**Title:** A systematic review of the application and psychometric properties of the graded Wolf Motor Function Test

**Short title:** A systematic review of the graded Wolf Motor Function Test

**Authors:**

1. Beverley Turtle. Centre for Health and Rehabilitation Technologies, Institute of Nursing and Health Research. Ulster University.
2. Alison Porter-Armstrong. School of Health Sciences, Institute of Nursing and Health Research. Ulster University.
3. May Stinson. School of Health Sciences. Institute of Nursing and Health Research. Ulster University.

**Corresponding author:** Dr 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](mailto:a.porter@ulster.ac.uk)

**Abstract**

**Introduction:** Adapted from the Wolf Motor Function Test, the graded Wolf Motor Function Test (gWMFT) is an upper limb activity assessment for use following stroke and brain injury. The aim of this systematic review was to identify and appraise evidence where the gWMFT has been used or has undergone psychometric evaluation.

**Method:** A systematic review of five databases was conducted to identify studies reporting the gWMFT using a keyword search. Intervention and clinical measurement studies were eligible for inclusion. Data quality was assessed using adapted Critical Appraisal Skills Programme questions and the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) Risk of Bias checklist.

**Results:** Twelve studies, of mostly low quality, were included. Studies included one randomised controlled trial, ten pre-and post-studies and one clinical measurement study. All studies involved participants following stroke. Reliability was the only measurement property assessed in two studies, which were of a ‘doubtful’ and ‘poor’ quality.

**Conclusion:** Low quality studies impede the ability of clinicians and researchers to best determine the applicability of the gWMFT to patient groups and research contexts. Further exploration of the psychometric properties of the gWMFT is recommended across stroke populations using rigorous design methods.

**Key words:** stroke, outcome assessment (health care), upper limb, psychometrics.

# Introduction

Measurement of upper limb function following stroke is complex; the full potential of upper limb motor control cannot be captured through the use of one outcome measure alone (Lang et al., 2013). Upper limb dysfunction can involve paresis, abnormal movement patterns and somatosensory deficits occurring in varying degrees and patterns across stroke survivors (Lang et al., 2013). Therefore, outcome measures must, at least in part, reflect the relevant individualised deficits, as well as demonstrate the expected changes resulting from upper limb treatment (Lang et al., 2013).

The use of standardised outcome measures is encouraged as part of occupational therapy assessment, with emphasis on those that best represent an individual’s performance in everyday activities (College of Occupational Therapists, 2017). Using the International Classification of Functioning, Disability and Health (ICF) as a framework to categorise outcome measures (World Health Organisation, 2001), activity-level outcome measures are often viewed as integral to demonstrating meaningful patient outcomes. These are believed to highlight an individual’s ability to complete everyday tasks (Lang et al., 2013), which are aligned with the goals of occupational therapy.

The Wolf Motor Function Test (WMFT) (Taub et al., 2011) and the graded Wolf Motor Function Test (gWMFT) (Constraint-Induced Movement Therapy Research Group, 2002) are activity-level assessments of upper limb function (see Table 1). The WMFT is one of the most frequently reported activity-level outcome measures used in investigations of upper limb interventions following stroke (Bushnell et al., 2015; Santisteban et al., 2016). The WMFT was designed to assess the upper limb function of individuals with hemiplegia following stroke or brain injury (Taub et al., 2011). The WMFT consists of 15 items, where individuals are scored on their speed and quality of performance, and includes two strength tasks (Taub et al., 2011). Items increase in difficulty from assessing joint-specific movements, through to items requiring the performance of functional tasks. The inclusion of functional tasks has led test authors to surmise this assessment may mirror an individual’s functional use of their more affected upper limb in everyday life (Morris et al., 2001).

The WMFT has undergone extensive evaluation of its psychometric properties to support its use in stroke rehabilitation. The WMFT has demonstrated high levels of inter-rater reliability (intraclass correlation coefficients >0.9) and test-retest reliability (Pearson’s product moment correlation >0.9) for functional ability and performance time when used with individuals more than 12 months’ post-stroke (Morris et al., 2001). Adequate responsiveness was found for the WMFT when used with individuals at least six months’ post-stroke (Hsieh et al. 2009). Aspects of validity such as construct, predictive and criterion have been demonstrated with significant corrleations found between the WMFT and commonly used outcome measures, including the Functional Independence Measure, Action Research Arm Test and Fugl-Meyer Assessment (Hsieh et al., 2009; Wolf et al., 2001; Wolf et al., 2005). Validated for use with a chronic stroke population, a multidisciplinary panel recommended the WMFT as an outcome measure for use in intervention studies (Bushnell et al., 2015). In a systematic review of upper limb outcome measures used in stroke research, the WMFT was the second most reported outcome measure across 477 studies and the most commonly reported activity-level outcome measure (Santisteban et al., 2016).

However, the WMFT was designed to capture the upper limb capabilities of those with mild to moderate deficits, and primarily those within a chronic stroke population. In response, the test authors developed the gWMFT (Constraint Induced Movement Therapy Research Group, 2002) to capture the activity of individuals with moderate to severe upper limb impairment and provide more accurate assessment of individuals in the acute or sub-acute stages of stroke.

Table 1

*Description of the gWMFT*

The gWMFT consists of 13 items, and progresses hierarchically, in a similar pattern as the WMFT (Constraint Induced Movement Therapy Research Group, 2002). Each item consists of two levels (level A and level B), meaning the item can be adjusted to an individual’s level of ability. A manual is available for the gWMFT with detailed instructions on test administration and scoring, promoting standardisation (Constraint Induced Movement Therapy Research Group, 2002).

*Scoring the gWMFT*

Individuals are scored based on their quality of movement and speed of performance. Quality of movement is scored using the functional ability scale (FAS). The FAS is an eight-point ordinal scale, with scores ranging from zero, representing no active movement, through to seven, representing normal movement. Test authors recommend videotaping the assessment and scoring the FAS at a later time to improve accuracy (Constraint Induced Movement Therapy Research Group, 2002).

Scores for performance time and the FAS are determined by the level of item completed. For items completed at Level A individuals must complete the item within 30 seconds and can score between four and seven on the FAS. If an individual is unable to complete a test item within 30 seconds, they are then able to attempt the level B version for that item. An extra 60 seconds is added onto the performance time score for items completed at level B. In addition, individuals are only able to score between zero and three on the FAS. A maximum time of 120 seconds is allowed. The test authors recommend the median time and the mean FAS scores to be reported as summary scores for each individual (Constraint Induced Movement Therapy Research Group, 2002).

However, in comparison to the WMFT, there remains limited uptake of the gWMFT and limited evaluation of its psychometric properties. Such studies are required to determine the suitability of the outcome measure across patient groups and upper limb interventions, as an essential component of evidence-based practice. A broad approach to scoping the literature was adopted, as no synthesis existed which reported upon the clinical use of the gWMFT. This was deemed necessary in order to determine where the gWMFT is being used and with whom, to guide where investigations of its psychometric properties are necessitated. A systematic literature review encompassing examination of the gWMFT is warranted to build on the evidence base for upper limb assessment following stroke.

The aim of this systematic literature review was to explore how the gWMFT has been utilised and reported in the literature. The objectives were:

* To identify and evaluate studies where the gWMFT has been used as a primary and/or secondary outcome measure.
* To summarise how the gWMFT has been reported.
* To identify and evaluate evidence for the measurement properties of the gWMFT.

# Method

A systematic literature review was completed by searching the following electronic databases in October 2018: CINAHL (Cumulative Index to Nursing and Allied Health Literature) (1937 to present), Ovid MEDLINE (1966 to present), AMED (1985 to present), PsycINFO (1872 to present) and Pubmed (1947 to present), for the purpose of locating published research regarding the gWMFT. The literature search was developed and completed by the first author with advice from the specialist subject librarian. The search strategy was formulated using the following keywords in combination: “graded wolf motor function test” OR “gwmft”. The search strategy used for Ovid MEDLINE is detailed in the Appendix.

This systematic literature review was completed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009). This approach was used to enable a transparent and structured search of the literature.

## Inclusion and exclusion criteria

Articles were included if they were in the English language and featured the gWMFT as a primary or secondary outcome measure. Due to the limited review scope, and the focus on the use and application of the gWMFT, any studies regardless of patient population, clinical intervention or methodological design were included. Review articles and those where only an abstract was available were excluded.

*Study selection*

Search results were transferred to the Refworks reference management programme and duplicates were removed. Titles and abstracts for all retrieved studies were screened and examined for any reference to the gWMFT by the first author. The reference list for each relevant publication was also searched, which led to the retrieval of one additional article. The full-text format of papers were reviewed where the outcome measures were not reported in the abstract, and for further examination of study criteria by the first author. The studies retrieved were independently checked by the remaining two review authors, and eligibility confirmed. Differences in opinion were resolved through discussion between the three review authors. One of the full text papers retrieved involved the grade 5 Wolf Motor Function Test (Bowman et al., 2006) and the consensus decision was made to exclude this article.

## Data collection and analysis

Data for all included studies were extracted by the first author and recorded using Excel spreadsheets. The data extracted included participant characteristics (age, gender); time post-stroke; study design; intervention applied; and psychometric properties. Aspects of the gWMFT extracted included version reported; and scoring attributes.

The quality of included studies was assessed using an adapted version of the Critical Appraisal Skills Programme (CASP) for cohort studies (Critical Appraisal Skills Programme, 2018), which includes questions used to assess the quality of studies examining outcome measures (Jerosch-Herold, 2005). This combination was chosen due to the varied type of studies found and the focus of the review on outcome measurement.

The quality assessment was as follows:

1. Did the study address a clearly focused issue?
2. Was the sample recruited in an acceptable way?
3. Is the sample size adequate (is there a power calculation)?
4. Is the instrument described and accurately measured to minimise bias?
5. Are the testers trained in test administration?
6. Have the authors identified and taken into account confounding factors?

Where studies examined the psychometric properties of the gWMFT, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias checklist was used (Mokkink et al., 2018). Designed originally for health-related patient-reported outcomes, the COSMIN can be applied to observer-reported outcome measures, to aid selection and reporting (Hernaez 2015). The measurement property evaluated was rated ‘good’, ‘fair’, ‘doubtful’, ‘poor’ or ‘not applicable’. The lowest rating achieved determined the methodological quality for each applicable study.

Each study was assessed on these attributes by the first author, and scores were agreed upon by the other two authors.

**Results**

The search results are summarised in the PRISMA flow diagram in Figure 1 (Moher et al., 2009). Thirty-five articles were identified from database and reference list searching. Following the removal of duplicates and application of the inclusion/exclusion criteria, 12 studies were reviewed.

Figure 1

## Characteristics of included studies

The characteristics of the included papers are summarised in Table 2. Most studies were performed in the United States of America (n=9), two were completed in India, and one was completed in Brazil. Eleven of the studies were clinically-based, exploring the application of an intervention and these were mostly of a pre- and post-test design. All interventions were aimed at improving upper limb function following stroke. One study was a psychometric investigation examining the inter-rater reliability and agreement of the Brazilian Portuguese version of the gWMFT (Pereira et al., 2015).

The sample sizes for included studies were low; nine studies consisted of 20 participants or less (Bonifer et al., 2005; Bonifer and Anderson, 2003; Demirtas-Tatlidede et al., 2015; Fischer et al., 2016; Flinn et al., 2009; Iwamuro et al., 2011; Pereira et al., 2015; Triandafilou et al., 2011, 2014). Participants were analysed separately according to time post-stroke in the study by Triandafilou and Kamper (2014); 12 participants were two to six months’ post-stroke, and 15 participants were more than six months’ post-stroke. The study with the largest sample size was completed by Arya et al. (2012) which consisted of 103 participants.

All studies were completed with individuals following stroke, and more than half of the studies were completed with individuals six months or more post-stroke (Bonifer et al., 2005; Bonifer and Anderson, 2003; Demirtas-Tatlidede et al., 2015; Flinn et al., 2009; Iwamuro et al., 2011; Pereira et al., 2015; Triandafilou et al., 2011). Five of the studies came from the same research team, using a device called the X-Glove to passively stretch the fingers of the more affected hand (Fischer et al., 2016; Iwamuro et al., 2011; Triandafilou et al., 2011, 2014; Triandafilou and Kamper, 2014).

Most studies included additional upper limb outcome measures; these are demonstrated in Table 2. The Fugl-Meyer upper extremity scale was most commonly reported, followed by measures of hand and arm strength.

Table 2

## Quality of included studies

The quality of studies was generally low (Table 3). The highest scoring study was a randomised controlled trial which completed a sample size calculation, an intention to treat analysis, and an examination of between-group differences in baseline characteristics (Arya et al., 2012).

Most pre- and post-test study designs did not account for how participants were recruited, nor report how bias was reduced. Studies with low sample sizes, and potentially insufficient statistical power were used to determine treatment effects. Most studies did not describe the gWMFT adequately and did not report how the gWMFT was administered and scored. The third most common reason for allocating low scores was due to lack of clarity regarding how test authors reduced confounding factors, including whether participants and assessors were blinded.

There were two case reports which clearly described the participant’s health status and the intervention delivered (Bonifer and Anderson, 2003; Flinn et al., 2009).

Table 3

## Quality of studies reporting psychometric evaluation

Two studies assessed reliability of the gWMFT (Bonifer et al., 2005; Pereira et al., 2015), and methodological quality was appraised using the COSMIN Risk of Bias checklist (Table 4). Intra-rater reliability of the 14-item gWMFT was assessed as part of the intervention study by Bonifer et al. (2005) and showed a good level of reliability (Pearson’s product moment correlation, *r* = 0.96). Paired raters scored functional ability; a physiotherapist and an occupational therapist viewed recorded videos of participants completing the gWMFT. This study received a rating of ‘poor’ due to the limited methodological detail reported regarding the time interval between scoring sessions and how a final score was achieved between raters (Table 4). In addtion, the Pearson product moment correlation was used, which is not an advised method of analysis (Mokkink et al., 2018).

An inter-rater reliability study of the Brazilian Portuguese version of the 13-item gWMFT was completed by Pereira et al. (2015). Inter-rater reliability and agreement for functional ability and performance time were reported. An excellent level of inter-rater reliability was found for the FAS (intraclass correlation coefficient = 0.98 [95% confidence interval = 0.92-0.99]) and performance time (intraclass correlation coefficient = 0.99 [95% confidence interval = 0.95-1.00]). An adequate amount of agreement was found for the FAS (limits of agreement = -0.68-0.6). Although not noted by study authors, the limits of agreement for performance time indicated inadequate agreement, with limits of agreement between -0.68 and 16.1 seconds. The mean difference between rater scores was 5.5 seconds.

Two raters independently administered the gWMFT and scored performance time and FAS through direct observation to 10 participants more than six months post-stroke. This study received a rating of ‘doubtful,’ due to lack of clarity in reporting aspects of the reliability analysis (Table 4).

Table 4

## Reporting of the gWMFT

The version of the gWMFT used by the studies is demonstrated in Table 5. Although not reported, Pereira et al. (2015) adapted 13 items from the gWMFT into Brazilian Portuguese, indicating the use of the latest version. Similarly, Iwamuro et al. (2011) did not reference the gWMFT manual in their study. However authors report the gWMFT consisted of 13 items, indicating the use of the 2002 version.

The 14-item version of the gWMFT referenced by Bonifer and Anderson (2003) and Bonifer et al. (2005) had an additional item, ‘drop golf ball or washcloth’ and required participants to stand to complete items 10 to 14. Fischer et al. (2016) cite the study by Bonifer and Anderson (2003), indicating the use of the 14-item version.

In the study by Flinn et al. (2009) it was not clear if a variation of the gWMFT was used. The study authors reported that tasks requiring fine motor control were removed; the number of tasks requiring gross motor function were reduced; and tasks which required pronation and supination were increased. There was insufficient reporting in the studies by Anandabai and Gupta (2013) and Demirtas-Tatlidede et al. (2015) to determine the version used.

Table 5

*Minimum level of function*

Eight of the studies required research participants to have a minimum level of function as part of their eligibility criteria. A significant level of hand impairment was required by Fischer et al. (2016), Iwamuro et al. (2011), Triandafilou and Kamper (2014) and Triandafilou et al. (2014), which was determined by the Chedoke McMaster Stroke Scale for the hand.

It was unclear what criteria were applied in the study by Anandabai and Gupta (2013) and further exploration was not possible. Arya et al. (2012) used the Brunnstrom stages of arm recovery and those at stages two to five were eligible for their study. This represents a wide variation in ability from basic limb synergies present at stage two to the emergence of more complex movement patterns at stage five. The Fugl-Meyer upper extremity scale was used by Demirtas-Tatlidede et al. (2015), participants were required to score ≤16, indicating a severe level of impairment.

*Measurement of the gWMFT*

There were variations in how studies reported the scoring criteria for the gWMFT (Table 5), with many not stating the differences in scoring according to level of item completed (Anandabai and Gupta, 2013; Bonifer et al., 2005; Demirtas-Tatlidede et al., 2015; Fischer et al., 2016; Flinn et al., 2009; Iwamuro et al., 2011). Only studies by Arya et al. (2012), Bonifer and Anderson (2003) and Pereira et al. (2015) detailed how the FAS and performance time were scored according to the level of item completed.

Iwamuro et al. (2011), Triandafilou et al. (2011, 2014) and Triandafilou and Kamper (2014) assessed participants on three tasks of the gWMFT. The three tasks included: lifting a pen, lifting cotton balls and lifting a washcloth. Triandafilou et al. (2014) reported the use of three hand-specific items from the gWMFT, which were not detailed.

Scoring criteria was also adapted. In the study by Triandafilou and Kamper (2014) participants received an additional 60 seconds for not using the appropriate grasp, with a maximum of 120 seconds, and the sum score for the three tasks was reported. Triandafilou et al. (2014) reported each task was completed three times for each assessment period, the averages for each task were then summed and used to summarise each assessment session. Triandafilou et al. (2011) reported that each task was completed in sets of three, with a maximum time of 60 seconds, and did not report what summary score underwent logarithmic transformation.

In the study by Arya et al. (2012) each participant’s performance time score was that of their less affected arm subtracted from the score for their more affected arm. Flinn et al. (2009) used the summation of performance time scores as a summary score.

# Discussion

This systematic literature review has provided an overview of how the gWMFT has been reported in intervention studies and includes the limited appraisal of its psychometric properties. Although the gWMFT was designed for individuals following stroke or brain injury all included studies involved stroke survivors only. Assessing the quality criteria of the studies, most were of a poor or dubious quality due to the inconsistent administration and scoring of the gWMFT. Due to minimal investigation, evaluation of the psychometric properties of the gWMFT was not used to guide the review. The current review identified one inter-rater reliability study of the Brazilian Portuguese version of the gWMFT (Pereira et al., 2015) and an intra-rater reliability study of the gWMFT FAS (Bonifer et al., 2005).

Hierarchies of evidence are used within healthcare to aid the interpretation of research studies of varying designs (Evans, 2003). Within studies which evaluate the effectiveness of interventions, systematic reviews and randomised controlled trials are viewed as delivering the highest quality of evidence and interpreted as trusted contributors to evidence-based practice (Evans, 2003).

There was one randomised controlled trial included in the review which examined the effectiveness of meaningful task-specific training using the gWMFT as a secondary outcome measure (Arya et al., 2012) and scored highly across all quality criteria. In the study by Arya et al. (2012) the gWMFT was explained in detail; the scoring criteria for performance time and descriptors of the ordinal scale used to score functional ability were reported aiding study replication.

The remaining intervention studies included in the review consisted of pre- and post-test designs. Non-randomised controlled trials are viewed as liable to increased bias and as such register lower on the hierarchy of evidence (Higgins et al., 2011). The description of outcome measures included was generally of low quality, with some simply stating the gWMFT was used and no further examination provided (Anandabai and Gupta, 2013; Fischer et al., 2016; Iwamuro et al., 2011). Also, how test authors reduced confounding factors were infrequently reported, with many not reporting whether assessors were blinded.

Case reports are descriptive and provide little insight into the efficacy of a treatment and are rated poorly in the hierarchy of evidence (Evans, 2003). However, the case report by Bonifer and Anderson (2003) provided an in-depth description of the training provided to raters, and of the gWMFT administration and scoring. While this study may not score highly in determining effectiveness, this study scored highly in the current review due to how the gWMFT was described and administered. In contrast, the case study by Flinn et al. (2009) poorly described the gWMFT and inaccurately reported the inter-rater reliability study which was completed by Bonifer et al. (2005).

## Psychometric properties of the gWMFT

Reliability was the only measurement property assessed for the gWMFT. While the results reported a high level of inter- and intra-rater reliability, the methodological quality of these studies were of a low standard.

In the study by Bonifer et al. (2005), two raters scored functional ability using the videotapes of participants completing the gWMFT. Although the level of training provided was detailed, it was not clear how the two raters came to a final agreed score for each participant. This study also scored poorly on the COSMIN checklist for not reporting the time interval between the repeated measurements. Reporting the time interval is integral to determine that a long enough period has elapsed to prevent the raters from remembering their previous scores (Kottner et al., 2011). Finally, this study reported reliability using Pearson’s product moment correlation which is not an advised method of analysis, with no further exploration of differences between the two-time points (Mokkink et al., 2018).

The study by Pereira et al. (2015) investigated the Brazilian Portuguese version of the gWMFT, which limits its applicability to the English language version. This study scored ‘doubtful’ on the COSMIN checklist due to lack of clarity regarding the type of intraclass correlation coefficient completed. How an intraclass correlation coefficient is interpreted relies on the type analysed, with assumptions made regarding the number of raters involved and how raters score participants (Kottner et al., 2011).

Pereira et al. (2015) also assessed agreement. Bland and Altman plots were used to determine the limits of agreement for scoring functional ability and performance time. Approximately 95% of the difference in scores between raters will lie between the limits of agreement (Bland and Altman, 1999). Ideally, this should be close to zero, indicating minimal differences. A large degree of measurement error was found for scoring performance time, illustrated through wide limits of agreement in the Bland and Altman plot.

In contrast, Fritz et al. (2009) assessed the minimal detectable change for performance time on the WMFT and found that change of at least 0.7 seconds indicated an improvement in ability. The wide degree of measurement error demonstrated by Pereira et al. (2015) would make it difficult to discern whether a change in a participant’s score was the result of actual change in recovery or the result of error. Continued validation of the gWMFT is necessary to determine its ability to accurately measure and document change in upper limb function.

## Application of the gWMFT

Across studies, there was heterogeneity in the version of the gWMFT used, how it was applied and scored. Most of these studies did not provide an adequate description of the complexity involved in delivering and scoring the gWMFT. In comparison to the 12 articles suitable for inclusion in this review, a search of the WMFT using Ovid MEDLINE elicited 384 studies. While increased use of the WMFT is to be expected, there remains a wide disparity in uptake between the two outcome measures. Poor reporting in studies of low quality is likely to play a role in whether clinicians or researchers choose to use the gWMFT.

The gWMFT could provide an appropriate alternative to the WMFT for use in the earlier stages of stroke and with those with a greater degree of impairment. However, included studies did not report floor and ceiling effects which would gauge the sensitivity of the gWMFT to measure severe upper limb deficits accurately. Further studies assessing the reliability, validity, and responsiveness of the gWMFT are required to ascertain its appropriateness across interventions and level of impairment.

## Adaptation of the WMFT

Through the preparation of this review, another adaptation of the WMFT was found called the Grade 5 WMFT (Uswatte et al., 2018). This test is comparable to the gWMFT in that each item consists of two levels, with similar scoring criteria. The Grade 5 version consists of 10 items, and the mean log score is reported for performance time. The gWMFT was developed to assess the upper limb motor function of individuals with moderate to severe deficits, while the Grade 5 WMFT was developed to assess the motor function of individuals with severe deficits.

The same research team responsible for the development of the WMFT devised a system for classifying the minimum active range of motion required at each joint of the upper limb, which can determine the appropriate WMFT to use (Uswatte et al., 2018; Uswatte and Taub, 2013). The gWMFT is suitable for individuals classified as grades 3/4. This requires stroke survivors to have a minimum of 45 degrees flexion and extension at the shoulder through to the ability to extend at least two fingers to a maximum of 10 degrees and extend the thumb by at least 10 degrees (Uswatte and Taub, 2013).

None of the studies in this review used this classification system as part of their inclusion criteria, with seven studies using three different standardised outcome measures to determine the minimum level of function required. This level of heterogeneity concerning the level of ability for which the gWMFT is appropriate limits its clinical utility.

## Implications for research and practice

In summary, this systematic literature review has the following implications for occupational therapists and researchers to consider:

This review has highlighted that in addition to the WMFT, gWMFT and grade 5 WMFT, there exist two versions of the gWMFT. As a result, clinicians and researchers must accurately report the version used and scoring criteria applied. In addition, multiple versions can lead to difficulties in determining the most appropriate outcome measure to use with various client groups. This review has reported that floor and ceiling effects for the gWMFT have not been investigated. Future research should focus on the applicability of the gWMFT to stroke survivors at different stages in their recovery to ascertain its clinical utility.

While there are studies which have reported the gWMFT to a high standard, these were in the minority. Therefore, future attention should be given to the development of high-quality research measuring upper limb function in stroke survivors using the gWMFT.

There is a paucity in psychometric analysis of the gWMFT, with reliability only examined in two different versions (13- and 14-item) of the outcome measure. However, these studies have limited applicability due to inadequate reporting and includes the assessment of a Brazilian Portuguese version of the gWMFT. Priority should be given to assessing the measurement properties of the gWMFT using rigorous design methods.

Whilst the manual for the WMFT is freely available online, the manual for the gWMFT is available upon request only from the test authors. Potentially impacting on uptake of the gWMFT.

The gWMFT is a multi-component outcome measure. Aspects of the testing kit can be purchased independently for self-assembly, and a template can be purchased directly from the test authors, with the option for self-assembly. However, this has the potential to lead to variations across testing kits, and errors in formation.

There is complexity in the delivery and scoring of the gWMFT. Training would be required to ensure accurate and consistent delivery of the outcome measure. Developed by a research team based in the United States of America, training internationally may not be available to all. These factors could impact reliability, leading to inconsistent delivery and scoring between therapists. As noted in the study by Pereira et al. (2015) there were prominent disagreements between raters scoring, which could have been mitigated by a standardised training protocol.

To improve reduce measurement error, test authors recommend video recording individuals completing the gWMFT for scoring purposes. This would require additional consent procedures to be in place and ensure compliance with general data protection regulation.

## Study limitations

This review was limited by the small number of studies identified, with wide heterogeneity impacting data synthesis, which limited the comparisons that could be made. Quality appraisal of the studies was completed using an amalgamation of two quality appraisal tools (CASP, 2018; Jerosch-Herold 2006). Although not a standardised tool, this was created to cope with the heterogeneity of the included studies.

## Conclusion

This review has demonstrated that while the gWMFT has limited uptake, researchers are continuing to use this outcome measure with limited evaluation of its measurement properties. To date, there has been an inter-rater reliability and agreement study of performance time and the FAS for the Brazilian Portuguese version, without similar evaluation completed for the English language version. While the intra-rater reliability study was a step in the right direction, this study lacked rigor and only considered the reliability of the FAS. The gWMFT has the potential to extend the applicability of the WMFT and generate meaningful results concerning the recovery of upper limb function in individuals undergoing stroke rehabilitation. However, further investigation is needed to ascertain its application within different stroke populations.

**Key findings**

* Lack of high-quality upper limb intervention studies which reported the gWMFT.
* Limited psychometric evaluation of the gWMFT, with only reliability assessed.

**What the study has added**

There is a need for rigorous assessment of the measurement properties and clinical utility of the gWMFT to determine the clinical and research context where it can be best applied.

**Acknowledgements:** The review team thanks the Ulster University School of Health Sciences librarian who provided advice on the search strategy.

**Research ethics:** Ethical approval was not required for this study.

**Consent:** Informed consent was not relevant as this is a systematic review.

**Declaration of Conflicting Interests:** The authors declared 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 and was funded by the Department for the Economy (DfE), Northern Ireland.

**Contributorship:** Beverley Turtle, Alison Porter-Armstrong and May Stinson contributed to the conceptualisation and design of the study. Beverley Turtle undertook the literature search and development of the spreadsheet information in preparation for data analysis. All authors were involved in data analysis, drafting of the manuscript and in continued revisions of the manuscript.

References

Anandabai J and Gupta M (2013) Effects of bimanual functional practice training versus unimanual functional practice training on functional performance of upper extremity in chronic stroke. Journal of Physiotherapy and Sports Medicine 2(1): 15–30.

Arya KN, Verma R, Garg RK, et al. (2012) Meaningful task-specific training (MTST) for stroke rehabilitation: a randomized controlled trial. Topics in Stroke Rehabilitation 19(3): 193–211.

Bland JM and Altman DG (1999) Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8(2): 135–160.

Bonifer N and Anderson KM (2003) Application of constraint-induced movement therapy for an individual with severe chronic upper-extremity hemiplegia. Physical Therapy 83(4): 384–398.

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.

Bushnell C, Bettger J, Cockroft K, et al. (2015) Chronic stroke outcome measures for motor function intervention trials: expert panel recommendations. Circulation: Cardiovascular Quality and Outcomes 8(6 suppl 3): S163–S169.

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.

Critical Appraisal Skills Programme (2018) CASP cohort study checklist. Available at: https://casp-uk.net/wp-content/uploads/2018/01/CASP-Diagnostic-Checklist-2018.pdf (accessed 11 August 2018).

Demirtas-Tatlidede A, Alonso-Alonso M, Shetty RP, et al. (2015) Long-term effects of contralesional rTMS in severe stroke: Safety, cortical excitability, and relationship with transcallosal motor fibers. NeuroRehabilitation 36: 51–59.

Evans D (2003) Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. Journal of Clinical Nursing 12(1): 77–84.

Fischer HC, Triandafilou KM, Thielbar KO, et al. (2016) Use of a portable assistive glove to facilitate rehabilitation in stroke survivors with severe hand impairment. IEEE Transactions on Neural Systems and Rehabilitation Engineering 24(3): 344–351.

Flinn NA, Smith JL, Tripp CJ, et al. (2009) Effects of robotic‐aided rehabilitation on recovery of upper extremity function in chronic stroke: a single case study. Occupational therapy international 16(3‐4): 232–243.

Fritz SL, Blanton S, Uswatte G, et al. (2009) Minimal detectable change scores for the Wolf Motor Function Test. Neurorehabilitation and Neural Repair 23(7): 662–667.

Hernaez, R. (2015) Reliability and agreement studies: a guide for clinical investigators. Gut64 (7): 1018-1027.

Higgins JPT, Altman DG, Gøtzsche PC, et al. (2011) The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ 343: d5928.

Hsieh Y, Wu C, Lin K, et al. (2009) Responsiveness and validity of three outcome measures of motor function after stroke rehabilitation. Stroke 40(4): 1386–1391.

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.

Jerosch-Herold C (2005) An evidence-based approach to choosing outcome measures: a checklist for the critical appraisal of validity, reliability and responsiveness studies. British Journal of Occupational Therapy 68(8): 347–353.

Kottner J, Audig 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.

Moher D, Liberati A, Tetzlaff J, et al. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339: b2535.

Mokkink LB, de Vet HCW, Prinsen CAC, et al. (2018) COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of Life Research 27(5): 1171–1179.

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.

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.

Santisteban L, Térémetz M, Bleton J-P, et al. (2016) Upper limb outcome measures used in stroke rehabilitation studies: a systematic literature review. PLOS ONE 11(5): e0154792.

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

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.

Triandafilou KM, Ochoa J, Kang X, et al. (2011) Transient impact of prolonged versus repetitive stretch on hand motor control in chronic stroke. Topics in Stroke Rehabilitation 18(4): 316–324.

Triandafilou KM, Ochoa J and Kamper D (2014) Effect of static versus cyclical stretch on hand motor control in subacute stroke. International Journal of Neurorehabilitation 01(03): 1–5.

Uswatte G and Taub E (2013) Constraint-induced movement therapy: a method for harnessing neuroplasticity to treat motor disorders. In: Progress in Brain Research. Elsevier, p. Vol. 207, pp. 379–401.

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.

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, SL, Thompson, PA, Morris, DM, et al. (2005) The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. *Neurorehabilitation and Neural Repair*; *19*(3): 194-205.

World Health Organisation (2001) International Classification of Functioning, Disability and Health: ICF. Geneva: World Health Organisation.

|  |  |  |  |
| --- | --- | --- | --- |
| Items | Wolf Motor Function Test | 13-item graded Wolf Motor Function Test | 14-item graded Wolf Motor Function Test |
| Raise forearm to table (side)ab | 🗸 | 🗸 | 🗸 |
| Raise forearm from table to box (side)ab | 🗸 | 🗸 | 🗸 |
| Extend elbow (side)ab | 🗸 | 🗸 | 🗸 |
| Extend elbow against 1 lb. weight (side)ab | 🗸 | 🗸 | 🗸 |
| Raise hand to table (front)ab | 🗸 | 🗸 | 🗸 |
| Raise hand to box (front)ab | 🗸 | 🗸 | 🗸 |
| Raise weighted hand to box (front)a | 🗸 |  |  |
| Reach and retrieve 1 lb. weight on table (front)ab | 🗸 | 🗸 | 🗸 |
| Raise can to moutha | 🗸 |  |  |
| Grasp and lift pencil from tablea | 🗸 |  |  |
| Lift paperclip from tablea | 🗸 |  |  |
| Stack three checkers on top of one anothera | 🗸 |  |  |
| Turn over three cardsa | 🗸 |  |  |
| Turn key in locka | 🗸 |  |  |
| Measure grip strength using dynamometera | 🗸 |  |  |
| Move foam stick through supination and pronationb |  | 🗸 | 🗸 |
| Grasp and lift washclothb |  | 🗸 | 🗸 |
| Flip light switchb |  | 🗸 | 🗸\* |
| Grasp and lift penb |  | 🗸 | 🗸\* |
| Grasp and lift cotton ballsb |  | 🗸 |  |
| Lift weighted basket (3 lb.), place onto raised table (standing)b |  | 🗸 | 🗸\* |
| Grasp and lift foam trianglesc |  |  | 🗸\* |
| Drop golf ball or washcloth with forearm supportedc |  |  | 🗸\* |

Table 1. Items assessed on the Wolf Motor Function Test, 13-item graded Wolf Motor Function Test and 14-item graded Wolf Motor Function Test

a Adapted from permission from Constraint Induced Movement Therapy Research Group. Taub E, Morris DM, Crago J, et al. (2011) Wolf Motor Function Test (WMFT) Manual. Birmingham: UAB CI Therapy Research Group.

b Adapted with permission from Constraint Induced Movement Therapy Research Group. Constraint Induced Movement Therapy Research Group (2002) Manual: Graded Wolf Motor Function Test. Birmingham: University of Alabama and Birmingham Veteran’s Administration Centre.

c Adapted with permission from Constraint Induced Movement Therapy Research Group. Constraint-Induced Movement Therapy Research Group (2000) cited in Bonifer N and Anderson KM (2003) Application of constraint-induced movement therapy for an individual with severe chronic upper-extremity hemiplegia. Physical Therapy 83(4): 384–398.

\*Task completed in standing.

Table 2. Characteristics of the studies included in the review

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Author (year) | Patient (n) | Country | Time post stroke | Age, years (mean ±SD),  Gender | Type of study | Intervention | Additional upper limb assessments included |
| Anandabai et al. (2013) | 30 | India | 3-4 months | Not reported  26 male, 6 female | Pre- and post-study | Bimanual (n=15) and unimanual (n=15) functional practice | Fugl-Meyer Assessment |
| Arya et al. (2012) | Intervention: 51  Control: 52 | India | 4-24 weeks | Intervention: 51.67 ±7.96  29 male, 22 female  Control: 50.21 ±7.60  33 male, 19 female | RCT | Meaningful task-specific training | Action Research Arm Test  Fugl-Meyer Assessment |
| Bonifer and Anderson (2003) | 1 | USA | 15 years | 53  1 female | Case report | Constraint-induced movement therapy | Fugl-Meyer Assessment  Motor Activity Log |
| Bonifer et al. (2005) | 20 | USA | >12 months | 57.5 ±16.6  13 male, 7 female | Pre- and post-study/Psychometric study | Constraint-induced movement therapy | Fugl-Meyer Assessment  Motor Activity Log |
| Demitas-Tatlidede et al., (2015) | 10 | USA | >1 year | 59.5 ±11  4 male, 6 female | Pre- and post-study. | Contra-lesional repetitive transcranial magnetic stimulation | Fugl-Meyer Assessment  Hand strength assessments  Modified Ashworth Scale |
| Fischer et al. (2016) | 15 | USA | 2-6 months | 63 ±12  10 male, 3 female | Pre- and post-study | Passive cyclical finger stretching with active-assisted task-oriented training using an orthotic glove | Action Research Arm Test  Chedoke Arm and Hand Inventory  Fugl-Meyer AssessmentHand strength assessments  Motor Activity Log |
| Flinn et al. (2009) | 1 | USA | 15 months | 48  1 female | Case report | Robot-assisted therapy | Active range of motion  Fugl-Meyer assessment  Motor Activity Log |
| Iwamuro et al. (2011) | 5 | USA | ≥9 months | 54 ±11  4 male, 1 female | Pilot pre- and post-study | Passive finger extension using an orthotic glove | Active range of motion  Box and Block Test |
| Pereira et al. (2015) | 10 | Brazil | >6 months | 53.2 ±11.39  6 male, 4 female | Psychometric study | N/A | N/A |
| Triandafilou and Kamper(2014) | Sub-acute: 12  Chronic: 15 | USA | Subacute: 2-6 months  Chronic: >7months | Sub-acute: 53 ±6  6 male, 6 female  Chronic: 57 ±8  7 male, 8 female | Pre- and post-study | Passive cyclical finger stretching using an orthotic glove | Box and Block Test  Hand strength assessments |
| Triandafilou et al. (2014) | 13 | USA | 2-6 months post | 51 ±12  6 male, 7 female | Pre- and post-study | Static finger stretching/passive finger stretching/rest using an orthotic glove | Hand strength assessments  Grip termination time |
| Triandafilou et al. (2011) | 15 | USA | ≥6 months post | 57 ±8  7 male, 8 female | Pre- and post-study | Prolonged and repetitive passive finger stretching using an orthotic glove | Hand strength assessments  Grip termination time |

Note: SD, standard deviation; RCT, randomised controlled trial.

Table 3. Critical appraisal scoring for each reviewed article

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Author (year)** | **Address focused issue** | **Acceptable recruitment** | **Adequate sample size** | **Instrument described and accurately measured** | **Testers trained in administration** | **Confounding factors identified and taken into account** |
| Anandabai et al. (2013) | Y | CT | N | N | CT | N |
| Arya et al. (2012) | Y | Y | Y | CT | Y | Y |
| Bonifer and Anderson (2003) | Y | N/A | N/A | Y | Y | N/A |
| Bonifer et al. (2005) | Y | Y | N | Y | Y | Y |
| Demitas-Tatlidede et al. (2015) | Y | CT | N | N | CT | N |
| Fischer et al. (2016) | Y | CT | N | N | CT | CT |
| Flinn et al. (2009) | Y | N/A | N/A | N | Y | N/A |
| Iwamuro et al. (2011) | Y | CT | N | N | N | CT |
| Pereira et al. (2015) | Y | CT | N | Y | CT | CT |
| Triandafilou and Kamper (2014) | Y | CT | N | N | CT | CT |
| Triandafilou et al. (2014) | Y | CT | N | N | CT | CT |
| Triandafilou et al. (2011) | Y | CT | N | N | CT | CT |

Note: Y, yes; N, no; CT, cannot tell; N/A, not applicable.

Table 4. COSMIN Risk of Bias checklist to assess the methodological quality of the included reliability studies

|  |  |  |
| --- | --- | --- |
| **Design requirements** | **Bonifer and Anderson**  **(2005)** | **Pereira et al.**  **(2015)** |
| Were patients stable in-between measurements? | F | F |
| Time interval appropriate? | D | G |
| Conditions similar for both measurements? | F | F |
| **Reliability analysis** | | |
| ICC for continuous scores? | F | F |
| Kappa for dichotomous, ordinal, or nominal scores? | N/A | N/A |
| Weighted kappa for ordinal scores? | N/A | N/A |
| Weighting scheme described for ordinal scores? | N/A | N/A |
| Any important flaws in design? | P | D |
| **Agreement analysis** | | |
| Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated for continuous scores | N/A | G |
| Percentage (positive and negative) agreement calculated for nominal/ordinal scores | N/A | N/A |
| Any important flaws in design | N/A | D |
| Final score | P | D |

Note: G, good; F, fair; D, doubtful; P, poor; N/A, not applicable

Table 5. Reporting of the gWMFT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author (year) | 13-item/ 14-item gWMFT | Domains of gWMFT scored | | Description of scoring criteria |
| FAS | Performance time |
| Anandabai et al. (2013) | Not reported | ✓ | ✓ | Not reported |
| Arya et al. (2012) | 13-item | ✓ | ✓ | ✓ |
| Bonifer and Anderson (2003) | 14-item | ✓ | ✓ | ✓ |
| Bonifer et al. (2005) | 14-item | ✓ | ✓ | Not reported |
| Demitas-Tatlidede et al. (2015) | Not reported | Not reported | Not reported | Not reported |
| Fischer et al. (2016) | Not clear, cited Bonifer and Anderson (2003) | ✓ | ✓ | Not reported |
| Flinn et al. (2009) | Not clear |  | ✓ | Not reported |
| Iwamuro et al. (2011) | Not clear, reported 13 items |  | ✓ | Not reported |
| Pereira et al. (2015) | 13-item | ✓ | ✓ | ✓ |
| Triandafilou and Kamper (2014) | 14-item |  | ✓ | ✓\* |
| Triandafilou et al. (2014) | 14-item |  | ✓ | ✓\* |
| Triandafilou et al. (2011) | 14-item |  | ✓ | ✓\* |

Note: ✓, reported; ✓\* reported with adaptations

Figure 1. Summary of the literature review search using the PRISMA group flow chart (Moher et al., 2009)

Abstracts screened  
(n = 14)

Duplicates removed  
(n = 21)

Additional records identified through other sources; paper identified through reference list search  
(n = 1)

Records identified through database searching  
(n = 34)

## Identification

## Screening

Records excluded  
(n = 1)

* Review article

## Eligibility

Full-text articles assessed for eligibility  
(n = 13)

Full text articles excluded, with reasons  
(n = 1)

* Alternate gWMFT version used

## Included

Studies included  
(n = 12)