British Journal of Occupational Therapy

Research Article

The reliability of the graded Wolf Motor Function Test for stroke

Beverley Turtle[1], Alison Porter-Armstrong[2], May Stinson[2]

British Journal of Occupational Therapy
2020, Vol. 83(9) 585–594

DOI: 10.1177/0308022620902697
journals.sagepub.com/home/bjot


Abstract

Introduction: The graded Wolf Motor Function Test assesses upper limb function following stroke. Clinical utility is limited by the requirement to video record for scoring purposes. This study aimed to (a) assess whether video recording is required through examination of inter-rater reliability and agreement; and (b) assess intra-rater reliability and agreement.

Method: A convenience sample of 30 individuals were recruited following stroke. The graded Wolf Motor Function Test was administered within 2 weeks of rehabilitation commencement and at 3 months. Two occupational therapists scored participants through either direct observation or video. Inter- and intra-rater reliability and agreement were examined for item-level and summary scores.

Results: Excellent inter-rater reliability ($n = 28$) was found between scoring through direct observation and by video (intraclass correlation coefficients >0.9), and excellent intra-rater reliability ($n = 21$) was found (intraclass correlation coefficients >0.9) for item-level and summary scores. Low agreement was found between raters at the item level. Adequate agreement was found for total functional ability, with increased measurement error found for total performance time.

Conclusion: The graded Wolf Motor Function Test is a reliable measure of upper limb function. Video recording may not be required by therapists. In view of low agreement, future studies should assess the impact of standardised training.

Keywords

Upper limb, outcome assessment, stroke, reliability, occupational therapy

Received: 23 July 2019; accepted: 8 January 2020

Introduction

Upper limb impairment is common following stroke (Lawrence et al., 2001), with survivors generally experiencing a combination of reduced motor control, reduced coordination and somatosensory deficits (Lang et al., 2013). With links to increased dependence in daily life activities (Lang et al., 2013), improvement in upper limb motor control and function is central to stroke rehabilitation (Pollock et al., 2014).

Choice of outcome measure has been identified as one of the top three research priorities for improving clinical trials (Smith et al., 2014). Currently, various upper limb outcome measures are recommended according to treatment modality (Sivan et al., 2011), sample group or setting (Langhorne et al., 2011), with no consensus demonstrated in the guidelines (Intercollegiate Stroke Working Party, 2016). The use of standardised outcome measures is essential for evidence-based occupational therapy practice and promoted across occupational therapy guidelines (Association of Canadian Occupational Therapy Regulatory Organizations, 2011; College of Occupational Therapists, 2017; Occupational Therapy Australia, 2018).

The Wolf Motor Function Test (WMFT) was developed to measure upper limb motor activity following stroke and traumatic brain injury (Wolf et al., 1989). Demonstrating adequate psychometric properties among people who have had a stroke (Lin et al., 2009; Morris et al., 2001; Wolf et al., 2001), the WMFT has become a widely used and recommended assessment of upper limb activity (Alt Murphy et al., 2015; Santisteban et al., 2016). The WMFT is recommended for individuals with mild to moderate upper limb impairment (Taub et al., 2011) and is most sensitive to those with a higher level of motor function (Thompson-Butel et al., 2014; Wolf et al., 2001), with floor effects found when used in the early stages of stroke (Lin et al., 2009). The graded Wolf Motor Function Test (gWMFT) was developed for accurate assessment of moderate to severe upper limb impairment (Constraint Induced Movement Therapy

[1]Centre for Health and Rehabilitation Technologies, Institute of Nursing and Health Research, Ulster University, Newtownabbey, UK
[2]School of Health Sciences, Institute of Nursing and Health Research, Ulster University, Newtownabbey, UK

Corresponding author:
Alison Porter-Armstrong, School of Health Sciences, Institute of Nursing and Health Research, Ulster University, Room 01F120, Shore Road, Newtownabbey, BT37 0QB.
Email: a.porter@ulster.ac.uk

Research Group, 2002). The WMFT and gWMFT are conducted in real time with performances video recorded to reduce measurement error when scoring this complex assessment (Constraint Induced Movement Therapy Research Group, 2002; Taub et al., 2011).

A systematic review explored the clinical application and psychometric properties of the gWMFT reported in the literature (Turtle et al., 2019). This review found that the gWMFT was a secondary outcome measure in 11 clinical trials, with two versions of the outcome measure reported: the 14-item gWMFT and the more recent 13-item gWMFT. The studies included in the review were predominantly of low quality due to inconsistencies in how the gWMFT was administered and scored, with some authors adapting it to meet study objectives (Bonifer et al., 2005; Iwamuro et al., 2011; Triandafilou and Kamper, 2014).

Reliability of the two versions of the gWMFT has been assessed across two studies. The 14-item gWMFT was assessed by Bonifer et al. (2005), who found a high level of intra-rater reliability for scoring functional ability in 20 individuals more than 12 months post-stroke. Pereira et al. (2015) found a high level of inter-rater reliability for scoring functional ability and performance time using a Brazilian Portuguese version of the 13-item gWMFT in 10 individuals in the chronic stage of stroke. With no further psychometric evaluation of the gWMFT reported, the gWMFT has limited utility in clinical practice and research. For a more detailed review of the application and psychometric properties of the graded Wolf Motor Function Test, see Turtle et al. (2019).

As noted previously, authors of the gWMFT recommend the use of video recording for scoring participants (Constraint Induced Movement Therapy Research Group, 2002). However, this adds to the burden of delivery and may not be appropriate for use in clinical practice, with evidence suggesting video recording the WMFT is not required for accurate scoring (Whitall et al., 2006).

Therefore, the aims of the current study were to investigate inter- and intra-rater reliability and agreement, and internal consistency for the gWMFT in a sub-acute stroke population (within 3 months of stroke onset).

Method

This study is presented based on the published guidelines for reporting reliability and agreement (Kottner et al., 2011). Ethical approval was granted by the Office for Research and Ethics Committees (Ref:14/NI/1149). All participants provided written informed consent.

Participants

Thirty individuals in the sub-acute phase of stroke recruited to an ongoing pilot randomised controlled trial formed the sample (ClinicalTrials.gov: NCT02276729).

Inclusion criteria were: adults aged 18 years or over and recently admitted to an inpatient rehabilitation ward; stroke diagnosis within 3 months with upper limb motor loss and upper limb rehabilitation a key component of treatment; able to understand and follow two-part verbal and written commands in the English language; and able to provide written consent. Exclusion criteria were: having had a previous stroke or gross cognitive impairment.

Raters

Rater one and rater two were research occupational therapists. The therapists were employed solely to collect outcome measures on the trial and had no clinical relationship with the participants. Training for both raters involved reviewing the manual (Constraint Induced Movement Therapy Research Group, 2002) and viewing training videos, the scoring of which was verified by occupational therapists experienced in the clinical administration of the outcome.

Outcome measure

The gWMFT assesses timed performance and quality of movement (Constraint Induced Movement Therapy Research Group, 2002). The gWMFT consists of 13 graded test items (Appendix 1) (Constraint Induced Movement Therapy Research Group, 2002) and takes approximately 40 minutes to administer. Video recording of the gWMFT is recommended to enable retrospective scoring of functional ability. A template can be purchased from the test's authors to standardise placement of the 13 test items.

Video recording

Test items 1 to 8 require placement of the video camera to the side of the template, 3 feet to the side of the participant being tested, allowing the view of their entire torso (Constraint Induced Movement Therapy Research Group, 2002). Test items 9 to 12 require the same placement of the video camera but zoomed in to detail the upper limb and fine finger movements. Test item 13 requires placement of the video camera to the front of the template and 3 feet in front of the participant (Constraint Induced Movement Therapy Research Group, 2002).

Scoring of the gWMFT

Quality of movement is assessed on the gWMFT using a functional ability scale (FAS). This is an eight-point ordinal scale, ranging from zero (not attempted) to seven (normal movement). Items are completed on two levels (A and B), where level A items are of a higher level of difficulty and are scored between four and seven. Level B items are of a lower level of difficulty and are scored between zero and three. Any items not completed are scored zero. For the assessment of performance time,

participants have 30 seconds to complete level A items, and if unable to do so have a second opportunity to complete the task at level B. Sixty seconds are added onto performance time for level B items, with a maximum time of 120 seconds. Table 1 presents the scoring procedure for level A and level B test items.

Procedure

The test was administered and video recorded according to protocol guidelines by one occupational therapist (rater one) (Constraint Induced Movement Therapy Research Group, 2002). To standardise placement of objects and participants, the template was devised from a plexiglass sheet according to protocol instructions and securely affixed to a table top (Appendix 2). The gWMFT was used to assess the participant's affected arm.

Assessments were completed at 2 weeks (T1) and 3 months (T2). The assessments completed at T1 took place in a private room used for research purposes on the hospital site. Assessments completed at T2 generally took place in the participant's own home.

For inter-rater analyses, rater one completed scoring through direct observation and rater two later viewed and scored participant videos for assessments completed at T1.

For intra-rater analyses, rater two scored assessment videos completed at T2 and re-scored one month later.

Internal consistency was assessed using rater two scoring at T1 and T2.

All recorded participant footage was viewed in a private room on hospital premises. Raters were blinded to each other's scoring.

Measurement constructs

Reliability and agreement determine the amount of measurement error in an outcome, and contribute to test validity (Kottner et al., 2011; Streiner et al., 2015). Reliability refers to the amount of variability between rater scores, while agreement assesses the degree to which allocated scores are identical (Kottner et al., 2011; Streiner et al., 2015). Internal consistency is a form of reliability that assesses the degree to which test items are inter-related and therefore indicative of measuring the same construct (Cronbach, 1951).

Data analysis

Descriptive statistics for age, gender and side of hemiparesis were recorded. The mean value was reported for the total FAS score, and the median value was reported for total performance time (Constraint Induced Movement Therapy Research Group, 2002). Score distributions were examined for both time points. Floor and ceiling effects were present if 15% or more of the sample achieved the minimum or maximum scores (McHorney and Tarlov, 1995).

Item-level reliability and agreement were completed to determine if there were any issues with individual items of the gWMFT. Inter-rater reliability for total and item-level functional ability and performance time were assessed using a two-way random consistency intraclass correlation coefficient ($ICC_{2,1}$) (Shrout and Fleiss, 1979). This enables generalisations to be made to other raters within the same population.

Intra-rater reliability for total and item-level functional ability and performance time were assessed using two-way mixed effects, consistency ICC ($ICC_{3,1}$) (Shrout and Fleiss, 1979). Intraclass correlation

Table 1. Scoring procedure for Level A and Level B items of the graded Wolf Motor Function Test.

	Performance time	Functional ability scale	
Level A	Score = actual time taken in seconds (0–30 seconds)	7	Task completed. Normal movement.
		6	Task completed. Reduced precision, consistency.
		5	Task completed. Noted compensatory movements, increased effort and/or time taken to complete.
		4	Task completed. Slight adjustments made by less affected arm, more than two attempts and/or completed very slowly.
Level B	Score = actual time taken in seconds (0–60 seconds) PLUS additional 60 seconds as Level B tariff	3	Task completed. Noted compensatory movements, increased effort and/or time taken to complete.
		2	Task completed. Slight adjustments made by less affected upper limb, more than two attempts and/or completed very slowly.
		1	No functional movement from more affected upper limb.
		0	Unable to complete. No active movement.

Adapted from Constraint Induced Movement Therapy Research Group (2002).

coefficients determine the level of consistency in the ranking of scores (Hallgren, 2012). A reliability score of 0.60 and above was considered acceptable (Cicchetti, 1994).

To examine item-level inter- and intra-rater agreement, proportion of agreement and proportion of agreement ±1 point were completed for functional ability. Standard error of measurement (SEM) (Stratford and Goldsmith, 1997) was completed for item-level performance time. Standard error of measurement was calculated for the total scores of both functional ability and performance time. The SEM portrays the amount of measurement error in scoring; the larger the value, the greater the variability between raters.

Internal consistency of functional ability and performance time were analysed using Cronbach's alpha. Values above 0.70 were considered indicative of test items measuring the same construct and correlating well together (Terwee et al., 2007).

All analyses were completed using SPSS Statistics (Version 24.0. IBM Corporation, Armonk, NY).

Results

A total of 30 participants were recruited (mean days post-stroke [SD], 14.73 [8.36]). Due to medical reasons, loss to follow-up and technical difficulties in viewing recorded videos, two and nine participants were not assessed at T1 and T2 respectively. Consequently, data from 28 participants yielded the analyses for inter-rater analyses (mean age [SD], 71.3 [9.85]; 18 males and 10 females) and data from 21 participants yielded the analyses for intra-rater analyses (mean age [SD], 70.5 [8.7]; 16 males and five females).

Technical difficulties prevented the scoring of one item for participant one and one item for participant two at T2. In order to utilise existing data, summary scores were calculated using the available items. Patient characteristics are presented in Table 2.

Floor and ceiling effects

Ceiling effects were not evident for either assessment session. At T1, floor effects were found for performance time and functional ability by both raters, with 35.7% and 21.4% of the sample achieving the maximum score of 120 seconds and minimum score of zero, respectively (Table 2).

At T2, floor effects were found for performance time, with 33.7% of the sample achieving the maximum score of 120 seconds (Table 2). Floor effects were also found for functional ability at both testing sessions, with 19% of the sample achieving the minimum score of zero (Table 2).

Inter-rater reliability and agreement

High levels of reliability were found between rater one scoring through direct observation and rater two scoring using recorded videos for item-level (Table 3) and total (Table 4) functional ability and performance time, with ICC values above 0.8.

The proportion of agreement for scoring functional ability at the item level ranged from 0.43 to 0.64 and proportion of agreement ±1 ranged from 0.56 to 0.96 (Table 3). Agreement based on SEM values for performance time at the item level ranged from 0.32 to 19.30, with greater differences found for scoring items 1 and 4 through 12 (Table 3). Standard error of measurement values for total scores was 0.33 for functional ability and 6.49 for performance time (Table 4). Larger differences for scoring performance time occurred where there were differences between raters in assigning participant performance to level A or level B tasks.

Intra-rater reliability and agreement

High levels of reliability were found for item-level (Table 3) and total (Table 4) functional ability and performance time, with ICCs above 0.9. The proportion of

Table 2. Participant characteristics and graded Wolf Motor Function Test scores.

	Two weeks (T1) ($n = 28$)		Three months (T2) ($n = 21$)	
Sex				
Male, *n*	18		16	
Female, *n*	10		5	
Age in years, mean (SD)	71.3 (9.6)		70.5 (8.7)	
Side of hemiplegia				
Left, *n*	18		15	
Right, *n*	10		6	
gWMFT FAS	Rater one	Rater two	Session one	Session two
Mean (SD)	3.74 (2.47)	3.16 (2.11)	3.45 (2.28)	3.53 (2.35)
Floor effect, *n* (%)	6 (21.4)	6 (21.4)	4 (19)	4 (19)
Ceiling effect, *n* (%)	0 (0)	0 (0)	0 (0)	0 (0)
gWMFT performance time				
Mean (SD)	51.79 (55.18)	53.94 (54.51)	47.74 (55.74)	46.39 (55.77)
Floor effect, *n* (%)	10 (35.7)	10 (35.7)	7 (33.3)	7 (33.3)
Ceiling effect, *n* (%)	0 (0)	0 (0)	0 (0)	0 (0)

gWMFT: graded Wolf Motor Function Test; FAS: functional ability scale.

Table 3. Item-level reliability and agreement for the graded Wolf Motor Function Test.

	Inter-rater					Intra-rater				
	Reliability $ICC_{(2,1)}$ (95% CI)		Agreement			Reliability $ICC_{(3,1)}$ (95% CI)		Agreement		
	FAS	Time	FAS Po	Po ± 1	Time SEM	FAS	Time	FAS Po	Po ± 1	Time SEM
1 Raise forearm to table (side)	0.884 (0.764–0.944)	0.943 (0.880–0.973)	0.43	0.82	11.11	0.976 (0.940–0.990)	1 (1–1)	0.8	1	0.56
2 Raise forearm from table to box (side)	0.967 (0.930–0.985)	1 (1–1)	0.64	0.93	0.32	0.982 (0.956–0.993)	1 (1–1)	0.81	1	0.54
3 Extend elbow (side)	0.967 (0.931–0.985)	1 (1–1)	0.54	0.96	0.70	0.969 (0.926–0.987)	0.970 (0.929–0.988)	0.71	0.95	9.29
4 Extend elbow against 1 lb weight (side)	0.866 (0.728–0.937)	0.853 (0.704–0.930)	0.44	0.56	19.30	0.948 (0.875–0.978)	0.967 (0.919–0.986)	0.81	0.95	9.25
5 Raise hand to table (front)	0.913 (0.820–0.959)	0.923 (0.841–0.964)	0.43	0.86	13.22	0.926 (0.827–0.969)	0.969 (0.924–0.987)	0.62	0.95	9.27
6 Raise hand to box (front)	0.926 (0.846–0.965)	0.969 (0.934–0.985)	0.57	0.86	9.89	0.988 (0.970–0.995)	1 (1–1)	0.86	1	0.20
7 Reach and retrieve 1 lb weight on table	0.953 (0.902–0.978)	0.919 (0.833–0.962)	0.57	0.86	13.65	0.967 (0.920–0.986)	1 (1–1)	0.57	1	0.10
8 Move foam stick through supination and pronation	0.900 (0.797–0.953)	0.901 (0.797–0.953)	0.43	0.93	17.16	0.984 (0.960–0.993)	1 (1–1)	0.76	1	0.07
9 Grasp and lift washcloth	0.946 (0.888–0.975)	0.939 (0.872–0.971)	0.5	0.89	13.43	0.973 (0.936–0.989)	0.972 (0.933–0.989)	0.67	0.90	9.24
10 Flip light switch	0.932 (0.858–0.968)	0.936 (0.867–0.970)	0.57	0.86	13.34	0.976 (0.941–0.990)	1 (1–1)	0.62	1	0.08
11 Grasp and lift pen	0.902 (0.797–0.954)	0.875 (0.745–0.941)	0.52	0.85	17.98	0.954 (0.890–0.981)	0.969 (0.924–0.987)	0.77	0.95	9.24
12 Grasp and lift cotton balls	0.913 (0.820–0.959)	0.914 (0.823–0.959)	0.57	0.82	15.08	0.953 (0.888–0.981)	0.957 (0.898–0.983)	0.86	0.95	9.25
13 Lift weighted basket (3 lb), place onto raised table (standing)	0.971 (0.939–0.987)	1 (1–1)	0.68	0.93	0.54	0.987 (0.968–0.995)	1 (1–1)	0.8	1	0.08

ICC: intraclass correlation coefficient; FAS: functional ability scale; Po: proportion of observed agreement; Po ± 1: proportion of agreement ± 1 point; SEM: standard error of measurement.

Table 4. Inter- and intra-rater reliability, standard error of measurement and internal consistency of gWMFT.

	Inter-rater reliability $ICC_{2,1}$ (95% CI) ($n = 28$)	Intra-rater reliability $ICC_{3,1}$ (95% CI) ($n = 21$)	SEM		Internal consistency	
			Inter-rater ($n = 28$)	Intra-rater ($n = 21$)	Two weeks ($n = 28$)	Three months ($n = 19$[a])
Functional ability	0.979 (0.955–0.990)	0.993 (0.983–0.997)	0.33	0.19	0.99	0.99
Performance time	0.986 (0.970–0.993)	0.996 (0.990–0.998)	6.49	3.64	0.98	0.98

CI: confidence interval; SEM: standard error of measurement.
[a]Due to technical difficulties one item was not scored for participants one and two, leading to their exclusion as part of the internal consistency analysis.

agreement ranged from 0.57 to 0.86 and proportion of agreement ±1 ranged from 0.90 to 1 for functional ability scores at the item level (Table 3). Agreement based on SEM values for item-level performance time ranged from 0.07 to 9.29, with greater differences found for scoring items 3, 4, 5, 9, 11 and 12 (Table 3). Standard error of measurement values for total scores were 0.19 for functional ability and 3.64 for performance time (Table 4).

Internal consistency

Internal consistency values for functional ability and performance time for both assessment points were above 0.9 (Table 4).

Discussion

This study estimated the psychometric properties of the gWMFT in a cohort of individuals with stroke and compared the results between scoring through direct observation and using video. Excellent inter-rater reliability was found for the FAS and performance time, and adequate agreement was found for scoring functional ability through direct observation and by video. However, unacceptable measurement error was found for scoring performance time. Excellent reliability was also found for intra-rater analyses. This is the first reported study to investigate the reliability and agreement properties of the gWMFT in the sub-acute phase of stroke. With limited psychometric evaluation existing, the ability to compare this study to previous literature is limited.

Substantial floor effects were found for performance time, with a high proportion of scores clustering at the maximum performance time allowed. Floor effects for the FAS were found by both raters at T1, and at both testing sessions at T2. Comparable findings were found for the WMFT when used with lower-functioning participants, with five participants unable to complete any item within 120 seconds (Thompson-Butel et al., 2015). Lin et al. (2009) found floor effects for the WMFT FAS when applied within 14 days of stroke onset. A large proportion of the current sample were unable to attempt all test items. With no recorded item available to score, participants scored 120 seconds and zero on the FAS. The pilot study, from which this sample was derived, did not preclude individuals with more severe upper limb impairment from recruitment procedures, potentially explaining the floor effects found. With participants demonstrating varying degrees of upper limb function, the gWMFT was not able to sensitively measure the range of motor capabilities exhibited.

The high levels of inter-rater reliability found between raters scoring through direct observation and by video indicates that scoring by video may not be a necessary adjunct. This was further substantiated by adequate agreement found between raters for scoring functional ability. While agreement for total FAS scores was adequate, exact agreement was poor across all items. The SEM for performance time highlighted greater discrepancies between raters. Examination of scores at the item level highlighted rater variations in assigning participant performance to level A or level B. Examining agreement at the item level, SEM values greater than 9 seconds were found for 10 items. Whilst the raters underwent training separately, the training content was consistent for both. This comprised reading the manual (Constraint Induced Movement Therapy Research Group, 2002), viewing training videos of an experienced occupational therapist administering the test with stroke survivors, and scoring in real time. This was augmented by a review of the scoring results with an experienced occupational therapist in a training session. In previous studies raters have been required to demonstrate approximate scoring to each other prior to study commencement (Morris et al., 2001; Whitall et al., 2006). This was not required in this study, potentially leading to measurement error and the disagreements demonstrated at the item level. Duff et al. (2015) recognised the issues of variability in ascribing the subjective aspects of the WMFT to patient performance and designed a quality process to ensure rater standardisation.

Excellent intra-rater reliability for total and item-level functional ability and performance time were found, indicating consistent scoring by one rater, over a 1-month interval. Intra-rater SEM values for functional ability displayed minimal variation between scoring sessions, indicating a good level of agreement. Adequate agreement was found for nine test items, with proportion of agreement greater than 0.7. However, similar to inter-rater agreement analyses, there were unacceptable differences in scoring performance time at both the item level and for total scores.

A previous study has reported good agreement between videotaped and observed scoring for the WMFT

based on $ICC_{2,1}$ agreement factor (greater than 0.9) (Whitall et al., 2006). However, the ICC is not a recommended agreement parameter, potentially obscuring the presence of wider variability (Kottner et al., 2011). Whilst differences in scoring modality may have impacted on rater differences in the current study, unacceptable measurement error was found for scoring performance time using video alone. This indicates the presence of additional factors impacting on measurement error. The study authors consider this the result of differences in accurately differentiating between a level A and level B performance by participants.

Although recommended by authors of the gWMFT and the WMFT (Constraint Induced Movement Therapy Research Group, 2002; Taub et al., 2011), the least affected limb was not tested. Scores for the less affected limb may act as a comparison for the more affected limb and help raters discern between FAS ratings accordingly.

Limitations and future research

As part of an ongoing pilot study, the sample size was small, limiting the amount of data available. This study examined participants in the sub-acute phase of stroke, with most experiencing difficulty attempting all test items. Therefore, consideration of reliability and agreement estimates should be applied with caution. Future study could stratify participants according to level of ability and examine use of the gWMFT in chronic stroke. In addition, the grade 5 Wolf Motor Function Test could be used, which was developed for individuals with more severe upper limb impairment (Uswatte et al., 2018).

Due to the discrepancies in rater agreement, provision of a standardised training programme throughout may reduce disagreement across level of item assigned, minimising error, and should be considered in future studies.

Implications for occupational therapy practice

The results of this study have the following implications for occupational therapy practice:

- The gWMFT is a reliable measure for assessing upper limb function post-stroke.
- Different therapists could potentially deliver the gWMFT with stroke survivors and score at different time points, leading to reliable results.
- Given the complexity of the assessment, training would be recommended prior to use, potentially using a fidelity check as developed by Morris et al. (2009) for the WMFT.
- Video recording may not be necessary when scoring the gWMFT, thereby increasing its clinical utility. This would also help to avoid technical errors in video recording and issues with obtaining consent and adhering to General Data Protection Regulations.
- The gWMFT showed floor effects. Therefore, caution should be applied in using the gWMFT with individuals who demonstrate more severe impairments following stroke. The level 5 WMFT could act as a suitable alternative (Uswatte et al., 2018).

Conclusion

The gWMFT demonstrated good levels of inter- and intra-rater reliability and internal consistency. There was acceptable agreement for functional ability, with greater measurement error found for performance time. This study demonstrates the potential use of the gWMFT in a sub-acute stroke population, without the additive strain of scoring individuals by video.

Key findings

- The graded Wolf Motor Function Test can be reliably scored by video and/or by direct observation.
- Inadequate agreement for scoring performance time and individual items indicates future studies should consider the impact of standardised training in the use of the assessment.

What the study has added

The graded Wolf Motor Function Test is a reliable measure of upper limb function in sub-acute stroke, and videotaping for scoring purposes may not be required.

Acknowledgements

We wish to thank all participants who took part in the study. We also wish to thank Nicola Gallagher for the acquisition of data and NICERN for supporting this project. Special thanks to Patricia McIlwaine, Lourene Abbi and Fiona Morrow, occupational therapists at Whiteabbey Hospital, for their assistance with training on outcome administration and scoring, and for their continued support. We would also like to thank Dr Ian Bradbury for his advice and feedback on statistical analyses.

Research ethics

Ethical approval was obtained from the Office for Research Ethics Committees Northern Ireland in 2015 (Ref: 14/NI/1149).

Consent

All participants provided written informed consent to participate in the study.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This work was completed as part of a PhD studentship, funded by the Department for the Economy (DfE), Northern Ireland. This study was also supported in part by the United Kingdom Occupational Therapy Research Foundation Research Priority Grant 2014, with additional funding from the Health and Social Care Research Fund, Northern Health and Social Care Trust, Northern Ireland.

Contributorship

All authors contributed to the study's conception and design. Alison Porter-Armstrong and May Stinson applied for ethical approval. Beverley Turtle carried out data collection and statistical analysis, and prepared the first draft of this manuscript. All authors were involved in the interpretation of the data, reviewed and edited the manuscript, and approved the final version.

ORCID iDs

Beverley Turtle https://orcid.org/0000-0001-5990-5216
Alison Porter-Armstrong https://orcid.org/0000-0002-3186-9599

References

Association of Canadian Occupational Therapy Regulatory Organizations (2011) *Essential Competencies of Practice for Occupational Therapists in Canada*, 3rd ed. Toronto: Association of Canadian Occupational Therapy Regulatory Organizations.

Bonifer NM, Anderson KM and Arciniegas DB (2005) Constraint-induced movement therapy after stroke: Efficacy for patients with minimal upper-extremity motor ability. *Archives of Physical Medicine and Rehabilitation* 86(9): 1867–1873.

Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4): 284–290.

College of Occupational Therapists (2017) *Professional Standards for Occupational Therapy Practice*. London: College of Occupational Therapists.

Constraint Induced Movement Therapy Research Group (2002) *Manual: Graded Wolf Motor Function Test*. Birmingham: University of Alabama and Birmingham Veteran's Administration Centre.

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3): 297–334.

Duff SV, He J, Nelsen MA, et al. (2015) Interrater reliability of the Wolf Motor Function Test-Functional Ability Scale: Why it matters. *Neurorehabilitation and Neural Repair* 29(5): 436–443.

Hallgren KA (2012) Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8(1): 23–34.

Intercollegiate Stroke Working Party (2016) *National Clinical Guideline for Stroke*, 5th ed. London: Royal College of Physicians.

Iwamuro BT, Fischer HC and Kamper DG (2011) A pilot study to assess use of passive extension bias to facilitate finger movement for repetitive task practice after stroke. *Topics in Stroke Rehabilitation* 18(4): 308–315.

Kottner J, Audigé L, Brorson S, et al. (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology* 64(1): 96–106.

Lang CE, Bland MD, Bailey RR, et al. (2013) Assessment of upper extremity impairment, function, and activity after stroke: Foundations for clinical decision making. *Journal of Hand Therapy* 26(2): 104–115.

Langhorne P, Bernhardt J and Kwakkel G (2011) Stroke rehabilitation. *The Lancet* 377(9778): 1693–1702.

Lawrence ES, Coshall C, Dundas R, et al. (2001) Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke* 32(6): 1279–1284.

Lin JH, Hsu MJ, Sheu CF, et al. (2009) Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. *Physical Therapy* 89(8): 840–850.

McHorney CA and Tarlov AR (1995) Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research* 4(4): 293–307.

Morris DM, Taub E, Macrina DM, et al. (2009) A method for standardising procedures in rehabilitation: Use in the extremity constraint induced therapy evaluation multisite randomised controlled trial. *Archives of Physical Medicine and Rehabilitation* 90(4): 663–668.

Morris DM, Uswatte G, Crago JE, et al. (2001) The reliability of the Wolf Motor Function Test for assessing upper extremity function after stroke. *Archives of Physical Medicine and Rehabilitation* 82(6): 750–755.

Alt Murphy M, Resteghini C, Feys P, et al. (2015) An overview of systematic reviews on upper extremity outcome measures after stroke. *BMC Neurology* 15(1): 29.

Occupational Therapy Australia (2018) *Evidence-Based Practice Position Statement*. Victoria: Occupational Therapy Australia.

Pereira ND, Vieira L, Pompeu FP, et al. (2015) Translation, cultural adaptation and reliability of the Brazilian version of the Graded Wolf Motor Function Test in adults with severe hemiparesis. *Fisioterapia em Movimento* 28(4): 667–676.

Pollock A, Farmer SE, Brady MC, et al. (2014) Interventions for improving upper limb function after stroke. *The Cochrane Database of Systematic Reviews* 11(11): CD010820.

Santisteban L, Térémetz M, Bleton JP, et al. (2016) Upper limb outcome measures used in stroke rehabilitation studies: A systematic literature review. *PLoS One* 11(5): e0154792.

Shrout PE and Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2): 420–428.

Sivan M, O'Connor R, Makower S, et al. (2011) Systematic review of outcome measures used in the evaluation of robot-assisted upper limb exercise in stroke. *Journal of Rehabilitation Medicine* 43(3): 181–189.

Smith CT, Hickey H, Clarke M, et al. (2014) The trials methodological research agenda: Results from a priority setting exercise. *Trials* 15(1): 32.

Stratford PW and Goldsmith CH (1997) Use of the standard error as a reliability index of interest: An applied example using elbow flexor strength data. *Physical Therapy* 77(7): 745–750.

Streiner DL, Norman GR and Cairney J (2015) *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press.

Taub E, Morris DM, Crago J, et al. (2011) *Wolf Motor Function Test (WMFT) Manual*. Birmingham: UAB CI Therapy Research Group.

Terwee CB, Bot SD, de Boer MR, et al. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology* 60(1): 34–42.

Thompson-Butel AG, Lin GG, Shiner CT, et al. (2014) Two common tests of dexterity can stratify upper limb motor function after stroke. *Neurorehabilitation and Neural Repair* 28(8): 788–796.

Thompson-Butel AG, Lin G, Shiner CT, et al. (2015) Comparison of three tools to measure improvements in upper-limb function with poststroke therapy. *Neurorehabilitation and Neural Repair* 29(4): 341–348.

Triandafilou KM and Kamper DG (2014) Carryover effects of cyclical stretching of the digits on hand function in stroke survivors. *Archives of Physical Medicine and Rehabilitation* 95(8): 1571–1576.

Turtle B, Porter-Armstrong A and Stinson MD (2019) A systematic literature review of the application and psychometric properties of the graded Wolf Motor Function Test. *British Journal of Occupational Therapy*. Epub ahead of print 9 October 2019. https://doi.org/10.1177/0308022619879074

Uswatte G, Taub E, Bowman MH, et al. (2018) Rehabilitation of stroke patients with plegic hands: Randomised controlled trial of expanded constraint-induced movement therapy. *Restorative Neurology and Neuroscience* 36(2): 225–244.

Whitall J, Savin DN, Harris-Love M, et al. (2006) Psychometric properties of a modified Wolf Motor Function Test for people with mild and moderate upper-extremity hemiparesis. *Archives of Physical Medicine and Rehabilitation* 87(5): 656–660.

Wolf SL, Catlin PA, Ellis M, et al. (2001) Assessing Wolf Motor Function Test as outcome measure for research in patients after stroke. *Stroke* 32(7): 1635–1639.

Wolf S, Lecraw D, Barton L, et al. (1989) Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Experimental Neurology* 104(2): 125–132.

Appendix 1. List of gWMFT test items and graded options[a]

	Task	Graded options
1	Raise forearm to table (side)	*Level A:* No cushion. *Level B:* Addition of 2.5cm cushion on seat.
2	Raise forearm from table to box (side)	*Level A:* Box at shoulder height. *Level B:* Box at half of shoulder height.
3	Extend elbow (side)	*Level A:* Extend hand to 40cm line. *Level B:* Extend hand to 28cm line.
4	Extend elbow against 1 lb weight (side)	*Level A:* Extend weight to 40cm line. *Level B:* Extend weight to 28cm line.
5	Raise hand to table (front)	*Level A:* No cushion. *Level B:* Addition of 2.5cm cushion on seat.
6	Raise hand to box (front)	*Level A:* Box at shoulder height. *Level B:* Box at half of shoulder height.
7	Reach and retrieve 1 lb weight on table	*Level A:* Starting point beyond 40cm line. *Level B:* Starting point beyond 28cm line.
8	Move foam stick through supination and pronation	*Level A:* Participant moves foam stick through supination, touching a box at 5cm, and pronation, touching a box at 2.5cm. *Level B:* Participant moves foam stick through pronation only.
9	Grasp and lift washcloth	*Level A:* Raking grasp is used. *Level B:* Alternate grasp is used.
10	Flip light switch	*Level A:* Lateral pinch grasp is used. *Level B:* Alternate grasp is used.
11	Grasp and lift pen	*Level A:* Tripod grasp is used. *Level B:* Alternate grasp is used.
12	Grasp and lift cotton balls	*Level A:* Tripod grasp is used. *Level B:* Alternate grasp is used.
13	Lift weighted basket (3 lb), place onto raised table (standing)	*Level A:* Raised table at 22cm above desk. *Level B:* Raised desk lowered to rest upon desk.

[a]Adapted from Constraint Induced Movement Therapy Research Group (2002).

Appendix 2. Photographic layout of gWMFT