini-BESTest FTSTS Quadriiceps Dynamometry 6MWT 9HPT FOGQ ABC PDQ-39-M UPDRS	Inter-rater reliability Test-retest reliability Cut-off scores to predict falls	FTSTS does not relate to the PD disease stage. FTSTS is reliable and easy to assess falls risk in people with PD.
Duncan (2012) ²⁵	PD Sample Size: 51(6), 40(12)	BBS FGA BESTest	Predictive validity	BESTest was slightly better than the other tests in predicting balance. None of the tests eliminated false-

	Mean age: 67.5±8.8(6); 67.3±9.5(12) Disease severity: H&Y stage: 2.4±0.6(6), 2.3±0.6(12) Gender: Not reported	Mini-BESTest		positive and false negative predictions.
Duncan (2015) ²⁶	PD Sample size: 171 Mean age: 65.91±9.13(non-fallers) 68.55±9.53(fallers) Disease severity: H&Y scale stage 1-4 Gender: 97M & 74F	FOGQ UPDRS-ME	External validity Discriminant validity	The results validated the utility of a tool to accurately identify falls risk in people with PD
Duncan (2015a) ²⁷	PD Sample size: 80 Mean age: 68.2±9.3 Disease severity: H&Y scale range 1-4 Gender: 47M & 33F	BESTest Mini-BESTest BBS UPDRS-ME FOGQ Falls history PASE	Predictive validity	BESTest or Mini-BESTest may be used to assess postural instability once in 6 months in people with PD. BESTest and Mini-BESTest may be preferred over the BBS.
Duncan (2013) ²⁸	PD Sample size: 81 (off:28, on:53) Mean age: 70±7.4 (off), 68±8.5(on) Disease severity: H&Y scale range 1-5 Gender: 46M & 35F	FSST UPDRS-ME Mini-BESTest FTSTS 6MWT 9HPT	Inter-rater reliability, Test-retest reliability Discriminant validity	FSST is not recommended in lieu of other balance measures such as the Mini-BESTest in people with PD.

Foreman (2011) ²⁹	PD Sample size: 36 Mean age: 70.95±11.41 (Fallers) 66.64±10.05 (Non-fallers) Disease severity: Modified H&Y scale median(range): 2.5(1.5-4)(Faller-ON) 3(2.5-4)(Faller-OFF) 2.25(1.5-2.5)(Non-faller-ON) 2.5(1.5-3)(Non-faller-OFF) Gender: 24M &12F	UPDRS-ME FGA Pull Test TUG	Discriminant validity	The pull test was less accurate as a predictor of falls compared to FGA and the TUG.
Franchignoni (2005) ³⁰	PD sample size: 70 Mean age: 71 (range 41-81) Disease severity: median 3 (1.5-4) Gender: 37M & 33F	BBS FFM PCS Tandem Romberg SLST FRT TUG UPDRS II and III S & E	Internal consistency Construct validity	This study confirms people with more severely affected balance or fear of falls tend to perform poorly on other balance, mobility and posture tests.
Franchignoni (2014) ³¹	PD Sample size: 217 Mean age: 71 (range 48-83) Disease severity: Not reported Gender: Not reported	ABC 6-items ABC by Peretz (ABC-6P) 6 items ABC by Oude Nijhuis (ABC-6OM) 5 items ABC by Lohnes (ABC-5L) FOF measure BBS UPDRS-ADL UPDRS-ME H & Y S & E ADL scale	Internal consistency Convergent and divergent validity Structural validity Rasch analysis	ABC showed adequate unidimensionality and Rasch analysis showed minor psychometric weaknesses of the 3 short versions when compared to the ABC scale.

Freeman (2018) ³²	PD Sample size: 26 Mean age: 66±6.9 Disease severity: H&Y 2.0 (range 2-3) Gender: 18M & 8F	SOT mCTSIB	Concurrent validity Discriminant validity	Instrumentation using body-worn movement monitors was shown to be efficient, quantitative alternative in the measurement of postural control in PD. The mCTSIB discriminated between fallers and non-fallers.
Giladi (2000) ³³	PD patient Sample size: 40 Mean(SD): 72.3±9.3 Disease severity: H&Y "OFF" phase mean(SD) 2.85±0.84 Gender: 26M & 14F	FOGQ UPDRS-Total UPDRS-Mental UPDRS-ADL UPDRS-Motor H&Y scale	Internal consistency Construct validity	The FOGQ a reliable and simple questionnaire that assesses freezing of gait in parkinsonian patients, which can be used in future clinical trials of FOG assessment.
Goetz (2008) ³⁴	PD patients Sample size: 877 Mean age: Nil Disease severity: H&Y range 1-5 Gender: 560M & 317F	MDS-UPDRS	Internal consistency Criterion-related validity	The combined clinimetric results of this study support the validity of the MDS-UPDRS for rating PD.
Harro (2016) ³⁵	PD Sample size: 42 Mean age: 66.21±7.92 Disease severity: H&Y 2.33±0.77 Gender: 22M & 20F	Balance measures: Mini-BESTest FGA Gait measures: 10MWT 6MWT Force platform measures: LOS MCT SOT	MDC Test re-test reliability SEM Convergent validity	The SOT, LOS and MCT have excellent test re-test reliability, the SOT and LOS have fair to good correlation with the clinical balance measures.

Holmes (2012) ³⁶	PD Sample size: 20 Mean age: 67±8 Disease severity: H&Y scale range 2-3 Gender: 13M & 7F	Balance task using Nintendo Wii© during Eyes open feet apart Eyes closed feet apart Eyes open feet together Eyes closed feet together	Concurrent validity	Balance board can be adopted as a new clinical tool for the assessment of postural instability in patients with PD.
Huang (2011) ³⁷	PD Sample size: 78 Mean age: 67.5±11.6 Disease severity: H&Y scale range 1 to 3 Gender: 48M & 30F	TUG DGI	MDC Test re-test reliability Convergent validity	Both TUG and DGI have acceptable reliability and random measurement error in the assessment of balance in patients with PD.
Jacobs (2006) ³⁸	PD and age-matched healthy control Sample size PD (n=67), control (n=65) Mean age (PD): 67±12 Disease severity: Not reported Gender: 47M & 20F	Pull test SLST FRT UPDRS items 27-29 ABC NOFs	Discriminant validity Construct validity	Multiple balance tests and UPDRS items 27-29 improved the assessment of balance and falls in PD. Combination of the One-legged stance test, pull test and the UPDRS items 27-29 provided the most informative balance confidence and falls assessment.
Jacobs (2006a) ³⁹	PD and age-matched healthy adults Sample size: PD (n=88), healthy adults (n=75) Mean age: (PD) 67±12 Disease severity: N/A Gender; 61M & 27F	Push and Release test Pull test ABC	Inter-rater reliability Sensitivity and specificity Construct validity	Push and release test provides high sensitivity and Pull test provides high specificity. They both can be used together. Third trial of the Push and Release test provides more valid, consistent and sensitive assessment of postural stability in PD
Jenkins (2010) ⁴⁰	PD and age-matched control	PST FRT	Construct validity	UPDRS may not be sufficient to evaluate postural instability, it is

	Sample size: PD (n=20), healthy control (n=20) Mean age (PD): 65.95±8.32 Disease severity: H&Y median score 2.25 Gender: 12 M & 8F	'Real-World' MRM		recommended to add FRT as a standard assessment to identify falls risk in people with PD
Joanna DiFrancisco (2016) ⁴¹	PD Sample size: 39 Mean age: 70.8±9.9 Disease severity: Not reported Gender: 29M & 10F	UPDRS-ME Mini-BESTest SOT	Discriminant validity Sensitivity, specificity and cut-off scores	SOT is found to be reliable in differentiating fallers and non-fallers with PD
Jonasson (2014) ⁴²	PD sample size: 102 Mean age: 73 Disease severity: Not reported Gender: 68M & 43F	mSAFFE FES-I FES-I (S) ABC	Discriminant validity Ceiling and floor effect Test re-test reliability SEM	All 4 scales had acceptable psychometric properties. Since ABC and FES (S) had many outliers this study favors FES-I or mSAFFE for assessing fear of falls.
Jonasson (2017) ⁴³	PD sample size: 101 Mean age: 73 Disease severity: Not reported Gender: 68M & 43F	FES-I FES-I short version	Discriminant validity Ceiling and floor effect Test re-test reliability SEM	Both scales have good psychometric properties. Owing to the high floor effect of the short FES-I this study favors the original FES-I for longitudinal follow-ups in PD
Kegelmeyer (2007) ⁴⁴	PD Sample size: 156 Mean age: 68.8±11.04 Disease severity: H&Y 2.5 (range 1-5) Gender: 99M & 57F	TMT UPDRS-ME Comfortable gait speed	Inter-rater reliability Intra-rater reliability Concurrent validity Criterion validity Discriminant validity	TMT is a reliable and valid tool for assessing balance, gait and falls risk of individuals with early and middle stages of PD
King (2012) ⁴⁵	PD Sample size: 97 Mean age: 65.6±7.1	Mini-BESTest BBS UPDRS	Criterion validity Discriminant validity	Mini-BEST might be more useful than BBS especially in patients with mild PD and subtle balance deficits

	Disease severity: H&Y 2.3±0.6 Gender: 59M & 38F	H & Y		
Kleiner (2018) ⁴⁶	PD Sample size: 30 Mean age: 69±7.02 years Disease severity: H&Y scale (mean): 2.85 ± 0.32; UPDRSIII: 30.16±6.50; MMSE: 27.03±2.38; Gender: 15M & 15F	TUG stopwatch TUG optoelectronic system TUG-IMU	Reliability (reproducibility by different systems)	The IMU shows excellent reliability, accuracy and precision in quantifying the TUG Test completion times, similar to those obtained using the optoelectronic system (the gold standard). As compared to the stopwatches widely used in clinical settings, IMU allows more information about patient's performance to be obtained and reduces subjectivity in outcome measures.
Koop (2018) ⁴⁷	Idiopathic PD Sample size: 30 Mean age: 61.9±9.0 Disease severity: H&Y range 2-3 Gender: 18M & 12F	TUG CC-MB (Cleveland Clinical Mobility and Balance Application) MDS-UPDRS III Motor scores	Construct validity Test-retest reliability	Using a single IMU in conjunction with the CC-MB application, we detected significant improvements from anti-PD medications in kinematic and timing measures during the Gait and Turning components of the TUG, that demonstrated excellent test-retest reliability while increasing objectivity for the assessment of mobility status.
Leddy (2011) ⁴⁸	PD Sample size: 80 Mean age: 68.2±9.3 Disease severity: H&Y scale 2.45±0.64 Gender: 47M & 33F	BESTest Mini-BESTest	Inter-rater reliability Test re-test reliability Discriminant validity	Mini-BESTest is a reliable measure of balance in PD. The mini-BESTest is comparable to the BESTest in its ability to discriminate between fallers and non-faller and more

feasible in clinical use as it takes shorter duration to complete.

Leddy (2011a) ⁴⁹	PD Sample size: 80 Mean age: 68.2±9.3 Disease severity: H&Y scale 2.5±0.64 Gender: 47M & 33F	FGA BESTest BBS ABC	Inter-rater reliability Test re-test reliability Criterion validity Discriminant validity with sensitivity and specificity	BBS is not suitable for early stages of postural instability due to significant ceiling effect. Both FGA and BESTest are reliable and valid for measuring balance throughout stages 1-4 on H&Y scale. Both tests can discriminate between fallers and non-fallers.
Lindholm (2016) ⁵⁰	PD patients Sample size: 138 Mean age: 67±9.8years Disease severity: H&Y median(min-max) 2(1-4) Gender: 74M & 64F	Three-steps falls prediction model	Cut-off scores to predict falls	This study confirms the value of the Three-step model as a clinical fall prediction tool and illustrates that it outperforms the use of single predictors.
Lofgren (2014) ⁵¹	PD Sample size; 27 Mean age: 73±4 Disease severity: H&Y (range 2-3) Gender: 18M & 9F	Mini-BESTest	Inter-rater reliability Test re-test reliability	The mini-BESTest showed good inter-rater and test re-test reliability in people with mild to moderate PD.
Lofgren (2017) ⁵²	PD and healthy control Sample size: PD (n=104) and control (n=47) Mean age: 73±5.5 Disease severity: H&Y (range 2-3) Gender: 27M & 20F	Mini-BESTest TUG Dual task TUG UPDRS part II	Hypothesis testing (known group difference) Convergent validity Divergent validity	The findings suggests Mini-BESTest to have adequate psychometric strength to assess balance in people with PD

Maia (2013) ⁵³	PD and age matched healthy elders. Sample size; PD (n=35) and control (n=35) Mean age: 66.5±10.3 Disease severity: H&Y 2.02±0.8 Gender: Not reported	BESTest Mini-BESTest	Test re-test reliability Rasch analysis	Both BESTest and Mini-BESTest are reliable and valid to assess balance in PD
McKee (2014) ⁵⁴	PD and healthy age matched peers Sample size: PD (31) and older adults (24) Mean age: 69.65±7.7 Disease severity: H&Y scale 2±2.0 Gender: 12F & 19M	FSST UPDRS-ME TUG 6MWT Thirty second chair stand BBS Gait speed Dual TUG ABC	Inter-rater reliability Concurrent validity	FSST involving negotiation of raised obstacles is feasible, reliable and a valid measure of motor performance in PD.
Mehdizadeh (2019) ⁵⁵	PD Sample size: 124 Mean age: 60.33±12.59 Disease severity: H&Y range 1-3 Gender: 90M & 34F	FES-I VAS-FOF BBS FRT PDQ-39 UPDRS-ADL H & Y Scale	Internal consistency Test-retest reliability Convergent validity Discriminant validity	Our study suggests that the FES-I is a reliable and valid scale for measuring fear of falling in individuals with PD. Also, it is an appropriate scale for clinicians and researchers to use during the On- and Off-drug phases.
Morris (2001) ⁵⁶	PD and age-matched healthy peers Sample size: 12 Mean age: 68.8±10.4 Disease severity: N/A Gender: 5M & 7F	TUG	Test re-test reliability Agreement between experienced and inexperienced assessors. Sensitivity of TUG detecting change	Reliability of the TUG was high and the measurement reflected changes in performance according to the use of Levodopa.

Nilsson (2010) ⁵⁷	PD Sample size: 79 Mean age: 64.5±7.2 Disease severity: H&Y scale not reported. Gender: 56M & 23F	FES SAFFE SF-36 UPDRS Part II and III Fast gait speed TUG Physical function (part of SF-36)	Internal consistency Test re-test reliability SEM Floor and ceiling effect Construct validity	This study supports the reliability and validity of the FES and SAFFE in people with PD
Nilsson & Hagell (2009) ⁵⁸	PD patient Sample size: 37 Median(range): 67(56-73) Disease severity: H&Y scale median(range) 3.0(1-5). Gender: 29M & 8F	FOGQ UPDRS H&Y FES TUG SF36 Gait speed SF-36	Internal consistency Cross-cultural validation	In conclusion, we have replicated and extended the findings in the original validation study of the FOGQ and provide initial support for the measurement properties of its Swedish version
Ozinga (2017) ⁵⁹	PD and age-matched peers without PD Sample size: 14 Mean age 62.9±8.3 Disease severity: H&Y 2.4±0.5 Gender: 7M & 7F	SOT COM using mobile device	Test re-test reliability	Use of acceleration measures using the mobile device increases the sensitivity to discriminate between people with and without PD.
Parashos (2015) ⁶⁰	PD Sample size: 340 Mean age: 71.40±9.06 Disease severity: H&Y (range 1-5) Gender: Not reported	ACM/PIGD H & Y Scale ABC NOF FOG-Q FTSS TUG GV (Gait velocity)	Internal consistency Construct validity	We conclude that the ACM, a score comprised of UPDRS items 13, 14, 15, 29, and 30, and the PIGD, consisting of the average score of the same UPDRS items, are valid measures of gait and ability to ambulate for PD patients with HY stages 1–4.

Park (2018) ⁶¹	PD Sample size: 24 Mean age: 72.33±6.95 Disease severity: H&Y scale 2.39±7.37 Gender: 14M & 10F	TMT BBS TUG 10m walk test UPDRS-ME	Concurrent validity Intra and inter-rater reliability Predictive validity	Korean version of TMT has acceptable reliability and validity for assessing gait and balance in PD
Paul (2012) ⁶²	PD Sample size: 31 Disease severity: H&Y 2.0±0.8 (range 1-4) Disease severity: H&Y stage(mean) 1-4(2.0±0.8); MMES mean 29.6±0.9 Gender: 22M & 9F	Sway on the floor with eyes open Sway on the floor with eyes closed Sway on foam with eyes open Sway on foam with eyes closed Maximal balance range Coordinated stability Choice stepping reaction (R) lateral reach (L) lateral reach Five-repetition sit-to-stand TUG TUG-cognitive	Test re-test reliability Standard error of measurement	High reproducibility for the mobility measures confirms their suitability as outcome measures for people with PD
Paul (2013) ⁶³	Idiopathic PD Sample size: 205 Mean age: Nil (59% fall atleast once over 6 months follow-up) Disease severity: H&Y stage(mean) 1-4(2.60±0.6) Gender: 55M & 35F (non-recurrent fallers), 63M & 57F (recurrent fallers)	5TSTS Motor section of UPDRS UPDRS-32&33 UPDRS-28 FRT Tande Stand FES-I Near Tandem Stand Alternate S+G9tep Test Postural sway	Cut-off scores to predict falls	The simple clinical prediction tool developed in this study can help clinicians to identify individuals with PD who are at particularly high risk of falling to enable the timely delivery of preventive and minimization strategies.
Paul (2016) ⁶⁴	PD, healthy young adults and healthy older adults	Push and release task Sit to stand	Criterion validity Inter-rater reliability	2D system of video analysis was found to be a reliable and valid

	Sample size: 15 PD and 15 in each control Mean age: (PD) 71±71. Disease severity: UPDRS-ME 36.1±9.7 Gender: 9M & 6F	Single leg stance ± acutely induced dizziness	Test-retest reliability	method of quantifying postural outcomes in healthy and PD subjects.
Peretz (2006) ⁶⁵	PD, higher level gait disorders (HLDG) and control Sample size: PD (19), HLDG (70) and control (68) Mean age: (PD) 72 ±6 Disease severity range: H&Y stage 1-3 Gender: (PD) 12M & 7F	ABC (16 items) Short version ABC (6 items)	Discriminative validity Internal consistency Reliability	The short version of ABC has properties comparable to the parent long version ABC.
Petersen (2017) ⁶⁶	PD Sample size: Mean age: 72±8.5 Disease severity: H&Y scale ranged from 1 and 3 Gender: 15M & 7F	FGA 5×STS 30sSTS	Test re-test reliability MDC Criterion validity Internal consistency	The 30sSTS and the FGA can reliably assess change in balance over time. A lower FGA score is associated with higher chance of falls.
Porta (2015) ⁶⁷	PD Sample size: 285 Mean age: 71.2±7 Disease severity: Modified H&Y scale 46±9.7 Gender: 130M & 155F	BBS Modified H & Y UPDRS part II and III	Content validity	The analysis revealed with some modifications, the BBS had adequate internal construct validity and reliability for measuring PD patients.
Qutubuddin (2005) ⁶⁸	PD Sample size: 38 Mean age: 71.1±10.5 Disease severity: H&Y range 1.5-4	BBS UPDRS S & E ADL scale Modified H & Y scale	Criterion-related validity	Results supports the criterion-related validity of the BBS in people with PD

	Gender: 38M & 0F			
Richard (1994) ⁶⁹	PD Sample size: 24 Mean age: 70.7±9.4 Disease severity: Not reported Gender: 13M & 11F	UPDRS-ME	Inter-rater reliability	Overall the reliability of the scale is satisfactory.
Ross (2017) ⁷⁰	PD Sample size: 159 Mean age: 66.6 (10.3) Disease severity: H&Y scale range 1-4 Gender: 105M & 54F	FRAC	Inter-rater reliability	The FARC is a reliable tool for classifying PD fallers.
Scalzo (2009) ⁷¹	PD Sample size: 53 Mean age: 62±7.9 Disease severity: H&Y scale, mean (range)- 2.5 (1-3) Gender: 37M & 16F	BBS (Brazilian version) UPDRS (I, II and III) H & Y staging scale S & E ADL Scale	Inter-rater reliability Internal consistency Criterion-related validity Structural validity	Brazilian version of the BBS is suitable for use. BBS correlated with severity of symptoms, disease stage and level of independence
Schenkman (1997) ⁷²	PD Sample size: 34 Mean age: 74.5 ±5.7 Disease severity: H&Y scale 2 and 3 Gender: 32M & 2F	FRT 360-degree turn 6MWT 10MWT	Variation by day or week of testing Test re-test reliability	The measures of physical performance are relatively stable for early and middle stage PD patients.
Schlenstedt (2015) ⁷³	PD Sample size: 85 Mean age: 67.8±9.8 Disease severity: H&Y scale 2.7±0.7	FAB BBS Mini-BESTest UPDRS VAS	Concurrent validity Test re-test reliability inter-rater reliability	FAB is a reliable and valid tool for assessing postural control in PD. In comparison to BBS and Mini-BESTest the FAB is shorter and takes less time to perform.

Gender: 57M & 28F

Schlenstedt (2016) ⁷⁴	PD Sample size: 66 Mean age: Fallers (68.1±7.5) and non-fallers 66±11.6. Disease severity: H&Y scale mean(range) 2.5(1-4). Gender: 45M & 21F	FAB BBS Mini-BESTest	Predictive validity: sensitivity and specificity to discriminate between fallers and non-fallers	FAB, Mini-BESTest and BBS showed moderate predictive validity to differentiate between fallers and non-fallers with PD for a prospective period of 6 months
Spagnuolo (2018) ⁷⁵	PD patients Sample size: 30 Mean age: 65.53±6.45 Disease severity: H&Y (range 1-4) Gender: 13M & 17F	TUG FTSTS BMT (Bed mobility test)	Responsiveness	The present study shows that TUG, FTSTS and BMT, were responsive to an 8-week evidence-based GPTI applied to PD providing important information about the clinical and research application of these tests.
Steffen (2008) ⁷⁶	PD Sample size 35 Mean age: 71±12 Disease severity Average H&Y 2 (range 1-4). Gender: 26M & 11F	BBS Forward and backward reach RSR ABC 6MWT Comfortable and fast gait speed TUG SF-36 UPDRS	Test re-test reliability Minimal detectable change	There is a wide range of outcome measures available for assessing balance and ambulation. The minimal detectable change will help therapist to determine if the change in score is due to measurement error or a true difference in the domain tested.
Stocchi (2018) ⁷⁷	PD Sample size: 194 Mean age: 66.51±9.34 Disease severity: H&Y scale 57.2% stage 1 or 2, 36.6% stage 3 and 6.2% stage 4 or 5 Gender: Not reported	UPDRS-MDS PDCS (Others) PD Sleep Scale Version 2 UPDRS-MDS-IV for complications SEND-PD CISI-PD	Acceptability: floor and ceiling effect Internal consistency Dimensionality (structural validity) Reproducibility	PDCS is a feasible, acceptable, reproducible, valid and precise measure of disease severity including motor, non-motor and disease complication in PD.

			Hypothesis testing: convergent, internal and known-groups validity	
Taghizadeh (2018) ⁷⁸	Idiopathic PD Sample size: 98 Mean age: 59.19±10.88 Disease severity: H&Y (range 1-4) Gender: 73M & 25F	BBS FES-I FRT UPDRS-II PDQ-39 S & E ADL Scale	Internal consistency Inter-rater reliability Intra-rater reliability Discriminant validity Convergent validity	The results of our study suggest that the BBS has acceptable reliability and validity to evaluate the functional balance during the drug off-phase in idiopathic PD.
Thomas (2004) ⁷⁹	PD Sample size: 35 Age range: 50 and 75 Disease severity: H&Y scale range 1-3 Gender: 15M & 20F	GABS Balance Master GAITRite	Intra-rater reliability Concurrent validity	GABS is found to be highly reliable, valid, well-constructed, comprehensive, easy to use clinical scale for gait, balance, posture, freezing of gait and gait cycle in PD
van Lummel (2016) ⁸⁰	PD Sample size: 28 Mean age: 67.1 ± 8.3 Disease severity: Median score 3 on Modified H& Yahr score. Gender: 22M & 6F	(Instrumented) i-TUG: timed up and go test measured with inertial sensor measurement system to record accelerational angular velocity in three directions	Intra-rater reliability Inter-rater reliability test re-test reliability	The i-TUG has excellent to good for total and turning duration.
Visser (2003) ⁸¹	PD and control Sample size PD (42) and control (15). Mean age: PD-Stable 62.7 ± 9.5 (13M & 7F) PD-unstable 66.3±11.9 (9M & 13F) Control 64.3±9.5 (7M & 8F)	1. (Nutt et al protocol) Reaction to an unexpected shoulder pull scored on a 4-point scale 2. (Bloem et al protocol) Reaction to an unexpected shoulder pull scored on a 4-point scale, this test included speed of restoring balance 3. (UPDRS) Reaction to an unexpected shoulder pull scored on a 5-point scale	Inter-rater reliability Concurrent validity Criterion validity	Unexpected shoulder pull, executed once, with taking more the 2 corrective steps backwards is considered the most valid test for postural instability in PD. Retropulsive test scored according to Nutt has good inter-rater reliability.

	Disease severity: Not reported Gender: N/A	4. (SPES) Evaluating the reaction to an expected shoulder pull, and scored on a 4-point scale 5. (Pastor et al) evaluating the reaction to an expected shoulder pull, and scored on a 5-point scale 6. Rating steady-stance positions		
Wallen (2016) ⁸²	PD Sample size: 112 Mean age (72.8±5.5) Disease severity: H&Y (range 1-3), UPDRS motor mean 32±11 Gender: (64M & 48F)	Mini-BESTest	Structural validity	Until a better test emerges, the Mini-BESTest may indeed be the best instrument for assessing balance in people with PD.
Yang (2014) ⁸³	PD Sample size: 121 Mean age (61.8 ±11.5) Disease severity: H&Y (range 1-4) Gender: 82M & 39F	FGA BBS FAC DGI TUG ABC UPDRS-ME Fast walking speed Modified H & Y Falls	Construct validity, concurrent validity and predictive validity	The FGA demonstrated good construct validity to evaluate balance and gait instability in patients with PD. It showed moderate to strong correlation with other balance and gait measures. An FGA cutoff score of 18/30 provides optimum predictive validity for falls in patients with PD within the 6 months after hospital discharge.
Yang (2016) ⁸⁴	PD Sample size: 121 Mean age (61.8 ±11.5) Disease severity: H&Y (range 1-4) Gender: 82M & 39F	FGA	Inter-rater and intra-rater reliability and internal consistency	FGA has high external and internal reliability.

Abbreviations: ABC: Activities-specific Balance Confidence, FES-I: Falls Efficacy Scale-International, BBS: Berg Balance Scale, FRT: Functional Reach Test, TUG: Timed Up

and Go Test, DGI: Dynamic Gait Index, GABS: Gait and Balance Scale, UPDRS-ME: Unified Parkinson's Disease Rating Scale- Motor Examination, FOGQ: Freezing of Gait Questionnaire, H & Y: Hoehn and Yahr scale, S & E: Schwab & England scale, NUDS: Northwestern University Disability Scale, S & E ADL scale: Schwab & England Activities of Daily Living Scale, PDQ – Short form: Parkinson's Disease Questionnaire - Short form, SSE: Stanford Self-Efficacy for Managing Chronic Disease 6-item Scale, BESTest: Balance Evaluation System Test, FES: Falls Efficacy Scale, QP: Quantitative Posturography, FFR: Forward Functional Reach Test, BFR: Backward Functional Reach Test, SLST: Single Leg Stance Test, RAPID: Rapid assessment of postural instability in Parkinson's Disease Questionnaire, ADL: Activities of Daily Living, FOF: Fear of Falling, NOFs: Number of Falls, BDL: Backstrand Dahlberg Liljenas Balance Scale, CTUG: Timed Up and Go cognition, 6MWT: 6 Minute Walk Test, PDQ 39-M: Parkinson's Disease Questionnaire-39-M, TUG – ABS: Timed Up and Go Assessment of Biomechanical Strategies, FTSTS: Five Times Sit-to-Stand Test, 9HPT: Nine Hole Peg Test, PASE: Physical Activity Scale for the Elderly, UPDRS: Unified Parkinson's Disease Rating Scale, FGA: Functional Gait Assessment, FSST: Four Square Step Test, FFM: Fear of Falls Measure, PCS: Postural Change Scale, ABC-6P: 6-items ABC by Peretz, ABC-6OM: 6 items ABC by Oude Nijhuis, ABC-5L: 5 items ABC by Lohnes, UPDRS-ADL: Unified Parkinson's Disease Rating Scale-Activities of Daily Living, H & R-ADL: Hoehn and Yahr Activities of Daily Living scale, SOT: Sensory Organization Test, mCTSIB: Modified Clinical Test for the Sensory Integration of Balance, 10MWT: 10 Meter Walk Test, LOS: Limit of Stability Test, MCT: Motor Control Test, UPDRS items 27-29: Unified Parkinson's Disease Rating Scale items 27-29, P&RT: Push and Release Test, PST: Postural Stability Test, MRM: Maximal Reach Measurement, mSAFFE: Modified Short Version of Survey of Activities and Fear of Falling in the Elderly, FES-I (S): Swedish Falls Efficacy Scale-International, TMT: Tinetti Mobility Test, SAFFE: Short Version of Survey of Activities and Fear of Falling in the Elderly, SF-36: 36-item Short-form Health Survey, COM: Center of Mass, SLS: Single Leg Stance, 5×STS: 5 Times Sit to Stand, 30sSTS: 30 Second Sit to Stand, FRAC: Fall-related Activity Classification, FAB: Fullerton Advanced Balance scale, VAS: Visual Analog Scale, RSR: The Romberg and the sharpened Romberg with eyes open and closed, UPDRS-MDS: Movement Disorder Society Unified Parkinson's Disease Rating Scale items, PDCS: Parkinson's Disease Composite Scale, SEND-PD: Scale for Evaluation of Neuropsychiatric Disorders in PD, CISI-PD: Clinical impression of Severity Index, iTUG: Instrumented Timed Up and Go Test, SPES: Short Parkinson's Disease Scale, FAC: Functional Ambulatory Category

Appendix 3: Outcome measures assessing balance followed by falls risk prediction at the body, structure and function level. COSMIN quality of evidence reported as *. Poor-*, fair-, Good-*** and Excellent-******

Measures of balance (n=10), domain tested and interpretation	References	Key findings of psychometric analysis
Balance Master <i>Domain:</i> To measure the position of center of gravity and assess postural control. <i>Interpretation:</i> Shorter time and accuracy to reach a target indicates better postural control.	<i>Thomas (2004)⁷⁹</i>	<i>Concurrent validity:</i> GABS significantly correlated with Balance Master data (posture, postural stability, balance during stance, single limb stance, tandem stance, turning, toe walking and FR (Spearman correlation coefficient r 0.46 to 1)**.
Center of Mass (COM) using mobile device. <i>Domain:</i> Measures the changes in the center of gravity in standing using the inverted pendulum model. <i>Scoring:</i> 0 to 100%. <i>Interpretation:</i> Higher scores indicate lower postural sway.	<i>Ozinga (2017)⁵⁹</i>	<i>Test-retest reliability:</i> Reliability of mobile device exceeded the reliability of NeuroCom (SOT) for all four metric characterizing COM acceleration. Mobile device and NeuroCom were equally effective in differentiating postural stability across population*.
Limits of Stability Test (LOS) <i>Domain:</i> Assesses postural control by quantifying	<i>Harro (2016)³⁵</i>	<i>Test-retest reliability</i> ICC: LOS average EPE: 0.87, LOS no of falls: 0.63**. <i>Measurement error:</i> LOS average EPE: 5, LOS no of falls: 0.5**.

individual's active limit of stability as well as movement excursion.

Interpretation: Scores close to 100% are indicative of normal balance.

Maximal Balance Range (MBR)

Paul (2012)⁶²

Domain: It uses the sway-meter to measure the maximal anterior-posterior displacement from the ankle joint.

Interpretation: Longer distance reached in "mm" indicating better postural control.

Motor Control Test (MCT)

Harro (2016)³⁵

Domain: Assesses the ability of a person to recover quickly and automatically from an external perturbation.

Interpretation: Shorter latency and fast amplitude scaling signify better postural control.

Nintendo Wii©

Holmes (2012)³⁶

Domain: Measure of standing balance.

Interpretation: Narrower center-of-pressure path length (in cm) indicates better

Convergent validity:

Correlation between balance measures and SOT composite equilibrium:

r range 0.44 to 0.51, LOS average EPE: 0.34 to 0.48, MCT average latency: -0.14 to -0.48

MCT average amplitude: -0.01 to 0.031, MDC: LOS average EPE: 13.8, LOS no of falls: 0.5**.

Test-retest reliability:

ICC: 0.81**.

Measurement error: SEM: 17**.

Test-retest reliability:

ICC: MCT average latency: 0.92, MCT average amplitude: 0.92**.

Measurement error:

SEM: MCT average latency: 2.7, MCT average amplitude: 1**.

Hypothesis testing:

Convergent validity: Correlation between balance measures and SOT composite equilibrium: r range 0.44 to 0.51, LOS average EPE: 0.34 to 0.48.**

MDC: MCT average latency: 7.4, MCT average amplitude: 2.7**.

Hypothesis testing: Concurrent validity was shown to be excellent across all 4 balance tasks (ICC 0.92 to 0.98)*.

balance.

Pull Test

Foreman (2011)²⁹

Domain: To Measure postural instability.

Jacobs (2006a)³⁸

Jacobs (2006)³⁹

Scoring: 0 to 4 points.

Interpretation: Lower scores indicate greater postural stability.

Intra-rater reliability: First trial: 0.45, third trial: 0.74*.

Construct validity: Except the One-leg stance times the other 3 tests had a significant difference between PD and control subjects ***.

Hypothesis testing: Discriminant Validity Except the One-leg stance times the other 3 tests had a significant difference between PD and control subjects***.

Responsiveness: Pull test was significantly different between ON and OFF medication. Pull test ON mean (sd) 0.94 (0.83), Pull test OFF Mean (sd) 1.28 (0.74) $p < 0.006$ **.

Push and Release Test (P&RT)

Jacobs (2006a)³⁹

Paul (2016)⁶⁴

Domain: Measure for early detection of balance impairments in patients with PD before experiencing falls.

Scoring: 0 to 4 points.

Interpretation: Lower scores indicate better balance.

Inter-rater reliability: Push release, Single leg stance and single leg stance with acutely induced dizziness had excellent inter-rater reliability (ICC 0.99 to 1.00)*.

Intra-rater reliability: First trial: 0.84, third trial: 0.83*.
(ICC 0.77 to 1.00)*.

Criterion validity: Agreement between 2D and 3D was excellent (ICC 0.96 to 0.99) for Push and release task, sit to stand*.

Construct validity: Correlation between Push and release test and ABC was high ($r=0.60$) for first trial and ($r=0.55$) for the third trial. Push and release test was more sensitive for identifying fallers but less specific to non-fallers***.

Interpretability: Sensitivity and specificity: The Push and release test was more sensitive to low balance confidence but less specific to high balance confidence***.

Quantitative Posturography (QP)

Browne (2002)¹⁴

Domain: Measure of dynamic balance.

Interpretation: Narrower center-of-pressure indicates lower degree of sway.

Hypothesis testing:

Browne (2002) ¹⁴: Discriminant validity: there was a significant difference in the sway parameters between elderly with history of falls and patients with PD. Concurrent validity: Moderate negative correlation ($r=-0.60$) found between sway parameters of the quantitative posturography and functional reach test*.

Sway on floor and foam <i>Domain:</i> Measure of static and dynamic balance. <i>Interpretation:</i> Wider sway (in mm) indicates poorer balance performance.	<i>Paul (2012)⁶²</i>	<i>Test-retest reliability:</i> ICC: Sway on floor with eyes open (mm): 0.04, Sway on floor with eyes closed (mm): 0.29, Sway on foam with eyes open (mm): 0.42 and Sway on foam with eyes closed (mm): 0.51**. <i>Measurement error:</i> SEM: Sway on floor with eyes open (mm): 172, Sway on floor with eyes closed (mm): 136, Sway on foam with eyes open (mm): 108 and Sway on foam with eyes closed (mm): 123**.
--	---------------------------------	---

Measures of balance and falls risk prediction (n=4)	References	Key finding of psychometric analysis
Modified Clinical test for the Sensory Integration of Balance (mCTSIB) <i>Domain:</i> To assess the usage of sensory inputs (vision, vestibular and somatosensory) when one or more sensory mechanism is affected. <i>Scoring:</i> Maximum score=120 seconds. <i>Interpretation:</i> Longer time is better.	<i>Freeman (2018)³²</i>	<i>Hypothesis testing:</i> Concurrent validity: SOT scores significantly related with the mCTSIB for all sway directions and almost all condition. Correlations ranged from fair ($r=-0.43$) for C to good ($r=-0.64$) for composite. Significant inverse correlation was observed between i-mCTSIB and SOT composite ($r=-0.64$). <i>Hypothesis testing:</i> Discriminant validity: Score for faller and non-fallers significantly differed across both the scales*.

Sensory Organization Test (SOT)

Domain: It measures the ability to use the visual, vestibular and proprioceptive inputs to maintain postural stability in stance.

Interpretation: Wider center of gravity sway signifies lower postural stability. Also, excessive movement of the ankles and hips used to maintain balance indicates lower postural stability.

Freeman (2018)³²

Harro (2016)³⁵

Joanna DiFrancisco (2016)⁴¹

Ozinga (2017)⁵⁹

Test-retest reliability:

Harro (2016)³⁵:

ICC: SOT composite scale: 0.90, SOT vestibular ratio: 0.80 and SOT no of falls: 0.78**.

Measurement error:

SOT composite scale: 4.2, SOT vestibular ratio: 0.1 and SOT no of falls: 1**.

Hypothesis testing:

Concurrent validity: SOT scores significantly related with the mCTSIB for all sway directions and almost all condition. Correlations ranged from fair ($r=-0.43$) for C to good ($r=-0.64$) for composite. Significant inverse correlation was observed between i-mCTSIB and SOT composite ($r=-0.64$)*.

Discriminant validity: Score for faller and non-fallers significantly differed across both the scales*.

Convergent validity: Correlation between balance measures and SOT composite equilibrium:

r range: 0.44 to 0.51**.

Interpretability:

Discriminant validity: Significant differences in scores between fallers and non-fallers in SOT, cut-off score <67 (Sensitivity=0.63 Specificity=0.81)**.

Balance Evaluation System Test (BESTest) <i>Domain:</i> Assesses balance constraints with biomechanical, stability limits, sensory orientation, postural responses, anticipatory postural adjustment and dynamic balance during walking and cognitive effect factors. <i>Scoring:</i> 0 to 108 points. <i>Interpretation:</i> Higher scores indicating better balance.	<i>Duncan (2015a)</i>²⁷ <i>Leddy (2011)</i>⁴⁸ <i>Leddy (2011a)</i>⁴⁹ <i>Maia (2013)</i>⁵³	<i>Inter-rater reliability:</i> ICC: 0.96*. <i>Test-retest reliability:</i> ICC: 0.88*. <i>Criterion validity:</i> Spearman correlation co-efficient: BESTest versus ABC (r= 0.757), BBS (r= 0.87), FGA (r=0.882), UPDRS-ME (r=-0.75). UPDRS (r=-0.78), H&Y (r=-0.73) ***. <i>Structural validity</i> (Rasch analysis): Revealed both BESTest and Mini-BESTest were having stable items and are reproducible**. <i>Hypothesis testing:</i> Predictive validity: Provided greater accuracy and falls prediction than a random guess. BESTest had the best predictive ability with highest AUC***. <i>Responsiveness:</i> Six months change in score of BESTest significantly correlated with the UPDRS-ME, H&Y scale, freezers status and falls history***. 12 months change in score of BESTest significantly correlated with the UPDRS-ME and freezers status***. <i>Interpretability:</i> Discriminant validity (fallers vs non-fallers) Cut-off score, sensitivity and specificity BESTest: ≤69%, 84% and 76%***.
Mini-Balance Evaluation Systems Test (Mini-BESTest) <i>Domain:</i> A short-version of the BESTest assessing dynamic balance. <i>Scoring:</i> 0 to 28 points. <i>Interpretation:</i> Higher scores indicating better dynamic balance.	<i>Benka (2016)</i>¹¹ <i>Bergstrom (2012)</i>¹² <i>Combs (2014)</i>²⁰ <i>Duncan (2011)</i>²⁴ <i>Duncan (2013)</i>²⁸ <i>Duncan (2015a)</i>²⁷ <i>Duncan 2012)</i>²⁵ <i>Joanna DiFrancisco (2016)</i>⁴¹ <i>King (2012)</i>⁴⁵ <i>Leddy (2011)</i>⁴⁸ <i>Lofgren (2014)</i>⁵¹ <i>Maia (2013)</i>⁵³ <i>Schlenstedt (2015)</i>⁷³	<i>Inter-rater reliability:</i> ICC:>0.95*. <i>Test-retest reliability:</i> ICC: >0.95*. <i>Internal consistency:</i> Cronbach's alpha: 0.87*. <i>Criterion validity:</i> The Mini-BESTest and BBS significantly correlated (r=0.79). UPDRS moderately correlated with BBS (-0.39) and Mini-BESTest (-0.51)***. <i>Structural validity:</i> Rasch analysis***. <i>Structural validity:</i> Items 1 (standing up from a seated position) and 7 (standing on a firm surface with eyes open) have ceiling effect. Similarly, outfit mean square errors indicated that items 1 and 7 were redundant. In

Schlenstedt (2016)⁷⁴

Wallen (2016)⁸²

addition, a PCA on the residuals retrieved from the Rasch analysis disqualified the assumption of unidimensionality, although the eigenvalue of 2.04 was close to the upper threshold of 2***.

Hypothesis testing:

Convergent validity (Pearson's correlation): FWT versus Mini-BESTest (0.55)***.

(Predictive Validity) Mini-BESTest score difference correlated with the UPDRS-ME and age***.

-PD subjects' performance was significantly worse than control

-Significant difference in score was found between mild and moderate PD subjects Mini-BESTest did not discriminate between recurrent and non-recurrent fallers. Convergent*** and divergent***

validity (Spearman correlation). Moderate correlation between Mini-BESTest and TUG ($r=-0.47$).

Poor correlation between UPDRS II and Mini-BESTest ($r=-0.21$).

Concurrent validity (Spearman correlation (b) FAB versus Mini-BESTest (0.87) **.

Concurrent validity (Spearman correlation co-efficient): High to very high correlation between Mini-BESTest and BBS ($r=0.94$). Mini-BESTest versus TUG ($r=-0.81$) and FES ($r=0.26$)*.

Predictive validity: Provided greater accuracy and falls prediction than a random guess***.

Concurrent validity (Spearman correlation co-efficient): High to very high correlation between Mini-BESTest and BBS ($r=0.94$). Mini-BESTest versus TUG ($r=0.81$) and FES ($r=0.26$)*.

Responsiveness:

Duncan (2015a): Mini-BESTest score difference correlated with the UPDRS-ME and age***.

Interpretability:

Discriminant validity (differentiate people with and without postural deficits): Cut-off score, sensitivity and specificity ≥ 21 , 89% and 81% ***.

Accuracy of falls prediction using AUC of the ROC was 0.65 for Mini-BESTest. Cut-off scores, sensitivity and specificity: Mini-BESTest: 19, 0.52 and 0.70***.

Consensus Based Standards for the selection of health status Measurement Instruments (COSMIN) scores: *poor, **fair, ***good and ****excellent, GABS: Gait and Balance Scale, EPE: Endpoint Excursion, ABC: Activities-specific Balance Confidence, ICC: Intra-class Correlation Coefficient, SEM: Standard Error of Measurement, MDC: Minimal Detectable Change, FRT: Functional Reach Test, SOT: Sensory Organization Test, mm: Millimeter, sd: Standard deviation and MCT: Motor control test.

Appendix 4: Outcome measures assessing balance and falls risk predictions at the activity level. COSMIN quality of evidence reported as *. Poor-*, fair-, Good-*** and Excellent-******

Measures of balance (n= 33)	References	Key findings of psychometric analysis
Ambulatory Capacity Measure (ACM) <i>Domain:</i> Measure of burden and Parkinson's disease severity. <i>Scoring:</i> 0-20 points. <i>Interpretation:</i> Lower scores show better ambulatory performance.	<i>Parashos (2015)⁶⁰</i>	<i>Internal consistency:</i> Cronbach's alpha: 0.78*. <i>Construct validity:</i> Significant correlation was found (Spearman $r=0.823$) between the ACM/PIGD and summed-ranks of H&Y, NOF, ABC, FOG, FTSS, TUG, GV and BBS ($p < 0.0001$). ACM/PIGD correlated with FOG, BBS, ABC, TUG and GV: $r=0.72, 0.70, 0.69, 0.67$ and 0.58 respectively***.
Backward Functional Reach Test (BFR) <i>Domain:</i> Measures dynamic stability by assessing the maximum distance reached backwards from a fixed standing position. <i>Interpretation:</i> Longer distance reached indicates better stability.	<i>Brusse (2005)¹⁵</i> <i>Franchignoni (2005)³¹</i> <i>Steffen (2008)⁷⁶</i>	<i>Inter-rater reliability:</i> ICC: 0.87*. <i>Test-retest reliability:</i> ICC: 0.73**. <i>Construct validity:</i> Correlation between performance-based balance and mobility test: BBS and SLST: 0.75***. <i>Hypothesis testing:</i> Concurrent validity (Spearman correlation co-efficient) UPDRS-ME versus BFR ($r= -0.33$), BBS versus BFR ($r=0.51$)*.
Bäckstrand Dahlberg Liljenäs Balance Scale (BDL) <i>Domain:</i> Measure of balance deficits due to neurological disorders. <i>Scoring:</i> 0 to 44 points. <i>Interpretation:</i> Higher scores indicate better balance.	<i>Claesson (2017)¹⁹</i>	<i>Internal consistency:</i> Cronbach's alpha: 0.83*. <i>Concurrent validity:</i> Correlation between the BDL and BBS ($r=0.70$) was high. BDL versus UPDRS-ME ($r=0.28$), TUG ($r=-0.32$), CTUG ($r=-0.36$) were low*.

Bed Mobility Test (BMT) <i>Domain:</i> Evaluates bed mobility performance. <i>Scoring:</i> 1-4 points. <i>Interpretation:</i> Higher scores are better.	<i>Spagnoulo (2018)</i>⁷⁵	<i>Responsiveness:</i> Cut-off point greater than 1.4 seconds on BMT was able to discriminate individuals reporting clinical change (60%) from the ones who did not.**
Bloem et al Protocol <i>Domain:</i> Assesses the reaction to an unexpected shoulder pull. <i>Scoring:</i> 0 to 3 points. <i>Interpretation:</i> Lower scores indicate better postural control.	<i>Visser (2003)</i>⁸¹	<i>Inter-rater reliability:</i> Weighted kappa (κ) was high (>0.80) for most of the ratings*. <i>Hypothesis testing:</i> Concurrent criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.
Choice Stepping Reaction Time (CSRT) <i>Domain:</i> It assesses risk of falling by measuring the ability to respond and step fast and appropriately. <i>Interpretation:</i> Shorter time to complete the task indicates better balance.	<i>Paul (2012)</i>⁶²	<i>Test-retest reliability:</i> ICC: 0.74**.
Cleveland Clinical Mobility and Balance Application (CC-MB) <i>Domain:</i> A mobile device platform that measures TUG performance. <i>Scoring:</i> It detects total trial time, gait, and turning performance on or off medication. <i>Interpretation:</i> Shorter time to	<i>Koop (2018)</i>⁴⁷	<i>Test-retest reliability:</i> ICC: (Total trial time) between the On and Off medication states were 0.96 and 0.90 respectively. Average turn velocity had an ICC of 0.83 (Off state) and 0.85 (On state)*.

complete the task indicates lower risk for falls.

Comfortable 10 Meters walk Test (CMT) *Combs (2014)*²⁰

Domain: Measure of walking speed over a short distance in meter per second.
Interpretation: Higher scores are indicative of better walking performance.

Test-retest reliability:
ICC: 0.98***.

Hypothesis testing:
Convergent validity (Pearson's correlation): FWT versus CMT (0.69). Discriminant validity: scores of CMT and FWT did not significantly differ from stage 1 and 2 of H&Y or stage 3 and 4 of H&Y. MDC: 0.09***.

Coordinated Stability Task (CST) *Paul (2012)*⁶²

Domain: Measure of dynamic balance.
Interpretation: Lower degree of sway (measured using swaymeter) indicates better balance.

Test-retest reliability:
ICC: 0.50**.

Fast 10 meters Walk Test (FWT) *Combs (2014)*²⁰

Domain: Measure of walking speed calculated as the duration to complete 10 metres of walking.
Interpretation: Lower the time required to complete better the performance.

Test-retest reliability:
ICC: 0.99***.

Hypothesis testing:
Convergent validity (Pearson's correlation): FWT versus CMT (0.69), PDQ39-M (-0.42), ABC (0.36), FoF (-0.32), Mini BESTest (0.55), 6MWT (0.75) & H&Y (-0.32). Discriminant validity: scores of CMT and FWT did not significantly differ from stage 1 and 2 of H&Y or stage 3 and 4 of H&Y. MDC: 0.13***.

Fear of Falls Measure (FFM) *Franchignoni (2005)*³⁰
*McKee (2014)*⁵⁴
Domain: Measure of disease severity, physical functioning and balance among people with

Internal consistency:
Cronbach's alpha: 0.95**.
Construct validity:
BBS and FFM: moderate correlation ($r = -0.67$) FFM and Tandem Romberg: -0.36, FFM to TUG:

PD.

Scoring: 0 to 38 points.

Interpretation: Lower the score obtained better the function.

0.58***.

Hypothesis testing:

Concurrent validity (Spearman correlation) FSST versus Fear of falls ($r=0.41$)**.

Fullerton Advanced Balance scale (FAB)

*Schlenstedt (2015)*⁷³

Domain: To assess dynamic and static balance under different sensory conditions.

Scoring: 0 to 40 points.

Interpretation: Higher scores indicate better balance.

Inter-rater reliability:

ICC: 0.99*.

Test-retest reliability:

ICC: 0.99*.

Gait speed

*Brusse (2005)*¹⁵

Domain: Measure of walking and functional capacity.

*McKee (2014)*⁵⁴

Interpretation: Higher scores indicate better functional and walking capacity.

*Paul (2013)*⁶³

Inter-rater reliability:

(ICC) Gait speed: 0.90, fast gait speed: 0.94*.

Hypothesis testing: Concurrent validity (Spearman correlation co-efficient) UPDRS-ME versus Comfortable gait speed ($r=-0.12$) and Fast gait speed ($r=-0.18$). BBS versus comfortable gait speed (0.73) and fast gait speed ($r=0.64$)*.

Hypothesis testing: Concurrent validity (Spearman correlation) FSST versus forward preferred gait speed (-0.46), backward preferred gait speed (-0.57), Fast gait speed ($r=-0.56$)**.

Interpretability: The absolute probability of falling in the next 6 months for people with low, medium, and high risk using the simple, 3-test tool was 17%, 51% and 85%, respectively***.

Instrumented Timed Up and Go (iTUG)

*van Lummel (2016)*⁸⁰

Domain: Measures mobility and risk of falls.

Interpretation: Shorter time to complete the task indicates lower risk for falls.

Inter-rater reliability: (Day 1) iTUG: 0.95, (Day 2) iTUG: 0.96*.

(Day 1) iTUG: 0.95, (Day 2) iTUG: 0.98*.

Test-retest reliability:

(Day 1) iTUG: 0.88, (Day 2) iTUG: 0.89*.

Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease

*Goetz (2008)*³⁴

*Koop (2018)*⁴⁷

Internal consistency:

Conbach's alpha: 0.79-0.93 across parts***.

Criterion-related validity: MDS-UPDRS correlated with the original UPDRS ($\rho=0.96$)*.

Rating Scale (MDS-UPDRS)
Domain: Measures the burden of Parkinson's disease progression over time.
Scoring: 0-4 points.
Interpretation: Lower score are better.

Construct validity: Percentage improvement in MDS-UPDRSIII scores correlated with percentage improvement in the Normalized Jerk scores ($\rho=0.41$, $p=0.0025$)**.

Nutt et al
Protocol
Domain: An estimate of reaction to an unexpected shoulder pull without prior warning.
Scoring: 0 to 3 points.
Interpretation: Lower scores are better.

Visser (2003)⁸¹

Inter-rater reliability
Weighted kappa (κ) was high (>0.80) for most of the ratings*.
Hypothesis testing: Predictive validity: predictive accuracy to discriminate between PD-stable and PD unstable was 0.75 for Nutt rating, Concurrent: criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.

Parkinson's Disease Composite Scale (PDCS)
Domain: Measure of postural stability, motor and non-motor symptoms in PD
Scoring: 0 to 94 points.
Interpretation: Lower scores indicate better function.

Stocchi (2018)⁷⁷

Internal consistency:
Motor: Cronbach's alpha 0.64****.
Structural validity
Five factors explaining 56% of the variance were identified by the Kaise-Meyer-Olkin 0.806; Bartlett test of sphericity, $p<0.001$)****.
Hypothesis testing:
Acceptability: No relevant floor or ceiling effect was observed. Correlation between PDCS and motor-HY ($r=0.73$) and correlation between PDCS and other measures tested ranged between 0.59 and 0.78)****.

Pastor et al
Protocol
Domain: Measure of reaction to an expected shoulder pull (patient instructed not to step backwards).
Scoring: 0 to 4 points.

Visser (2003)⁸¹

Inter-rater reliability:
Weighted kappa (κ): 0.98*.
Hypothesis testing:
Pastor rating has the lowest specificity (0.69). Concurrent criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.

Interpretation: Lower scores indicate better function.

Postural Stability Score (PSS) *Jenkins (2010)⁴⁰*

Domain: Measure of postural instability.

Scoring: 0 to 4 points.

Interpretation: Lower scores indicate higher stability.

Construct validity:

Correlation between FRT and the 'Real-World' maximal reach were: Top reach ($r=0.72$), Middle reach (0.76) and Bottom reach (0.73)*.

Postural Change Scale (PCS) *Franchignoni (2005)³⁰*

Domain: Measure of ability to change postures.

Scoring: 0 to 12 points.

Interpretation: Higher scores indicate better postural response.

Internal consistency:

Cronbach's alpha: 0.82^{**} .

Construct validity: PCS moderately correlated with the BBS ($r=0.82$) and FFM (-0.65). Correlation between performance-based balance and mobility test ranged from 0.51 (Functional reach) to 0.75 (single-limb stance) for BBS, from -0.36 (Tandem Romberg) to 0.58 (TUG) for FFM and from 0.40 (Tandem Romberg) to -0.59 (TUG) for PCS***.

Reach test *Paul (2012)⁶²*

Domain: Measure of balance assessing left and right lateral reach.

Interpretation: Longer the distance reached in "cm" greater the balance.

Test-retest reliability:

ICC: (R) lateral reach (cm): 0.81 , (L) lateral reach (cm): 0.62^{**} .

Short Parkinson Evaluation Scale (SPES) *Visser (2003)⁸¹*

Domain: Measure evaluating reaction to unexpected shoulder pull.

Scoring: 0 to 3 points.

Interpretation: Lower scores indicate better balance.

Inter-rater reliability:

Inter-rater reliability: weighted kappa (κ) was high (>0.80) for most of the ratings*.

Hypothesis testing:

Concurrent criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.

Single-Leg-Stance-Test (SLST) *Chomiak (2015)¹⁷*
Domain: Measure of balance *Jacobs (2006)³⁹*
 assessing the ability to stand on *Paul (2016)⁶⁴*
 single leg with eyes open or
 closed.
Interpretability: Longer the
 time sustained (in seconds)
 better the balance.

Inter-rater reliability:
 Inter-rater reliability of 2 different raters were excellent for all three outcomes
 (ICC 0.77 to 1.00)*.
Test-retest reliability:
 ICC: Reliability of 2D system cross 3 raters in a sub-set of 15 (5 from each group) was excellent for
 Push release, Single leg stance and single leg stance with acutely induced dizziness (ICC 0.99 to
 1.00)*.
Construct validity:
 Except the One-leg stance times the other 3 tests had a significant difference between PD and control
 subjects***.
Hypothesis testing:
 Concurrent validity: SLST significantly correlated with the UPDRS-ME. UPDRS bradykinesia sub
 score is significantly correlated with the SLST performance*.
Hypothesis testing:
 Discriminant Validity Except the One-leg stance times the other 3 tests had a significant difference
 between PD and control subjects***.

Single Leg Stance (SLS) ± *Paul (2016)⁶⁴*
 acutely induced dizziness
 Measure of balance on a single
 leg standing with eyes open or
 closed.
Interpretability: Longer the
 time sustained (in seconds)
 better the balance.

Inter-rater reliability:
 Inter-rater reliability of 2 different raters was excellent for all three outcomes (ICC 0.77 to 1.00)*.
Test-retest reliability:
 ICC: Reliability of 2D system across 3 raters in a sub-set of 15 (5 from each group) was excellent for
 Push release, Single leg stance and single leg stance with acutely induced dizziness (ICC 0.99 to
 1.00)*.

Six-Spot Step test *Brincks (2019)¹³*
 Measure of balance requiring
 the user to walk crisscross
 across 6 spots.
 Interpretation: Lesser the time
 required to complete the task,
 better the balance.

Concurrent validity:
 A strong correlation ($p=0.81$, $P<0.001$) between the Six-Spot Step test and the TUG indicates high
 concurrent validity.***
 Moderate Spearman's Rank Correlation Co-efficient (-0.64 , $P<0.001$) between the Six-Spot Step test
 and the mini-BESTest describes moderate validity.***

Timed Up and Go Assessment *de Silva (2017)²¹*

Intra-rater reliability:

of Biomedical Strategies (TUG-ABS)

Domain: Measure assessing the biomechanical strategies while performing the TUG test.

Scoring: 15 to 45 points.

Interpretation: Higher scores indicate better performance.

ICC: 0.99***.

Test-retest reliability:

ICC: 0.96***.

Internal consistency:

Cronbach's alpha: 0.98**.

Construct validity:

TUG-ABS versus TUG ($p=-0.78$), UPDRS-ME ($p=-0.62$) and BESTest section IV ($p=0.72$).

Correlations were moderate to high. TUG-ABS scores significantly differed across the participants of different stages of H&Y scale***.

Hypothesis testing:

Ceiling effect: 22% Inter examiner agreement: 6% and 10.2% variation between 2 evaluations.

Discriminant validity: TUG-ABS showed a canonical correlation of 0.74, with 60% of individuals correctly classified into quick, moderate and slow performers***.

Tinetti Mobility Test (TMT) *Kegelmeyer (2007)⁴⁴*

Domain: Measure of balance and gait. *Park (2018)⁶¹*

Scoring: 0 to 28 points.

Interpretation: Higher scores indicates better balance.

Inter-rater reliability:

Correlation between the scores of student and therapist rating was excellent ($r=0.82$)**.

Inter-rater reliability:

Tinetti balance scale: 0.97, Tinetti gait scale: 0.94*.

Intra-rater reliability:

ICC of 6 raters on day 1 and 1 week later was moderate to high ($r=0.69-0.88$)*.

Unified Parkinson's Disease Rating Scale- Motor

Examination (UPDRS-ME)

Domain: Measure of PD severity relating to balance, gait and posture.

Scoring: 0 to 108 points.

Interpretation: Lower scores indicate lesser severity.

Almeida (2017)⁶

Brusse (2005)¹⁵

Claesson (2017)¹⁹

Duncan (2015a)²⁷

Schlenstedt (2015)⁷³

Stocchi (2018)⁷⁷

de Silva (2017)²¹

Duncan (2013)²⁸

Qutubuddin (2005)⁶⁸

Yang (2014)⁸³

Richard (1994)⁶⁹

Steffen (2008)⁷⁶

Inter-rater reliability:

Total motor score had 0.82 ICC*.

Test-retest reliability:

ICC: 0.89**.

Internal consistency:

Cronbach's alpha: 0.87**.

Criterion validity:

Significant correlation between FSST and UPDRS-ME ($r=0.61$).

Criterion validity:

(Pearson correlation co-efficient) BBS versus UPDRS-ME (-0.58), Modified H&Y (-0.45), Modified ADL (0.55) *.

Construct validity:

The FGA significantly correlated with BBS, FAC, TUG, ABC, UPRSD-ME, BI, fast waling speed and modified H&Y scale (Correlation co-efficient range: 0.57 to 0.85)***.

Unified Parkinson's Disease Rating Scale (UPDRS) *Visser (2003)*⁸¹
Domain: Measure of PD severity.
Scoring: 0 to 199 points.
Interpretation: Lower scores indicate lesser severity.

Unified Parkinson's Disease Rating Scale (UPDRS) items 27-29 *Jacobs (2006)*³⁸
Domain: Measure of PD severity relating balance and posture.
Scoring: 0 to 12 points.
Interpretation: Lower scores indicate lesser severity.

Construct validity:

TUG-ABS versus UPDRS-ME ($p=-0.62$)***.

Hypothesis testing:

Concurrent validity (Spearman correlation co-efficient) UPDRS-ME versus BBS ($r=-0.48$), FFR ($r=-0.46$), BFR ($r=-0.33$), TUG ($r=0.34$), Comfortable gait speed ($r=-0.12$) and Fast gait speed ($r=-0.18$)*.

Predictive Validity: Six months change in score of BESTest significantly correlated with the UPDRS-ME, H&Y, freezers status and falls history. 12 months change in score of BESTest significantly correlated with the UPDRS-ME and freezers status***.

Correlation between PDCS and motor-H&Y ($r=0.73$) and correlation between PDCS and other measures tested ranged between 0.59 and 0.78)****.

Responsiveness

Six months change in score of BESTest significantly correlated with the UPDRS-ME, H&Y, freezers status and falls history. 12 months change in score of BESTest significantly correlated with the UPDRS-ME and freezers status***.

Interpretability:

AUC of 0.75 (95% CI 0.69–0.80) in UPDRS-ME was reported***.

Inter-rater reliability:

Kappa: 0.63*.

Hypothesis testing:

Concurrent criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.

Construct validity:

Except the One-leg stance times the other 3 tests had a significant difference PD and control subjects***.

Hypothesis testing:

Discriminant Validity: Except the One-leg stance times the other 3 tests had a significant difference between PD and control subjects***.

Visual Analog Scale (VAS)
Domain: Measure of estimating the difficulties in daily life due to problems with balance.
Scoring: 10 cm VAS.
Interpretation: Lower scores indicate lesser interference of daily life activities due to problems with balance.

*Schlenstedt (2015)*⁷³

Hypothesis testing:
 Concurrent validity (Spearman correlation (b) FAB versus VAS (-0.49)**.

30 seconds sit-to-stand test (30sSTS)
Domain: Measure of functional lower limb endurance.
Interpretation: Higher number of sit-to-stand indicates better endurance.

*McKee (2014)*⁵⁴
*Peteresen (2017)*⁶⁶

Test-retest reliability:
 ICC: 0.94*.

Hypothesis testing:
 Concurrent validity (Spearman correlation) FSST versus 30 second chair stand (0.339) **.

360° turn
Domain: Measure of balance in older adults.
Interpretation: Lesser duration for completing the task indicate better balance.

*Schenkman (1997)*⁷²
*McKee (2014)*⁵⁴

Test-retest reliability:
 ICC ranged between 0.77 and 0.95 across the performance measures*.
Hypothesis testing:
 Concurrent validity (Spearman correlation) FSST versus 360 turn time (r=0.49)**.

Measures of falls risk prediction (n=5)	References	Key findings of psychometric analysis
Falls Efficacy Scale (FES) <i>Domain:</i> Assess perceived efficiency to avoid fall. <i>Scoring:</i> 0 to 130 points. <i>Interpretation:</i> Higher scores indicate better confidence in carrying out activities.	<i>Nilsson (2010)</i> ⁵⁷	<i>Test-retest reliability:</i> ICC: 0.87***. <i>Internal consistency:</i> Cronbach's alpha: 0.97/0.99**. <i>Measurement error:</i> SEM: 12.3***. <i>Construct validity:</i>

Falls Efficacy Scale- International (FES-I) <i>Domain:</i> Measure of balance confidence and falls risk. <i>Scoring:</i> 16 to 64 points. <i>Interpretation:</i> Lower scores indicate higher balance confidence and lower risk of falls.	<i>Almeida (2016)</i>⁵ <i>Almeida (2017)</i>⁶ <i>Jonasson (2014)</i>⁴² <i>Jonasson (2017)</i>⁴³ <i>Mehdizadeh (2019)</i>⁵⁵	FES strongly correlated with SAFFE ($r=-0.74$). Followed by strong correlation with Physical functioning a sub-scale of SF-36 ($r=0.66$), fast gait speed ($r=0.63$), TUG ($r=-0.61$), UPDRS-II ($r=-0.58$) and UPDRS-ME ($r=-0.46$) ***. <i>Hypothesis testing:</i> Floor and ceiling effect were below or close to the recommendation of <15-20%***. <i>Test-retest reliability:</i> Original FES-I: 0.92, Short FES-I: 0.91***. <i>Test-retest reliability:</i> ICC: total FES-I for On- and Off-drug phases were 0.94 and 0.91 respectively*. <i>Internal consistency:</i> Cronbach's alpha: 0.96 for On- & 0.98 for Off-drug phase**. <i>Measurement error:</i> SEM (% of possible scoring range) FES-I: 3.4 (7) ***. <i>Hypothesis testing:</i> Convergent validity: Correlation between original and short for FES-I was 0.97 Discriminant validity: Both versions of scale were able to discriminate between people who are afraid of falls, avoided activities, experienced falls and Ceiling and floor effect: was <20% for both the scales***. <i>Hypothesis testing:</i> Discriminant validity: FES-I scores for Fallers vs Non-fallers in On and in Off, Fallers vs Non-fallers were statistically significant ($p < 0.001$ & $p \leq 0.005$, respectively). Effect size for the FES-I in discriminating between Fallers and Non-fallers was 0.68 (On drug phase) and 0.48 (Off-drug phase)***. <i>Interpretability:</i> (Discriminant Validity) Cut-off score (non-recurrent fallers vs recurrent fallers) FES-I: >29 points, 77% and 62%****.
Fall-related Activity Classification (FRAC) <i>Domain:</i> Robust classification to identify falls for people with PD. <i>Scoring:</i> categorized falls- related activities as advanced,	<i>Ross (2017)</i>⁷⁰	<i>Inter-rater reliability:</i> (Kappa): Excellent agreement was reached $k=0.81$.*.

combined and transitional tasks
 Interpretation: transitional
 category accepts least
 challenges while advanced
 category accepts maximal
 challenges with low risk for
 falls

Functional Gait Assessment
 (FGA)
Domain: Measure of postural
 stability during walking.
Scoring: 0 to 30 points.
Interpretation: Higher scores
 indicate better stability.

Duncan (2012)²⁵
Leddy (2011a)⁴⁹

Inter-rater reliability:
 FGA: 0.93*.
Test-retest reliability:
 ICC: 0.91*.
Criterion validity:
 Spearman correlation co-efficient BESTest versus FGA ($r=0.882$)***.
Hypothesis testing:
 Predictive validity: Provided greater accuracy and falls prediction than a random guess. FGA was
 inferior than all other tests in falls reduction***.
Interpretability:
 Discriminant validity (fallers vs non-fallers) Cut-off score, sensitivity and specificity FGA: $\leq 15/30$,
 72% and 78%***.

Three steps falls prediction
 model
Domain: Measure of falls
 prediction and risk of falling in
 people with Parkinson's
 disease.
Scoring: 0-11 points.
Interpretation: Lower scores
 are better.

Lindholm (2016)⁵⁰

Interpretability:
 3-step model yielded an AUC (95 % CI) of 0.74 (0.65–0.84) to distinguish between individuals with
 and without future falls. Sensitivity of 0.57 and specificity of 0.86**.

Measures of balance and falls risk prediction (n=12)	References	Key findings of psychometric analysis
Activities-specific Balance	<i>Almeida (2016)⁵</i>	<i>Test-retest reliability:</i>

Confidence (ABC):	<i>Almeida (2017)</i> ⁶	ICC >0.80***.
Domain: Self-reported measure that assesses the confidence in carrying out a list of 16 activities without losing balance.	<i>Bello-Hasss (2010)</i> ¹⁰	Internal consistency:
	<i>Combs (2014)</i> ²⁰	Cronbach's alpha: - 0.95****.
	<i>Duncan (2011)</i> ²⁴	Measurement error:
	<i>Jacobs (2006a)</i> ³⁹	SEM (% of possible scoring range) ABC: 11.0 (11) ***.
	<i>Jonasson (2014)</i> ⁴²	Criterion validity:
Scoring: 0 to 100%.	<i>Franchignoni (2014)</i> ³¹	Correlation between FTSTS and ABC (r= -0.54, p<0.001)***.
Interpretation: Higher scores indicating better balance.	<i>Leddy (2011a)</i> ⁴⁹	Criterion validity:
	<i>Peretz (2006)</i> ⁶⁵	Spearman correlation co-efficient BESTest versus ABC (r= 0.757)***.
	<i>Peretz (2006)</i> ⁶⁵	Construct validity:
	<i>Steffen (2008)</i> ⁷⁶	Correlation between Push and release test and ABC was high (r=0.60) for first trial and (r=0.55) for the third trial. Push and release test was more sensitive for identifying fallers but less specific to non-fallers***.
	<i>Yang (2014)</i> ⁸³	Construct validity:
		The FGA significantly correlated with BBS, FAC, TUG, ABC, UPRSD-ME, BI, fast waling speed and modified H & Y scale (Correlation co-efficient range: 0.57 to 0.85)***.
		Hypothesis testing:
		Discriminant validity: the mean scores of ABC significantly differed across all stages of H&Y disease severity*. MDC: 11.12*
		Hypothesis testing:
		No floor or ceiling effect was found for the ABC while short versions patients scored bottom level****. Rasch analysis: Items showing fit and misfit were identified.
		Hypothesis testing: Convergent validity: correlation between ABC and short version of ABC (r>0.94) was excellent, ABC and short versions against UPRDS-M (r=0.39) and BBS (r=0.46) was moderate and correlation between ABC and short versions of ABC against S&E and H&Y was weak.
		Correlation between ABC and short versions of ABC against FFM was high (r=0.82) ****.
		Hypothesis testing: Jonasson (2014) ⁴² : Convergent validity: The correlation between the four scales was strong ranging from 0.83 and 0.90. Ceiling and floor effects: was <20% for all 4 scales***.
		Interpretability:
		(Discriminant Validity) Cut-off score (non-recurrent fallers vs recurrent fallers) ABC: ≤55%, 71% and 62%****.
		Sensitivity and specificity: Both long and short-versions of ABC had good sensitivity and specificity in discriminating between the groups*.

Berg Balance Scale (BBS) <i>Domain:</i> Measure of static and dynamic balance. <i>Scoring:</i> 0 to 56 points. <i>Interpretation:</i> Higher scores indicate better balance.	<i>Almeida (2017)</i>⁶ <i>Babaei-Ghazani (2017)</i>⁷ <i>Baggio (2013)</i>⁸ <i>Claesson (2017)</i>¹⁹ <i>Dibble (2008)</i>²³ <i>Dibble & Lange (2006)</i>²² <i>Duncan (2012)</i>²⁵ <i>Franchignoni (2005)</i>³¹ <i>King (2012)</i>⁴⁵ <i>Leddy (2011a)</i>⁴⁹ <i>McKee (2014)</i>⁵⁴ <i>Park (2018)</i>⁶¹ <i>Porta (2015)</i>⁶⁷ <i>Scalzo (2009)</i>⁷¹ <i>Schlenstedt (2015)</i>⁷³ <i>Taghizadeh (2018)</i>⁷⁸ <i>Qutubuddin (2005)</i>⁶⁸ <i>Schlenstedt (2016)</i>⁷⁴ <i>Yang (2014)</i>⁸³	<i>Inter-rater reliability:</i> ICC: 0.98***. <i>Test-retest reliability:</i> ICC and SEM: 0.95 and 2.71 respectively ***. <i>Internal consistency:</i> Cronbach's alpha: 0.92***. <i>Criterion validity:</i> Spearman correlation co-efficient BESTest versus BBS ($r=0.87$)***. <i>Criterion validity:</i> The Mini-BESTest and BBS significantly correlated ($r=0.79$). UPDRS moderately correlated with BBS (-0.39) and Mini-BESTest (-0.51)***. <i>Criterion validity</i> (Pearson correlation co-efficient) BBS versus UPDRS-ME (-0.58), Modified H&Y (-0.45), Modified ADL (0.55)*. <i>Construct validity:</i> Significantly different scores between fallers and non-fallers ***. <i>Construct validity:</i> BBS and FFM: moderate correlation ($r=-0.67$) BBS and PCS: moderately correlated ($r=0.82$). Correlation between performance-based balance and mobility test: BBS and FFR: 0.51 BBS and SLST: 0.75***. <i>Construct validity:</i> The FGA significantly correlated with BBS, FAC, TUG, ABC, UPRSD-ME, BI, fast walking speed and modified H&Y scale (Correlation co-efficient range: 0.57 to 0.85)***. <i>Structural validity:</i> Varimax rotation identified 2 factors with a total variance of 68.7%***. <i>Content validity:</i> BBS item 3 is linked to 'Limits of stability', 4 items (BBS 2,6,7, and 12) were linked to sensory orientation and all the remaining items were linked to 'anticipatory postural adjustment***'. Rasch analysis: there were no misfit items. <i>Cross-cultural validity:</i> The BBS is cross-culturally adopted to Persian language***. <i>Hypothesis testing:</i> Concurrent validity (Spearman correlation) FSST versus BBS ($r=-0.65$)**. Convergent validity: Correlation of the total BBS score with FES-I, FRT, PDQ-39 (mobility),
--	--	--

UPDRS-II, and S&E-ADL scores were 0.74, -0.70, -0.66, -0.62, and -0.61 respectively**.

Discriminant validity (Mann-Whitney test): The difference in BBS scores for faller vs. non-faller (mean $\pm$ SD, 45.42 $\pm$ 11.88 vs. 49.86 $\pm$ 12.05) was significant $p < 0.003^{**}$.

Predictive validity: Provided greater accuracy and falls prediction than a random guess***.

Interpretability:

Discriminant validity (differentiate people with and without postural deficits): Cut-off score, sensitivity and specificity BBS: ≥ 52 , 77% and 74%***.

Interpretability:

(Discriminant Validity) Cut-off score (non-recurrent fallers vs recurrent fallers) BBS: ≤ 49 , 74% and 74%****.

Discriminant validity (fallers vs non-fallers) Cut-off score, sensitivity and specificity BBS: $\leq 47/56$, 72% and 75%***.

Discriminant validity (falls prediction) Cut-off, sensitivity and specificity for BBS: 51, 68% and 63%***.

Accuracy of falls prediction using AUC of the ROC was 0.69 for BBS. Cut-off scores, sensitivity and specificity: BBS: 52, 0.64 and 0.67***.

Dynamic Gait Index (DGI)
Domain: Measure assessing the probability of falling in older adults.
Scoring: 0 to 24 points.
Interpretation: Higher scores indicate better level of function.

Almeida (2016)⁵
Dibble & Lange (2006)²²
Dibble (2008)²³
Huang (2011)³⁷

Test-retest reliability:

ICC: 0.84***.

Construct validity:

Significantly different scores between fallers and non-fallers***.

Hypothesis testing:

Convergent validity (Pearson correlation): Association between the mean scores of TUG and DGI was 0.54. MDC (MDC%): 2.9 (13.3%)***.

Interpretability:

Discriminant validity (non-recurrent fallers vs recurrent fallers). Cut-off score for DGI ≤ 19 , (sensitivity=0.73, specificity=0.72)****.

Cut-off score of 22 resulted in a sensitivity of 0.89, a specificity of 0.48, and a negative LR of 0.27**.

Five Times Sit-to-Stand Test (FTSTS)
Domain: Test for lower

Duncan (2011)²⁴
Paul (2012)⁶²
Peteresen (2017)⁶⁶

Inter-rater reliability:

ICC: 0.99*.

Test-retest reliability: ICC: 0.91**

extremity muscles strength, and functional balance.

Interpretation: Shorter time to complete the task indicating better extremity strength and functional balance.

*Spagnoulo (2018)*⁸⁵

Criterion validity:

Correlation between FTSTS and ABC ($r = -0.54$, $p < 0.001$) Mini-BEST ($r = -0.71$, $p < 0.001$) FOGQ ($r = -0.44$, $p < 0.001$)***.

Hypothesis testing:

MDC: 10seconds*.

Responsiveness:

Cut-off score of 2.5 seconds on FTSTS was able to discriminate those who rated some improvement (63.3% of the sample) from the ones who didn't report clinical changes (36.7%)**.

Interpretability:

Discriminant validity (fallers vs non-fallers): Cut-off score, sensitivity and specificity of FTSTS: 16 seconds, 75% and 68%***.

Forward Functional Reach (FFR)

Domain: Measures stability by assessing the maximum distance reached forwards from a fixed standing position
Interpretation: Higher distance reached indicates better stability.

*Brusse (2005)*¹⁵

*Franchignoni (2005)*³⁰

*Jacobs (2006)*³⁸

*Jenkins (2010)*⁴⁰

*Schenkman (1997)*⁷²

*Steffen (2008)*⁷⁶

Inter-rater reliability:

ICC: 0.87*.

Test-retest reliability:

ICC: 0.73**.

Construct validity:

Correlation between performance-based balance and mobility test: ranged BBS and FFR: 0.51***.

Correlation between FRT and the 'Real-World' maximal reach were: Top reach ($r = 0.72$), Middle reach (0.76) and Bottom reach (0.73)*.

Hypothesis testing:

Concurrent validity (Spearman correlation co-efficient) UPDRS-ME versus FFR ($r = -0.46$) BBS versus FFR ($r = 0.50$)*.

Hypothesis testing:: Discriminant Validity Except the One-leg stance times the other 3 tests had a significant difference between PD and control subjects***.

Hypothesis testing:

(Variation by day or week of testing) No variations in the scores were found across 4 different days of assessment.

MDC: 9**.

Four Square Step Test (FSST)

Domain: Test of dynamic balance.

Interpretation: Shorter time

*Duncan (2013)*²⁸

*McKee (2014)*⁵⁴

Inter-rater reliability:

ICC: 0.99***.

Test-retest reliability:

ICC: 0.90 (off phase) and 0.78 (on phase)***.

taken to complete the task
indicates better balance.

Freezing of Gait Questionnaire (FOGQ) *Candan (2019)¹⁶*
Duncan (2011)²⁴
Domain: Measures the impact of gait impairments on functional independence. *Duncan (2013)²⁸*
Scoring: 0 to 24 points. *Giladi (2000)³³*
Interpretation: Lower the score obtained better the function. *Nilsson & Hagell (2009)⁵⁸*
McKee (2014)⁵⁴

Gait and Balance Scale (GABS) *Baggio (2013)⁸*
Domain: Measure of essential components of balance and gait *Thomas (2004)⁷⁹*

Criterion validity:

Significant correlation between FSST and UPDRS-ME ($r=0.61$), Mini-BESTest ($r=-0.65$), FTSS ($r=0.58$), 6MWT ($r=-0.52$), 9HPT ($r=0.65$) and FOGQ ($r=0.44$)***.

Discriminant validity:

No significant difference in FSST between freezers and non-freezers and fallers and on-fallers***.

Hypothesis testing:

Concurrent validity (Spearman correlation) FSST versus TUG ($r=0.73$), TUG cognitive ($r=0.63$), TUG manual ($r=0.556$), BBS ($r=-0.65$), 360 turn time ($r=0.49$), forward preferred gait speed (-0.46), backward preferred gait speed (-0.57), Fast gait speed ($r=-0.56$), 6MWT ($r=0.47$), 30 second chair stand (0.339), Fear of falls ($r=0.41$), freezing gait questionnaire (0.216), UPDRS ($r=0.49$) **.

Inter-rater reliability:

ICC: 0.914^{**} .

Intra-rater reliability:

(measurement 2): ICC: 0.849^{**} .

Internal consistency:

Cronbach's alpha: 0.96^{**} .

Criterion validity:

Correlation between FTSTS and FOGQ ($r=-0.44$, $p<0.001$)***.

Hypothesis testing:

Concurrent validity (Spearman correlation) FSST versus freezing gait questionnaire (0.216)**.

Construct validity:

The Turkish version of FOGQ correlated with modified H&Y scale ($r_s=0.557$), UPDRS-total (0.649), BBS (0.643), FES (0.693), TUG (0.657) and FTSSST (0.579)**.

Cross-cultural validity: Nilsson & Hagell (2009)⁵⁸.*

Cross-cultural validity: Candan (2019)¹⁶.*

Hypothesis testing:

Duncan (2015)²⁶: (External and discriminant validity): Risk category based on Fall in previous 12 months, FOGQ score and self-selected gait speed $<1.1\text{m/sec.}$: Low risk, 0-8: Medium risk and >8 : High risk***.

Intra-rater reliability:

ICC: 0.94^*

Intra-rater reliability:

for patients with PD.
Scoring: 0 to 82 points.
Interpretation: Lower scores indicate better balance.

Good to fair kappa (κ)>0.41**.
Internal consistency:
 Cronbach's Alpha: 0.94 ***.
Structural validity:
 Kaiser's criteria: 7 factors explaining 70.5% variance for the GABS. 3 factors were found for BBS accounting for 77.45% variance***.
Cross-cultural validity
Hypothesis testing:
 Convergent Validity of GABS to:- BBS ($r_s=-0.93$; $p<0.001$), the Pull test ($r_s=0.60$; $p<0.001$), the H&Y ($r_s=0.69$; $p<0.001$), the SE ($r_s=-0.60$; $p<0.001$), and the FES-I ($r_s=0.61$; $p<0.001$).
 Posture testing of GABS had κ 0.83. Concurrent validity: GABS significantly correlated with Balance Master data (posture, postural stability, balance during stance, single limb stance, tandem stance, turning, toe walking and FR (Spearman correlation co-efficient r 0.46 to 1) Significant correlation was found between GABS and GAITRite variables**.
Interpretability:
 Discriminant validity- falls prediction to discriminate between patients with and without falls over last 12 months was found to have a Cut-off of 13 points, sensitivity of 75% and specificity of 60%.

Optoelectronic System Timed Up and Go (Opto-TUG)
Domain: Measure of dynamic balance and risk of falls.
Interpretation: Shorter the duration to complete the task, better the balance.

Kleiner (2018)⁴⁶

Reliability (reproducibility by different systems: ICC for single measurement: 0.996 (TUG-opto vs TUG-IMU); 0.997 (TUG-opto vs TUG-stopwatch) and 0.995 (TUG-IMU vs TUG-stopwatch).
 ICC for average measurement: 0.998 (TUG-optp vs TUG-IMU); 0.998 (TUG-opto vs TUG-stopwatch) and 0.997 (TUG-IMU vs TUG-stopwatch).*

Timed Up and Go (TUG)
Domain: Measure of dynamic balance and falls risk prediction.
Interpretation: Shorter the time required to complete the task better the balance and low the risk of falls.

Almeida (2016)⁵
Almeida (2017)⁶
Bello-Hasss (2010)¹⁰
Brusse (2005)¹⁵
Claesson (2017)¹⁹
de Silva (2017)²¹
Dibble (2008)²³
Dibble & Lange (2006)²²

Inter-rater reliability:
 (Day 1) TUG: 0.96, (Day 2) TUG: 0.95*.
Intra-rater reliability::
 (Day 1) TUG: 0.96, (Day 2) TUG: 0.97*.
Test-retest reliability::
 [ICC and MDC (MDC%)] TUG: 0.80 and 3.5 (29.8%)*
Construct validity:
 The FGA significantly correlated with BBS, FAC, TUG, ABC, UPRSD-ME, BI, fast waling speed

Foreman (2011)²⁹
 Franchignoni (2005)³⁰
 Huang (2011)³⁷
 McKee (2014)⁵⁴
 Nilsson (2010)⁵⁷
 Park (2018)⁶¹
 Paul (2012)⁶²
 Schlenstedt (2015)⁷³
 Steffen (2008)⁷⁶
 van Lummel (2016)⁸⁰
 Yang (2014)⁸³
 Lofgren (2017)⁵²

and modified H & Y scale (Correlation co-efficient range: 0.57 to 0.85)***.
Hypothesis testing:
 Concurrent validity (Spearman correlation co-efficient) Tinetti balance scale versus TUG (r=-0.64).
 Predictive validity: TUG was significantly different between fallers and non-fallers during OFF medication while pull test demonstrated no difference across fallers and non-fallers**.
 Convergent validity: the ABC significantly correlated with TUG.
 (r=-0.44). MDC: 4.85*.
 Convergent validity (Pearson correlation): Association between the mean scores of TUG and DGI was 0.54. MDC (MDC%):3.5 (29.8%)***.
 Cut-off point greater than 2.2 seconds on TUG could discriminate the ones rating improvements (43.3%) from the ones who didn't report clinical changes (56.7%)**.
Interpretability:
 Cut-off score of 7.95 seconds resulted in a sensitivity of 0.93, a specificity of 0.30, and a negative LR of 0.27**.
 (Discriminant validity) Cut-off score (non-recurrent fallers vs recurrent fallers) TUG: >15.2 sec, 63% and 74%****.
 AUC of 0.72 (95% CI 0.66–0.78) in TUG was reported***.

5-items Activities-specific
 Balance Confidence (ABC-5L)
Domain: Short version of the
 ABC assessing the confidence
 in carrying out 5 activities
 without losing balance.
Scoring: 0 to 100%.
Interpretation: Higher scores
 indicating better balance.

Franchignoni (2014)³¹

Internal consistency:
 Cronbach's alpha: 0.88****.

6-item Activities-specific
 Balance Confidence (ABC-6P)
Domain: Short version of the
 ABC assessing the confidence
 in carrying out 6 activities
 without losing balance.
Scoring: 0 to 100%.

Franchignoni (2014)³¹
 Peretz (2006)⁶⁵
 Visser (2003)⁸¹

Internal consistency:
 Cronbach's alpha: 0.89****.
Hypothesis testing:
 Concurrent criterion validity: significant difference between the performance of PD-stable, PD-unstable and control subjects**.
Interpretability:
 Sensitivity and specificity: Both the long and short-versions of the ABC had good sensitivity and

Interpretation: Higher scores indicating better balance.

specificity in discriminating between the groups*.

Consensus Based Standards for the selection of health status Measurement Instruments (COSMIN) scores: *poor, **fair, ***good and ****excellent, FFM: Fear of Falls measure, ABC: Activities-specific Balance Confidence, BI: Barthel Index, Mini-BESTest: Mini-Balance Evaluation Systems Test, H&Y: Hoehn and Yahr, PDQ39-M: Parkinson's Disease Quality of life-39, FoF: Fear of Falling, S&E-ADL: Schwab and England Activities of Daily Living scale, FES-I: Falls Efficacy Scale-International, FOGQ: Freezing of Gait Questionnaire, FAC: Functional Ambulation Category, FES: Falls Efficacy Scale, PIGD: Postural Instability and Gait Difficulty, NOF: No of falls, FOG: Freezing of gait, FTSS: Five times sit-to-stand, TUG: Timed up and go, GV: Gait velocity, BBS: Berg balance scale, UPDRS-ME: Unified Parkinson's Disease Rating Scale- Motor Examination, SLST: Single Leg Stance Test, BDL: Backstrand Dahlberg Liljenas Balance Scale, CTUG: Timed Up and Go cognition, BMT: Bed Mobility Test, CMT: Comfortable 10 Meters walk Test, FWT: Fast 10 meters Walk Test, 6MWT: 6 minute walk test, FSST: Four Square Step Test, iTUG: Instrumented Timed Up and Go Test, UPDRS-MDS-III: Movement Disorder Society Unified Parkinson's Disease Rating Scale items III, FRT: Functional Reach Test, PCS: Postural Change Scale, TUG-ABS: Timed Up and Go Assessment of Biomedical Strategies, TUG: Timed Up and Go, VAS: Visual Analog Scale, FAB: Fullerton Advanced Balance scale, SAFFE: Short Version of Survey of Activities and Fear of Falling in the Elderly, SF-36: 36-item Short-form Health Survey, PDQ 39-M: Parkinson's Disease Questionnaire-39-M, UPDRS-ADL: Unified Parkinson's Disease Rating Scale-Activities of Daily Living, FGA: Functional Gait Assessment, NUDS: Northwestern University Disability Scale, FFR: Forward Functional Reach Test, FFM: Fear of Falls Measure, GABS: Gait and Balance Scale, FES-I: Falls Efficacy Scale-International, ICC: Intra-class Correlation Coefficient, SEM: Standard Error of Measurement, MDC: Minimal Detectable Change, ROC: Receiver Operating Character and AUC: Area Under an ROC Curve.

Appendix 5: Outcome measures assessing balance and falls risk prediction at the participatory level. COSMIN quality of evidence reported as *. Poor-*, fair-, Good-*** and Excellent-******

Measures of falls risk prediction (n=5)	References	Key findings of psychometric analysis
Modified Short version of Survey of Activities and Fear of Falling in the Elderly (mSAFFE) <i>Domain:</i> Measure assessing the extent activity avoidance due to risk of falling. <i>Scoring:</i> 17 to 51 points. <i>Interpretation:</i> Higher scores indicate greater avoidance.	<i>Jonasson (2014)⁴²</i>	<i>Test-retest reliability:</i> ICC>0.80***. <i>Internal consistency:</i> Cronbach's alpha >0.90*. <i>Measurement error:</i> SEM (% of possible scoring range): 3.0 (9)***. <i>Hypothesis testing:</i> Convergent validity: The correlation between the four scales were strong ranging from 0.83 and 0.90. Ceiling and floor effects: was <20% for all 4 scales***.
Rapid assessment of postural instability questionnaire (RAPID) <i>Domain:</i> To assess postural instability. <i>Scoring:</i> 12 items in Activities of daily living sub-section (rated 0 or 1), Fear of falling rated on 1 to 10 points and Number of falls sub-section rated on 3 months falls recall. <i>Interpretation:</i> Higher scores in all sections are worst.	<i>Chong (2012)¹⁸</i>	<i>Interpretability:</i> Cut-off score, sensitivity and specificity: ADL: ≥6, 29% and 74%; FoF: ≥6, 29% and 74%; NoF: ≥3, 48% and 67%**.
Survey of Activates and fear of Falling in the Elderly (SAFFE) <i>Domain:</i> Measure assessing	<i>Nilsson (2010)⁵⁷</i>	<i>Test-retest reliability:</i> ICC: 0.92***. <i>Internal consistency:</i> Cronbach's alpha: 0.95/0.96**.

the extent activity avoidance
due to risk of falling.

Scoring: 17 to 51 points.

Interpretation: Higher scores
connote greater avoidance.

Measurement error:

SEM: 2.4***.

Construct validity:

SAFEe strongly correlated with PF ($r=-0.76$) followed by FES ($r=-0.74$), TUG ($r=0.67$), fast gait speed ($r=-0.64$), UPDRS II ($r=0.52$) and UPDRS-III ($r=0.50$). Floor and ceiling effect were below or close to the recommendation of <15-20%***.

Hypothesis testing:

Floor and ceiling effect were below or close to the recommendation of <15-20%***.

Timed Up and Go Cognition
(CTUG)

Claesson (2017) ¹⁹

Dibble & Lange, (2006)²²

Domain: Assesses the motor-
cognitive interaction while
walking.

Paul (2012) ⁶²

McKee (2014) ⁵⁴

Interpretation: Shorter time
indicates lower falls risk.

Test-retest reliability:

ICC and SEM: 0.55 and 6.7**.

Hypothesis testing:

Concurrent validity: BDL versus, CTUG ($r=-0.36$) were low*.

Concurrent validity (Spearman correlation) FSST versus TUG cognitive ($r=0.63$)**.

Concurrent validity (Spearman correlation) FSST versus D-TUG cognitive ($r=0.63$), D-TUG manual ($r=0.556$)**.

Interpretability:

Cut-off score of 8.5 seconds resulted in a sensitivity of 0.92, a specificity of 0.40, and a negative LR of 0.23**.

Consensus Based Standards for the selection of health status Measurement Instruments (COSMIN) scores: *poor, **fair, ***good and ****excellent, TUG: Timed Up and Go, UPDRS-II: Unified Parkinson's Disease Rating Scale-II, UPDRS-III: Unified Parkinson's Disease Rating Scale-III, FES: Falls Efficacy Scale, CTUG: Timed Up and Go cognition, FSST: Four Square Step Test, ICC: Intra-class Correlation Coefficient, SEM: Standard Error of Measurement.

63. Almeida LR, Valença GT, Negreiros NN, et al. Predictors of recurrent falls in people with Parkinson's disease and proposal for a predictive tool. *J Parkinsons Dis* 2017; 7(2): 313–324.
64. Baggio JA, Curtarelli Mde B, Rodrigues GR, et al. Validation of the Brazilian version of the clinical gait and Balance Scale and comparison with the Berg Balance Scale. *Arq Neuropsiquiatr* 2013; 71(9a): 621–626.
65. Behrman AL, Light KE, Flynn SM, et al. Is the functional reach test useful for identifying falls risk among individuals with Parkinson's disease? *Arch Phys Med Rehabil* 2002; 83(4): 538–542.
66. Brincks J, Callesen J, Johnsen E, et al. A study of the validity of the Six-Spot Step Test in ambulatory people with Parkinson's disease. *Clin Rehabil* 2019; 33: 1206–1213.
67. Browne J, O'Hare N, O'Hare G, et al. Clinical assessment of the quantitative posturography system. *Physiotherapy* 2002; 88(4): 217–223.
68. Candan SA, Çatıker A and Özcan TŞ. Psychometric properties of the Turkish version of the freezing of gait questionnaire for patients with Parkinson's disease. *Neurol Sci Neurophysiology* 2019; 36(1): 44.
69. Chomiak T, Pereira FV and Hu B. The single-leg-stance test in Parkinson's disease. *J Clin Med Res* 2015; 7(3): 182–185.
70. Chong RK, Lee KH, Morgan J, et al. Diagnostic value of the rapid assessment of postural instability in Parkinson's disease (RAPID) questionnaire. *Int J Clin Pract* 2012; 66(7): 718–721.
71. Dibble LE and Lange M. Predicting falls in individuals with Parkinson disease: a reconsideration of clinical balance measures. *J Neurol Phys Ther* 2006; 30(2): 60–67.
72. Giladi N, Shabtai H, Simon E, et al. Construction of freezing of gait questionnaire for patients with Parkinsonism. *Parkinsonism Relat Disord* 2000; 6(3): 165–170.
73. 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.
74. Holmes JD, Jenkins ME, Johnson AM, et al. Validity of the Nintendo Wii® balance board for the assessment of standing balance in Parkinson's disease. *Clin Rehabil* 2013; 27(4): 361–366.
75. Kegelmeyer DA, Kloos AD, Thomas KM, et al. Reliability and validity of the Tinetti mobility test for individuals with Parkinson disease. *Phys Ther* 2007; 87(10): 1369–1378.
76. Kleiner AFR, Pacifici I, Vagnini A, et al. Timed up and go evaluation with wearable devices: validation in Parkinson's disease. *J Bodywork Mov Ther* 2018; 22(2): 390–395.
77. Koop MM, Ozinga SJ, Rosenfeldt AB, et al. Quantifying turning behavior and gait in Parkinson's disease using mobile technology. *IBRO Report* 2018; 5: 10–16.

78. Lindholm B, Nilsson MH, Hansson O, et al. External validation of a 3-step falls prediction model in mild Parkinson's disease. *J Neurol* 2016; 263(12): 2462–2469.
79. Mehdizadeh M, Martinez-Martin P, Habibi S-A, et al. Reliability and validity of Fall Efficacy Scale-International in people with Parkinson's disease during on-and offdrug phases. *Parkinsons Dis* 2019; 2019: 6505232.
80. Nilsson MH and Hagell P. Freezing of gait questionnaire: validity and reliability of the Swedish version. *Acta Neurol Scand* 2009; 120(5): 331–334.
81. Parashos SA, Elm J, Boyd JT, et al. Validation of an ambulatory capacity measure in Parkinson disease: a construct derived from the Unified Parkinson's Disease Rating Scale. *J Parkinsons Dis* 2015; 5(1): 67–73.
82. Paul SS, Canning CG, Sherrington C, et al. Three simple clinical tests to accurately predict falls in people with Parkinson's disease. *Mov Disord* 2013; 28(5): 655–662.
83. Petersen C, Steffen T, Paly E, et al. Reliability and minimal detectable change for sit-to-stand tests and the functional gait assessment for individuals with Parkinson disease. *J Geriatr Phys Ther* 2017; 40(4): 223–226.
84. Ross A, Yarnall AJ, Rochester L, et al. A novel approach to falls classification in Parkinson's disease: development of the Fall-Related Activity Classification (FRAC). *Physiotherapy* 2017; 103(4): 459–464.
85. Taghizadeh G, Martinez-Martin P, Fereshtehnejad S-M, et al. Psychometric properties of the Berg Balance Scale in idiopathic Parkinson's disease in the drug off-phase. *Neurol Sci* 2018; 39(12): 2175–2181.
86. Visser M, Marinus J, Bloem BR, et al. Clinical tests for the evaluation of postural instability in patients with Parkinson's disease. *Arch Phys Med Rehabil* 2003; 84(11): 1669–1674.
87. Yang Y, Wang Y, Zhou Y, et al. Reliability of functional gait assessment in patients with Parkinson disease: interrater and intrarater reliability and internal consistency. *Medicine* 2016; 95(34): e4545.