HHS Public Access

Author manuscript

Neurorehabil Neural Repair. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

Neurorehabil Neural Repair. 2015 June ; 29(5): 436–443. doi:10.1177/1545968314553030.

Inter-rater Reliability of the Wolf Motor Function Test-Functional Ability Scale: Why it Matters

Susan V. Duff, EdD, PT, OTR/L, CHT,
Department of Physical Therapy, Thomas Jefferson University, Philadelphia, PA, USA

Jiaxiu He, PhD,
Kellogg School of Management, Northwestern University, Chicago Il, USA

Monica A. Nelsen, DPT, PT,
Division of Biokinesiology and Physical Therapy, Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA, USA

Christianne J. Lane, PhD,
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Veronica T. Rowe, MS, OTR/L,
Department of Occupational Therapy, University of Central Arkansas, Conway, AR, USA. PhD student, Texas Woman's University, Dallas, TX, USA

Steven L. Wolf, PhD, PT, FAPTA, FAHA,
Departments of Rehabilitation Medicine, Medicine and Cell Biology, Emory School of Medicine, Atlanta, GA, USA

Alexander W. Dromerick, MD, FAHA, and
Departments of Rehabilitation Medicine and Neurology, Georgetown University and MedStar National Rehabilitation Hospital, Washington, DC, USA

Carolee J. Winstein, PT, PhD, FAPTA
Division of Biokinesiology and Physical Therapy, Herman Ostrow School of Dentistry; Department of Neurology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Abstract

Background—One important objective for clinical trialists in rehabilitation is determining efficacy of interventions to enhance motor behavior. In part, limitation in the precision of measurement presents a challenge. The few valid, low-cost observational tools available to assess motor behavior cannot escape the variability inherent in test administration and scoring. This is especially true when there are multiple evaluators and raters as in the case of multi-site randomized controlled trials (RCT). One way to enhance reliability and reduce variability is to implement rigorous quality control (QC) procedures.

Corresponding author: Dr. Carolee J. Winstein, Department of Biokinesiology and Physical Therapy, University of Southern California, 1540 Alcazar St., CHP 155, Los Angeles, CA 90089, USA. winstein@usc.edu, 323-442-3290.

Objective—This paper describes a systematic QC process used to refine the administration and scoring procedures for the Wolf Motor Function Test (WMFT)-Functional Ability Scale (FAS).

Methods—The QC process, a systematic focus-group collaboration was developed and used for a phase III RCT, which enlisted multiple evaluators and an experienced WMFT-FAS Rater Panel.

Results—After three staged refinements to the administration and scoring instructions, we achieved a sufficiently high inter-rater reliability (weighted kappa = 0.8).

Conclusions/Implications—A systematic focus-group process was shown to be an effective method to improve reliability of observational assessment tools for motor behavior in neurorehabilitation. A reduction in noise-related variability in performance assessments will increase power and potentially lower the number needed to treat. Improved precision of measurement can lead to more cost effective and efficient clinical trials. Finally, we suggest that improved precision in measures of motor behavior may provide more insight into recovery mechanisms than a single measure of movement time alone.

Keywords

quality; observational; motor control; impairment; assessment; stroke

Introduction

Improvements in motor control or skill which can be causally linked to specific interventions could inform clinical practice and perhaps more importantly, advance knowledge about the mechanisms of recovery. The most robust measures of motor behavior use laboratory-based instruments to record spatial-temporal parameters of position (kinematics) and force (kinetics).[1–3] However, these instrumented assessments are typically costly, time consuming, and require specialized training and equipment for data collection and analysis. As technology advances, more low-cost options with real-time capabilities will become available.[4,5] Yet, most researchers and some practitioners still rely on inexpensive, low-technology measures such as the Wolf Motor Function Test (WMFT)[6] and the Action Research Arm Test (ARAT)[7]. These kinds of assessment tools can be subject to poor reliability (i.e., increased variability), especially when used by multiple raters. Ultimately, a diminished inter-rater reliability can reduce the statistical power for determining efficacy of rehabilitative intervention studies.

One way to control for the inherent risk of diminished inter-rater reliability is through quality control (QC) procedures. The WMFT[6] is a standardized, performance-based, upper extremity (UE) assessment of functional capability for adults post-stroke that includes 15 timed items and 2 strength items. From video-capture, a discrete WMFT-Functional Ability Scale (FAS) score is determined post-hoc to quantify the quality of movement (relative to the less affected side) for each of the 15 timed tasks. The score is based on characteristics of speed, precision, coordination and fluidity--metrics of skill.[8] The WMFT-FAS uses a 6-point ordinal rating scale that ranges from 0 (no use of the affected side attempted) to 5 (normal) for a maximum score of 75 for the fifteen tasks. The psychometric properties of the WMFT-FAS were previously explored in three related papers.[8–10] The minimal detectable change (MDC_{90}) was reported to be 0.37 (range 0.2–0.4) points per task[8] within an overall

change of 20 out of 75 points (27%) (MDC_{95}) across tasks.[10] Morris and colleagues[8] found the inter-rater reliability (ICC) to be ≥ 0.88 when task items were pooled, yet there was inconsistency for individual tasks with median ICCs that ranged from 0.36 to 0.93.

The Interdisciplinary Comprehensive Arm Rehabilitation Evaluation (ICARE) phase III multi-site randomized controlled trial[11] afforded an opportunity to implement a rigorous QC process. This effort was undertaken to strengthen inter-rater reliability of the Rater Panel scores and thereby improve construct validity of the WMFT-FAS for the ICARE trial. Our process resulted in a fine-tuned revision to the administration and scoring instructions (ASI, see Supplementary Material). Ultimately, our aim is to inform clinical investigators about the QC process and to provide the revised ASI to future users.

Methods

Organization of the Phase III Clinical Trial

The ICARE trial with targeted enrollment of 360 participants is a prospective phase III RCT that includes three regional centers (Los Angeles, Washington DC, and Atlanta) and 7 clinical sites (5 in Los Angeles, and 1 each in Washington DC and Atlanta).[11] The primary aim of ICARE is to compare arm and hand recovery in adults early post-stroke who are randomized to one of three groups. The experimental group participates in an outpatient structured therapy program termed the Accelerated Skill Acquisition Program (ASAP) and are compared to those randomized to a dose equivalent usual and customary therapy program.

The primary outcome is the change in the log-transformed WMFT time score at one-year post-randomization. The WMFT-FAS is but one of a large battery of secondary outcomes. As with other multi-site studies, the ICARE trial requires considerable coordination, training and standardization of multiple team members to ensure success. Each site has a team leader and usually one reserve, while each regional center has multiple blinded evaluators (BEs) who formally administer the ICARE assessment battery at each of four designated time points. The Administrative Core employs 3 experienced clinicians (Rater Panel) who rate the fifteen timed tasks from digitally captured footage. Data are collected and managed through the primary database hosted by the central Data Management and Analysis Center (DMAC) and the File Transfer Protocol (FTP) site hosted by the Administrative Core at the University of Southern California in Los Angeles. The flow of data from test administration to rating score submission is illustrated in Figure 1.

Licensed occupational and physical therapists blinded to group assignment, administer the WMFT at baseline and post-randomization at 3 pre-determined follow-up periods: post-therapy, 6 months, and 1 year. The WMFT is filmed using a digital camera (standardized across sites) according to detailed procedures outlined in the ICARE Manual of Procedures (MOP). Prior to study initiation, site-designated BEs attended a 3-day Clinical Research Evaluators Training Meeting in Los Angeles. Training materials and protocols, with video demonstrations are updated when necessary and made available on a secure ICARE web site for ongoing review and use by BEs and other study personnel. Test administration certification requires each BE to demonstrate at least 90% proficiency with the standardized

administration and scoring criteria. Re-certification is required every 6 months until data collection is complete.

Immediately after test administration, the local site edits the digital file and uploads it to a secure FTP server. A trained member of the Administrative Core evaluates and approves each de-identified digital file for quality (i.e., data completeness, audio and visual clarity) and consistency with the ASI essential elements (see Results for details on essential elements). Once the QC check is complete, the digital files are made available to the Rater Panel. Files that fail the QC check are retained and depending on the reason may go back to the BE for remediation. In any case, the test is not re-administered. Failed files are returned to BE's if deemed fixable. For example, if the problem was in editing, the BE can re-edit the digital file from the raw footage. Most importantly, feedback is given to sites to ensure that the same error does not recur, regardless of the reason for failure.

The Rater Panel scores each digitally recorded assessment using the 6-point scale. Initially, panel members independently determined a score for each task. This yielded 3 ratings per task for each digital file (i.e., 15 tasks × 3 raters = 45 scores/test). Thus, without attrition we estimated ~1440 digital files (360 participants × 4 time points) at ~ 30–45 minutes/each for accurate scores per rater.

Timeline for QC Process

To improve consistency in test administration across sites and to verify the reliability of WMFT-FAS scoring, the ICARE clinical leadership in collaboration with the BEs and Rater Panel implemented a QC process. Prominent events of the systematic QC process included: refinement of the WMFT template (i.e. table-top mat), clarification and refinement of the ASI, statistical analysis of the Rater Panel scores and initiation of monthly Rater Panel meetings. These events are chronicled in Figure 2.

Three revisions were made to the ASI to simplify instructions, specify the set-up and reduce scoring ambiguity (see Results for details). After each revision, the changes were reviewed in meetings held with the site team leaders, BEs and Rater Panel members. Each site and BE provided signed documentation that the revised protocols were reviewed and replaced in the MOP binders; master copies were also available through the secure ICARE web site.

In concert with the ASI revisions, the DMAC conducted five rounds of inter-rater reliability testing based on scores generated by the Rater Panel. The first two rounds used digital files from a prior study completed before ICARE and the later three rounds used digital files from the ICARE trial itself. Each panel member rated the digital files and submitted their scores to the DMAC through the secure FTP site.

Monthly web-based meetings of the Rater Panel began after round 2 of reliability testing. These meetings were initiated in an effort to hone the panel's objective rating skills and to identify scoring criteria that needed clarification. After independent review of sample digital files, discrepancies in rating were identified by the DMAC. Each rater shared the decision process they used to rate discrepant items and a meaningful discussion ensued. During meetings boundaries for determining whether a task was performed "very slowly (score 2),

slower (score 3), or slightly slower (score 4)" were discussed and differences between raters on specific task scores (identified by the DMAC) were reviewed. An examination of the boundaries for each scoring category included a review of the actual movement time of task performance and a comparison with the less-involved limb's performance. However, a specific time interval for each category was not designated because, in most cases, it was determined relative to the less-involved limb. Specific instructions about the WMFT FAS rating and the speed of movement are included in the ASI (Supplementary Material, Section VI.B.) in the section beginning, "For determination of normal…" The rater panel agreed that the difference between speed categories was best determined through examples and discussion during the monthly meetings. Discrepancies were adjudicated by majority vote.

Statistical Analysis

Inter-rater reliability was assessed using the weighted kappa for all five rounds of testing. Weighted kappa (k_w) is a measure of concordance for ordinal outcomes that weights larger disparities in ratings higher than smaller differences, [12–13] that was appropriate given the 6-point scaling. A quadratic weighting was used for these analyses in order to penalize greater discordance in ratings more severely than would occur with a linear weighting. For example, a difference in ratings of a 2 and a 3 would be seen as a much smaller disparity in rating than a difference in rating between a 2 and a 5. Weighted kappa was utilized, as the items under review were ordinal. While ICCs are appropriate for interval level data, they are inappropriate for ordinal level data, as they require calculation of parametric outcomes that are not meaningful in non-parametric data. To guide our efforts toward the achievement of inter-rater reliability, we looked to the "benchmarks" that Landis and Koch[14] used and set a lower bound of 0.80, which should represent "substantial" agreement.

Results

Revisions to Administration and Scoring Instructions

The initial version of the ASI adopted for the ICARE MOP was the same as that used in the EXCITE trial.[15] Prior to randomization for ICARE, Revision 1 was made to the original table-top template and the ASI (Figure 2). Instructions on positioning for task execution and camera placement were revised and placement markers for each item were labeled on the template. Participant instructions for the quick demonstrations were condensed for all items. For example, for Item #10 the instructions for the quick demonstration now reads, "Pick up the pencil as fast as you can."

Revision 2 of the ASI was completed after randomization commenced. Refinements included specifications on chair positions, participant/object proportions (i.e., box size) and filming strategies. For example, the filming position "Side-Close" was specified to zoom in on fine-motor skills. Instructions for task items were also clarified. For example, for Item #12 Stack checkers, we added, "Do it like this (demonstrate correct method), not like this (demonstrate incorrect method of at least one checker still touching the table surface in the stacked position)." Finally, a description of template placement during test administration was added.

Revision 3 was finished after round 4 of statistical testing and included clarification of the scoring criteria and the addition of 'essential' and 'desirable' elements for each item. Essential elements are the specific features that must be accomplished in order for the task to be deemed 'complete'. The desirable elements are other qualitative features that should be included in the task but are not necessary for completion. If a participant cannot complete the essential elements they are given a timed score of 120+ (coded as −7). An essential element in Item #15 Turing key in lock, reads, "Turns key fully each direction and back to vertical". Two desirable elements for Item #15 are, "lateral pinch and turns key to the instructed direction first". Revision 3 also included photos of task demonstrations. In the *General Comments* section details were added such as, "If a BE makes a mistake with set-up or timing, the task may be repeated an additional time". Consultation with investigators from the EXCITE trial and the BE team at Emory University guided the third ASI revision.

Reliability Testing—Reliability testing began during the Rater Panel's training period. The five rounds of testing were interspersed with ASI revisions, review of previous reliability findings, consultation with WMFT-FAS experts and quality checks of digital files (Figure 2). The timeline reveals that this process transpired over a 3-year period.

In round 1, some k_w values between raters were below our 0.8 criterion (0.67 to 0.83). Therefore, the ICARE leadership team consulted the reliability findings from Morris et al.[8] The pooled reliability was reportedly high,[8] but the inter-rater reliability for more than half of the tasks was lower than the 0.80 criterion set for the ICARE trial.[8] For at least one of two test periods, inter-rater reliability for 9/15 tasks was < 0.75 and ICCs for a number of tasks (i.e., #1, 4, 7, 10 and 11, see Figure 3 legend for task ID) were <0.50.[8] This important factor contributed to our decision to seek to improve inter-rater reliability to a greater level than round 1 of ICARE and that reported by Morris et al.[8]

For round 2, inter-rater reliability was inconsistent (k_w values 0.62 to 0.87). Only one rater pair (raters 2 and 3) reached the k_w criteria of 0.80 in the first two rounds of statistical testing (Figure 2). Therefore, the ICARE clinical leadership team, in collaboration with the DMAC, requested that each Rater Panel member score every ICARE digital file (i.e., 3 scores per file) and meet as a group once/month to review and adjudicate task items with between rater scoring discrepancies.

For round 3 all k_w values (0.76 to 0.89) were higher than for round 2. Yet, round 4 k_w values showed a reduction in agreement between rater 2 and the other two raters (0.51 and 0.66) compared with the first three rounds. Additional QC methods were then pursued to examine the basis for the diminished reliability. This involved a heightened screening process in which the digital files were reviewed for confirmation of clear visibility of the essential task elements and accuracy of editing. After round 4, the ASI underwent a 3rd and final refinement.

For round 5, the k_w pooled reliability was above criterion (0.81 to 0.86). Yet, the k_w between raters for the fifteen tasks in round 5 was less consistent (Figure 3). Three tasks fell below a k_w of 0.70 for at least two rater pairs including; "Forearm to table" (#1; 0.66), "Forearm to box" (#2; 0.66), and "Reach and retrieve" (#8; 0.63). Of note, these 3 tasks

were among those Morris et al.[8] found to have a low level of agreement on at least one of two test periods (#1 = 0.52; #2 = 0.57; #8 = 0.61). For ICARE, raters 1 and 3 had the highest agreement (k_w > 0.7 all tasks) and raters 2 and 3 had the lowest agreement for the "Reach and retrieve" task (#8 = 0.63). It is important to note that a kappa value of 0.80 is a very stringent cutoff, representing almost perfect alignment of scores. Kappa's > 0.60 are considered very highly associated.

Once an inter-rater weighted kappa ≥ 0.8 was achieved by round 5 for the pooled tasks, 90% of the digital files were each assigned to only one rater with ~10% (randomly selected by the DMAC) distributed to all 3 raters for independent scoring. This shift in allocation from three- to one-rater per digital file reduced the time and cost of the rating process. Periodic examination of rater reliability (10% of digital files) has demonstrated that inter-rater reliability is being maintained at levels > k_w = 0.70, with most > 0.80.

Discussion

Findings from recent multi-site definitive neurorehabilitation RCTs have been less optimistic than the phase II trials that preceded them;[16–19] primarily because it is difficult to replicate the group differences from smaller scale trials in multi-site large scale trials.[20–22] This phenomenon may be due at least in part, to the inherent confounding factors introduced when conducting bench to bedside work, and the lower methodological rigor often tolerated in smaller single-site compared with larger multi-site trials.[20–23] This paper focuses on one potential confounder--random error introduced during administration and scoring of an observationally-based motor behavior assessment. Improved inter-rater reliability is one way to increase the sensitivity of these types of measures in multi-site RCTs.

The ICARE trial was powered on the log WMFT-time score, the primary outcome variable that will be used to determine the efficacy of the experimental therapy protocol.[11] The QC process we describe here was implemented for one of the secondary outcome measures—one that will be important for interpreting changes in the WMFT time score. The systematic QC process included modifications to the ASI criteria and quality checks of digital files. This process likely elevated the construct validity of this secondary outcome measure. Clinicians and researchers who wish to establish substantial agreement in using the WMFT-FAS should find the details and knowledge gained by the ICARE team particularly helpful for future endeavors.

There is no doubt that the QC process we describe is time consuming, costly, and requires considerable resources to implement. Given that the WMFT-FAS was a secondary outcome measure, why implement such a rigorous, resource-consuming process? What might be the benefit of improved inter-rater reliability? Recently, See and colleagues[24] showed that a standardized training approach used with examiners for a phase II controlled trial significantly reduced variability in scoring on the Fugl-Meyer Assessment (FMA) UE Scale. Data analysis revealed that the improved reliability on the FMA decreased the variance in scoring by 20%. In turn, a 20% reduction in variance on the FMA would allow a reduction in sample size from 137 to 88 to detect group differences for a trial powered at 80%.[24] For the ICARE Trial, an improved WMFT-FAS inter-rater reliability could effectively

strengthen the sensitivity to detect group differences. However we cannot know the possible impact on ICARE until we are permitted to analyze group data (expected, August, 2014). For studies in which the WMFT-FAS is a primary measure,[25–27] an improved reliability could lead to increased power. As shown recently, even small decreases in variability can have a large effect on the sample size required to detect a statistically significant effect.[23] Furthermore, a decrease in sample size could have a very large effect on the cost of conducting a clinical trial. Use of the revised WMFT ASI (Supplementary Material) could minimize the need for an extensive QC effort, decrease the cost and increase the efficiency of future single- and multi-site clinical trials in stroke rehabilitation.

Recently, Woodbury and colleagues[28] used Rasch analysis to establish a hierarchy of item difficulty for 14/15 items based on the rating scale of the WMFT-FAS. From that analysis the authors discovered that Item #8, “Reach and retrieve”, had abnormally high missing values due to administration and filming errors. Without this item the authors were able to establish an item difficulty hierarchy to show that higher scores on difficult items were associated with higher UE function. For ICARE, greater reliability and precision of the WMFT-FAS score will be important for determining whether the structured intervention significantly contributed to an improvement in UE motor behavior and skill above that achieved for the dose equivalent usual and customary treatment group.

Limitations

One limitation was that initially we relied on the inter-rater reliability of the WMFT-FAS reported in Morris et al[8] and assumed that we would achieve the same if not a higher rater agreement level. After round 1 of reliability testing and closer examination of the findings for individual items in Morris and colleagues[8] we realized that this assumption was incorrect. In hindsight, we should have scrutinized previous results more carefully before conducting round 1. This may have allowed us to initiate strategies to improve inter-rater reliability earlier in the trial.

Another limitation is that we assumed that the initial in-person training provided to the blinded evaluators would be sufficient. However, over the course of the trial new BE’s who joined the team did not receive this in-person training; although they did have web site access to training materials, protocols, video demonstrations and in person local experienced BEs. In previous work, the investigators suggested that evaluator training may have been insufficient due to the low agreement level for many items[8]. See and colleagues[24] showed that standardized training and testing after training increased reliability of the FMA. We speculate that in-person training and greater scrutiny of the knowledge and skill of all BE’s could have improved consistency in WMFT execution and may have hastened the achievement of inter-rater reliability.

A final limitation pertains to the web-based group meetings of the Rater Panel. Although this process was deemed beneficial overall, we note that there was an initial familiarization period that may have adversely influenced individual scoring strategies and subsequent ratings assigned by the panel members. Initially, the raters reported a tendency to second guess their first responses and predict how the other raters would score. This context effect dissipated with time and familiarization with the process.

Recommendations for Clinical Researchers

From this systematic QC process we offer a few recommendations to investigators who plan to use the WMFT-FAS in their clinical trial research. 1) The quality of the visual media capture is critical. As such, close attention to the camera setup is strongly recommended to assure sufficient visualization of the essential and desirable elements (see Supplementary Material). 2) Frequent and consistent (pre-planned) web-based Rater Panel focus-group meetings are recommended to sharpen rater skills, foster consistency in scoring and maintain these skills over the entire course of the study. 3) Implementation of regular meetings for the blinded evaluators is recommended to maintain consistency in test execution over time and across sites. Remote meeting formats such as Go-to-Meeting, WebEx or Adobe Connect are useful for recommendations 2 and 3. 4) To further strengthen inter-rater reliability of the WMFT-FAS we recommend removal of the most problematic items in future revisions including: Forearm to table (#1); Forearm to box (#2); and Reach and retrieve (#8) (see rationale in Results).[8,11,28] While standardization procedures are common for implementation of clinical trials research, the first two recommendations are unique aspects of QC enforced in the ICARE study. Finally, we suggest that the first three recommendations can be generalized for use with comparable quality-based motor performance measures in the context of multi-site RCTs.

Conclusions/Implications

The effort expended to modify the ASI procedures and achieve a substantial level of inter-rater reliability likely enhanced the construct validity of the WMFT-FAS instrument. We detail the systematic QC process developed for ICARE so that others may benefit from a more sensitive and objective measure of motor behavior. We believe that the process of strengthening the psychometric properties of observationally based motor behavior measures is vital to advancing the science of our field and to enhancing our understanding of the mechanisms of recovery. This concern is timely given the recent fervent discussions surrounding the nature of sub-optimal motor recovery (recovery vs. compensatory) [29–32] and impairment-based vs. task-based intervention protocols.[1,33] As the number of clinical trials in neurorehabilitation grows and we attempt to demonstrate sufficient efficacy and effectiveness of our interventions, it is essential that we use reliable tools that provide information pertaining to restitution and substitution strategies.

Observational measures that provide reliable information about ‘how the movement was performed’ with the addition of temporal measures (e.g., movement time) could offer new insights about the recovery process. The WMFT-FAS complements the WMFT-time score. Thus, we would ideally like the scores from both measures to improve. Yet, discrepancies may provide greater insight.[34] For example, if movement time decreases while the WMFT-FAS score is unchanged, this may suggest that the improvement stems from sub-optimal compensatory strategies. Therefore, complementary measures of motor behavior and movement time should be used to better understand the mechanisms of recovery that are impacted by rehabilitative interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by grants from the National Institute of Neurological Disorders and Stroke and the Eunice Kennedy Shriver National Institute of Child Health and Human Development under award numbers U01 NS056256 and T32 HD064578.

Acronyms

ASAP	Accelerated Skill Acquisition Program
ASI	Administration and scoring instructions
DMAC	Data Management and Analysis Center
BE	Blinded Evaluator
FAS	Functional Ability Scale
ICARE	Interdisciplinary Comprehensive Arm Rehabilitation Evaluation
ID	Identity
ICCs	Interclass correlations
MOP	Manual of Procedures
QC	Quality control
UE	Upper Extremity
WMFT	Wolf Motor Function Test

References

1. Kitago T, Liang J, Huang VS, et al. Improvement after constraint-induced movement therapy: recovery of normal motor control or task-specific compensation? Neurorehabil Neural Repair. 2013; 27:99–109. [PubMed: 22798152]
2. van Kordelaar J, van Wegan EE, Kwakkel G. The impact of time on quality of motor control of the paretic upper limb after stroke. Arch Phys Med Rehabil. 2014; 95:338–44. [PubMed: 24161273]
3. Alberts JL, Butler AJ, Wolf SL. The effects of constraint-induced therapy on precision grip: a preliminary study. Neurorehabil Neural Repair. 2004; 18:250–8. [PubMed: 15537995]
4. Dobkin BH, Dorsch A. The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. Neurorehabil Neural Repair. 2011; 25:788–98. [PubMed: 21989632]
5. Wade E, Chen C, Winstein CJ. Spectral analysis of wrist motion in individuals poststroke: the development of a performance measure with promise for unsupervised settings. Neurorehabil Neural Repair. 2013; 28:169–78. [PubMed: 24213957]
6. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf Motor Function Test as outcome measure for research in patients after stroke. Stroke. 2001; 32:1635–1639. [PubMed: 11441212]
7. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and reseach. Int J Rehab Research. 1981; 4:483–492.

8. Morris DM, Uswatte G, Crago JE, et al. The reliability of the Wolf Motor Function Test for assessing upper extremity function after stroke. Arch Phys Med Rehabil. 2001; 82:750–755. [PubMed: 11387578]
9. Lin KC, Hsieh YW, Wu CY, Chen CL, Jang Y, Liu JS. Minimal detectable change and clinically important difference of the Wolf Motor Function Test in stroke patients. Neurorehabil Neural Repair. 2009a; 23:429–434. [PubMed: 19289487]
10. Lin JH, Hsu MJ, Sheu CF, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys Ther. 2009b; 89:840–50. [PubMed: 19556333]
11. Winstein CJ, Wolf SL, Dromerick AW, et al. Interdisciplinary comprehensive arm rehabilitation evaluation (ICARE): a randomized controlled trial protocol. BMC Neurol. 2013; 13:5. [PubMed: 23311856]
12. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20:37–46.
13. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213–220. [PubMed: 19673146]
14. Landis JR, Koch GG. The measurement observation agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
15. Wolf SL, Winstein CJ, Miller JP, et al. Effect of constraint induced movement therapy on upper Extremity function 3 to 9 months after stroke: the EXCITE randomized clinical trial. JAMA. 2006; 296:2095–2104. [PubMed: 17077374]
16. Dobkin B, Barbeau H, Deforge D, et al. Spinal Cord Injury Locomotor Trial Group. The evolution of walking-related outcomes over the first 12 weeks of rehabilitation for incomplete traumatic spinal cord injury: the multicenter randomized Spinal Cord Injury Locomotor Trial. Neurorehabil Neural Repair. 2007; 21:25–35. [PubMed: 17172551]
17. Duncan PW, Sullivan KJ, Behrman AL, et al. Body-weight-supported treadmill rehabilitation after stroke. N Engl J Med. 2011; 364:2026–36. [PubMed: 21612471]
18. Harvey RL, Winstein CH. Everest Trial Group. Design for the Everest randomized trial of cortical stimulation and rehabilitation for arm function following stroke. Neurorehabil Neural Repair. 2009; 23:32, 44. [PubMed: 18812431]
19. Lo AC, Guarino PD, Richards LG, et al. Robot-assisted therapy for long-term upper-limb impairment after stroke. N Engl J Med. 2010; 362:1772–83. [PubMed: 20400552]
20. Adkins DL, Schallert T, Goldstein LB. Poststroke treatment: lost in translation. Stroke. 2009; 40:8–9. [PubMed: 19038911]
21. Dobkin BH. Progressive staging of pilot studies to improve Phase III trials for motor interventions. Neurorehabil Neural Repair. 2009; 23:197–206. [PubMed: 19240197]
22. Dobkin BH, Duncan PW. Should body-weight supported treadmill training and robotic-assisted training trot back to the starting gate? Neurorehabil Neural Repair. 2012; 26:308–317. [PubMed: 22412172]
23. Krakauer JW, Carmichael ST, Corbett D, Wittenberg GF. Getting neurorehabilitation right: what can be learned from animal models? Neurorehabil Neural Repair. 2012; 26:923–31. [PubMed: 22466792]
24. See J, Dodakian L, Chou C, et al. A standardized approach to the Fugl-Meyer Assessment and its implications for clinical trials. Neurorehab Neural Repair. 2013; 27:732–41.
25. Caliandro P, Celletti C, Padua L, et al. Focal muscle vibration in the treatment of upper limb spasticity: a pilot randomized controlled trial in patients with chronic stroke. Arch Phys Med Rehabil. 2012; 93:1656–61. [PubMed: 22507444]
26. Fabbrini S, Casati G, Bonaiuti D. Is CIMT a rehabilitative practice for everyone? Predictive factors and feasibility. Eur J Phys Rehabil Med. 2014 Epub.
27. Patten C, Condliffe EG, Dairaghi CA, Lum PS. Concurrent neuromechanical and functional gains following power training post-stroke. J Neuroeng Rehabil. 2013; 10:1–19. [PubMed: 23336711]
28. Woodbury M, Velozo CA, Thompson PA, et al. Measurement structure of the Wolf Motor Function Test: implications for motor control theory. Neurorehabil Neural Repair. 2010; 24:791–801. [PubMed: 20616302]
29. Levin ML, Kleim J, Wolf SW. What do motor “recovery” and “compensation” mean in patients following stroke? Neurorehabil Neural Repair. 2009; 23:313–319. [PubMed: 19118128]

30. Lum PS, Mulroy S, Amdur RL, Requejo P, Prilutsky Bl, Dromerick AW. Gains in upper extremity function after stroke via recovery or compensation: Potential differential effects on amount of real-world limb use. Top Stroke Rehabil. 2009; 16:237–53. [PubMed: 19740730]

31. Michaelsen SM, Jacobs S, Roby-Brami A, Levin MF. Compensation for distal impairments of grasping in adults with hemiparesis. Exp Brain Res. 2004; 157:162–73. [PubMed: 14985899]

32. Shaikh T, Goussev V, Feldman AG, Levin MF. Arm-trunk coordination for beyond-the-reach movement in adults with stroke. Neurorehabil Neural Repair. 2013; 28:355–66.3. [PubMed: 24270057]

33. Corti M, McGuirk TE, Wu SS, Patten C. Differential effects of power training versus functional task practice on compensation and restoration of arm function after stroke. Neurorehabil Neural Repair. 2012; 26:842–54. [PubMed: 22357633]

34. Duff M, Chen Y, Cheng L, et al. Adaptive mixed reality rehabilitation improves quality of reaching movements more than traditional reaching therapy following stroke. Neurorehabil and Neural Repair. 2013; 27(4):306–315.

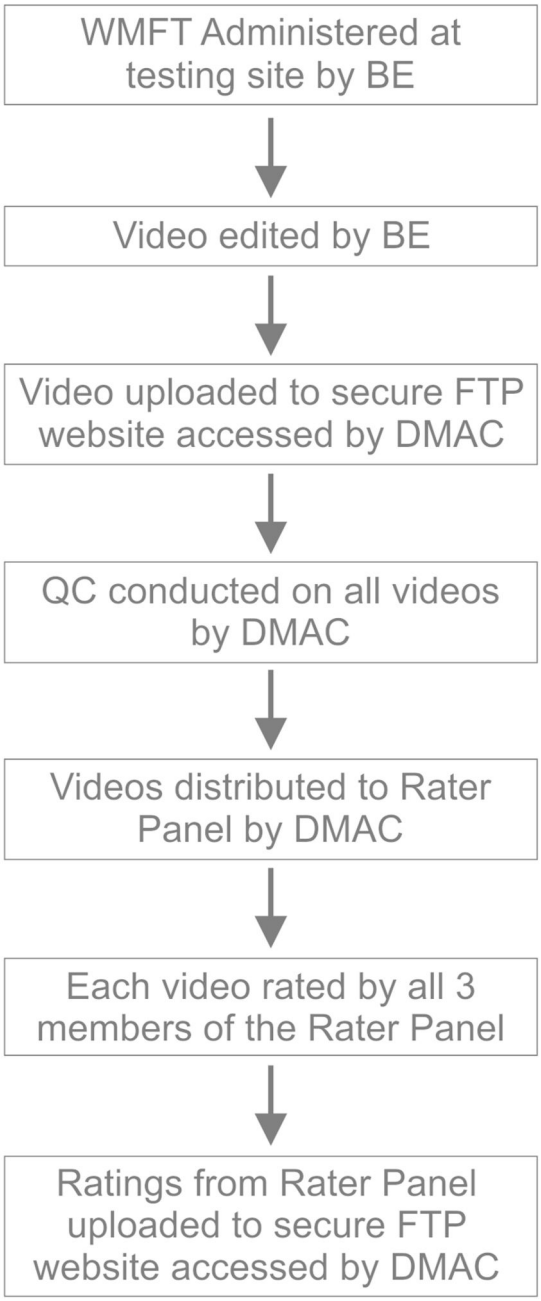

Figure 1.

Flow diagram. The flow of data from initial administration of the WMFT by the BE to uploading of scores to the FTP site by the Rater Panel member is shown. WMFT = Wolf Motor Function Test, BE = Blinded Evaluator, FTP = File Transfer Protocol.

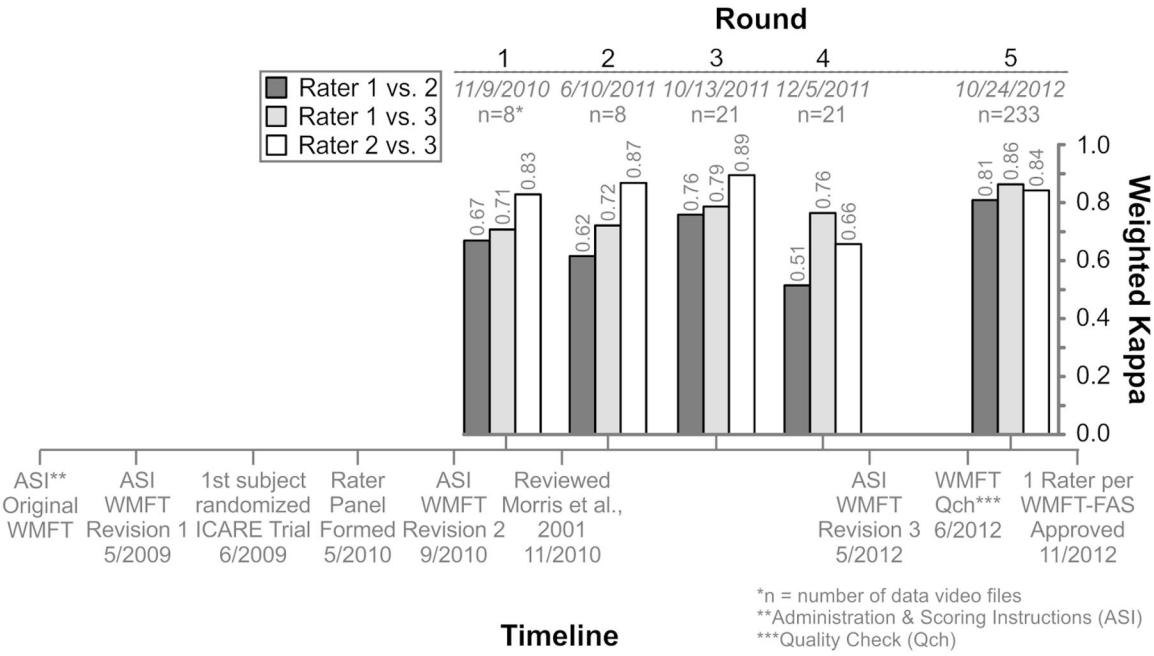

Figure 2.

Timeline for systematic quality control process. Top – Inter-rater reliability (quadratic weighted kappa) for the WMFT-FAS Rater Panel overall for each round of analysis for each of the three rater pairs. The number of digital files (n) included in each round is listed. Bottom – Timeline of events (with dates) for the ICARE trial including revisions to the WMFT-FAS administration and scoring instructions (ASI) and other notable events before, during and after the 5 rounds of analysis. QCh = quality check conducted to ensure essential elements were met for each digital file.

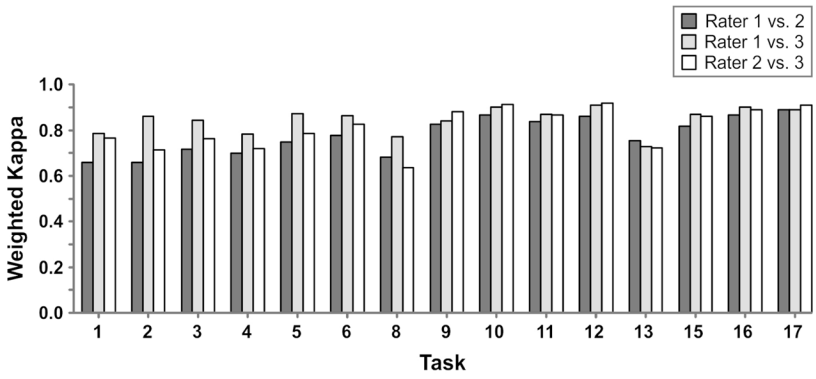

Figure 3.

Inter-rater reliability by task at Round 5. The quadratic weighted kappa values between members of the WMFT-FAS Rater Panel are shown for each task. The fifteen timed tasks are: 1) Forearm to table; 2) Forearm to box; 3) Extend elbow; 4) Extend elbow with weight; 5) Hand to table; 6) Hand to box; 8) Reach and retrieve; 9) Lift can; 10) Lift pencil; 11) Lift paper clip; 12) Stack checkers; 13) Flip cards; 15) Turning key in lock; 16) Fold towel; and 17) Lift basket. [Note that the two force tasks are not listed, but the full WMFT task numbering has been retained].