Reliability of the Test of Wheeled Mobility (TOWM) and the Short Wheelie Test
Fliess-Douer, Osnat; Van der Woude, Lucas H. V.; Vanlandewijck, Yves C.

Published in:
Archives of Physical Medicine and Rehabilitation

DOI:
10.1016/j.apmr.2012.09.023

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2013

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 18-09-2023
Reliability of the Test of Wheeled Mobility (TOWM) and the Short Wheelie Test

Osnat Fliess-Douer, PhD, Lucas H.V. Van Der Woude, PhD, Yves C. Vanlandewijck, PhD

From the Faculty of Kinesiology and Rehabilitation Sciences, Department of Rehabilitation Sciences, Catholic University of Leuven, Leuven, Belgium; and Center for Human Movement Sciences, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

Abstract

Objective: To assess the reliability of the Test of Wheeled Mobility (TOWM) and the Wheelie test.

Design: Cohort study.

Setting: Gymnasium.

Participants: Manual wheelchair users (N = 30, age 23–53y) with a spinal cord injury.

Intervention: Participants performed the 30 skills of the TOWM and the 8 skills of the Wheelie test twice. Ability, time, and anxiety scores were assessed on field. Quality scores were assessed by video analysis.

Main Outcome Measures: Test-retest reliability was evaluated for the ability, time, anxiety, and quality scores of both tests. Intrarater and interrater reliability were determined on the basis of quality scores of 20 participants. Intraclass coefficient and nonparametric statistics were applied, as well as standard error of measurement, method error (ME), coefficient variation of ME, minimal detectable change (95% confidence), and technical error of measurement.

Results: Test-retest reliability: no significant differences between t1 and t2 in the ability, quality, and time scores, except for anxiety scores. Standard error of measurement, ME, coefficient variation of ME, and minimal detectable change (95% confidence) values were low for the ability and quality total score and higher for the time and anxiety total score. Intrarater and interrater reliability interclass correlation coefficients of both tests ranged between .91 and .99. Interrater relative technical error of measurement for the TOWM and the Wheelie test total quality score was 3.7% and 6.3%, respectively, and intrarater relative technical error of measurement was 4.3% and 6.1%, respectively. Intraclass correlation coefficients per individual tasks ranged between .88 and 1.00, except for “level propulsion forward,” which showed low interclass correlation coefficient scores (interrater: .49; intrarater: .44; test-retest: .43).

Conclusions: Based on ability and quality total scores, the TOWM and the Wheelie test are reliable when assessing the wheeled mobility of manual wheelchair users with spinal cord injury. The quality criteria of 1 task from the TOWM and 3 tasks from the Wheelie test need to be refined.

Archives of Physical Medicine and Rehabilitation 2013;94:761-70

© 2013 by the American Congress of Rehabilitation Medicine

Wheeled mobility (WM) is a key to independence for a large group of people with a handicap, especially for those who will not return to walking again, such as individuals with a complete spinal cord injury (SCI). Learning wheelchair skills performance is an important part of the rehabilitation process, because for many people with SCI, the wheelchair will serve as the main device for mobility.

WM is defined by the International Classification of Functioning, Disability, and Health as “Moving around using equipment; moving the whole body from place to place, on any surface or space, by using specific devices designed to facilitate moving or create other ways of moving around, such as moving down the street in a wheelchair or a walker.” Participation is also an important rehabilitation outcome for persons with SCI. In the International Classification of Functioning, Disability and Health, participation is defined as “involvement in life situations,” including, for example, work and school, social relations, and community organizations. Participation restrictions are the problems that an individual may face in involvement in life situations. From the literature, it is known...
that persons with activity limitations experience participation restrictions in daily life and there is a positive relationship between WM skills and participation in persons with SCI.\textsuperscript{3} Therapists should have a valid, reliable, and sensitive measuring tool at their disposal to objectively and systematically assess their patient’s level of WM performances, before, during, and after interventions. Currently, several wheelchair skills tests are available, based on actual performance.\textsuperscript{4,16}

In a recently published systematic literature review on wheelchair skills tests,\textsuperscript{17} results showed that only a few tests focus explicitly on WM in persons with SCI.\textsuperscript{5,11,13,14} Wheelchair skills tests that were aimed at the general wheelchair users’ populations failed to differentiate between levels of performance and resulted in a “ceiling effect,” mainly in individuals with paraplegia.\textsuperscript{17} The review study revealed a lack of a broadly accepted wheelchair skills tests, and disclosed large inconsistencies among the current available tests, which made comparison of study results impossible.\textsuperscript{18} Furthermore, different scales were applied to express test scores. Some tests used qualitative scales, whereas others used quantitative scales.

A quantitative measurement can be, for example, the time necessary to complete a task or the percentage of a slope. The qualitative measurement has to be well defined. It is not sufficient to apply only a pass/fail scale, because the same level of difficulty can be completed with a different grade of maturity. In this case, it can be useful to incorporate the term “with difficulties,” or to also combine a performance time assessment (quantitative and qualitative measurements within the same test). Still, these scales might not be sensitive enough to detect small changes in WM because they only record whether a person can perform a particular task or whether the task is performed independently. Small improvements in the quality of the skill performance often cannot be scored. Nonetheless, these small changes can be of great importance. The way to overcome this problem would be through developing quality assessment criteria for each WM skill. These criteria (based on key components that compose the skill) will reflect the WM maturity proficiency.

Within the scope of a study aimed at promoting a standardized broadly accepted and applicable WM test, the ‘Test of Wheeled Mobility (TOWM) and the short Wheelie tests were developed.\textsuperscript{18} The development of these 2 tests was based on experts’ opinions and on the outcome of an international survey among users, aimed to create a sorted list of the most essential WM skills.\textsuperscript{19} In addition, a systematic critical literature review of available WM skills tests enabled the development of the new tests, relying on strengths of existing tools.\textsuperscript{17} The purpose of the TOWM and the Wheelie test is to assess WM skills in manual wheelchair users with SCI during and after clinical rehabilitation, allowing accurate monitoring and assessment of small changes in WM. The TOWM and the Wheelie test are primarily designed for clinical purposes, but they may also be used in a research setting. Both assessment tools were tested for their feasibility and validity with respect to duration, costs, content, construct, convergent, and predictive validity.\textsuperscript{18} The protocols of the TOWM and the Wheelie test with descriptions of the tasks as well as testing equipment and score sheets can be obtained at www.scionn.nl/inhoudp28.htm.

The reliability of any new measurement instrument is critical to ensure that the measurement error is small enough to detect actual changes in what is being measured.\textsuperscript{20} The nature of reality is such that because of instrument imprecision and human inconsistencies, measurements are not free of error (ie, perfectly reliable). The aim of this study was to assess the reliability and response stability of the TOWM and the Wheelie test. Responsiveness refers to the ability of a measure to detect clinically meaningful change over time, and provides a means for determining whether an individual’s change in score is related to true recovery, or to natural variation in repeated performances. Scale responsiveness is an important concept for clinicians in this time of evidence-based practice, and understanding and interpreting the responsiveness of a scale enables clinicians to discriminate true change from measurement error.\textsuperscript{21}

Therefore, the objectives of this study were (1) to determine the test-retest reliability and response stability of the TOWM and the Wheelie test on the basis of 4 scales total scores, (2) to assess the inter- and intrarater reliability of both tests’ tasks, on the basis of agreement between and within raters, as well as on technical error of measurement (TEM) index (accuracy assessment), and (3) to determine the minimum detectable change at 95% confidence (MDC\textsubscript{95}) for both tests in a group of manual wheelchair users with SCI.

**Methods**

**Participants**

Thirty wheelchair users with SCI (convenience sample) were recruited during the first 2 weeks of February 2011. Recruitment was performed by word of mouth, e-mail, and telephone calls. All participants were living in Belgium, were post—clinical rehabilitation, used a hand-rim wheelchair, and were between 18 and 65 years of age. Potential participants were not included if they had a current cardiorespiratory disorder or orthopedic or other medical complications that restricted them from performing the tasks required for the TOWM and the Wheelie test. One participant did not attend the retest (t2) because of a recurrence of an old shoulder injury unrelated to the WM testing; therefore, the results of 29 participants were included in the presented data analysis.

All procedures were performed in accordance with the guidelines of the Declaration of Helsinki. The study was approved by the Medical Ethics Committee of the Catholic University of Leuven, Belgium. Prior to participation, all participants signed an informed consent form. Participants were reimbursed for transportation costs.

**Measuring instruments**

The TOWM consists of 30 standardized tasks that are conditional to mobility in persons with SCI.\textsuperscript{18} The short Wheelie test includes 8 tasks that are related to the ability to perform a mature pattern of a wheelie (balancing on the rear wheels). The TOWM and the Wheelie test tasks present different difficulty levels and are

---

**List of abbreviations:**

- CI confidence interval
- CV\textsubscript{AE} coefficient of variation of method error
- ICC interclass correlation coefficient
- MDC minimal detectable change
- MDC\textsubscript{95} minimal detectable change (95% confidence)
- ME method error
- SCI spinal cord injury
- TEM technical error of measurement
- TOWM Test of Wheeled Mobility
- WM wheeled mobility
ability score refers to all the items that can be performed sufficiently and independently, and is being assigned as 1 point if the participant completes the task successfully in the first trial, 0.5 points if he or she succeeds on the second trial, and 0 score presents either a failure or avoid trying. The performance time score is the sum of the performance times (seconds) of 2 tasks of the TOWM and 4 tasks of the Wheelie test.

In addition to these 2 scoring methods that were already used in previous tests, an anxiety score (tested by using a visual analog scale [0–10] prior to each task performance) and a qualitative score (based on skill maturity criteria) were introduced. The latter are unique scoring methods and as far as we know were never used before in instruments aiming to assess wheelchair skills. The intentions of developing these methods were, first, to include the psychological impact (anxiety) on WM performance, and, second, to qualitatively assess each participant in detail, according to the maturity of his or her WM proficiency.

A specific focus of this study was targeting the reliability assessment of this newly developed quality scale; in this scale, scoring is based on a process-oriented assessment. For each skill of the TOWM and the Wheelie test, 5 components were selected as the most important quality “performance criteria.” If a participant perfectly accomplishes these 5 components while performing a task, it constitutes a “mature pattern of the skill.” This method provides meaningful feedback to both the therapist and the client, regarding how well one performed each WM task, what exactly were the mistakes, and how mature was the overall quality of his or her WM proficiency, and it allows later on to use the score sheet as a teaching aid and rehabilitation guide.

Testing procedures

A test-retest procedure was undertaken. On test occasion 1 (t1), the research team explained about the study aims and asked the participants to complete a personal information form and to sign the consent form. Body dimensions (participant’s weight, height, and body mass index [\(=\text{weight (kg)/height [m^2]}\]) were also taken. Following these procedures, the research team introduced the course of the tests and ran both tests (TOWM followed by the Wheelie test). After 1 week from t1, all the participants were tested again (t2) at the same place and time of the day, and by the same research team (fig 1). Both tests were performed while using participants’ own daily wheelchairs and the research team did not change the sitting position or configuration at the time of the test. Inspection of the wheelchair configuration (eg, the calculation of the wheelchair center of gravity) was performed at t1 and at t2 to ensure the use of the same wheelchair (configuration and condition) at both test occasions. Center of gravity was tested according to the formula \[x = (F1 \times d) \times (m \times g)^{-1}\], \[y = x \cot (\text{anglincl})\] (with \(F1\) as the weight on the front wheels, \(d\) as the horizontal distance between the rear and the front wheel axle, and \(m \times g\) as the total weight [person + wheelchair]). The \(y\) coordinate of the center of gravity (vertical coordinate) was calculated as follows: \[y = x \cot (\text{anglincl})\]. \(\text{anglincl}\) is determined by the angle over which the wheelchair-user system had to be inclined to decrease \(x\) to zero.

The reliability of the TOWM and Wheelie test

The participants were asked to refrain from smoking and drinking alcohol and caffeine products for at least 2 hours before each trial. For the quality assessment (done after the test), participants were videotaped by the same photographer, using the same camera (Canon Basler 100Hz), placed on a marked line. (The test map is presented in figure 2.)

Reliability assessment

Three raters were involved in the reliability analysis (see fig 1) (all with master’s degrees in physical therapy, with at least 6mo of experience working in the SCI unit, and they all received the same training on using the TOWM and the Wheelie test). For both the test-retest and intrarater reliability assessment, the exact interval between the first and second rating was 1 week.

For test-retest reliability, rater number 1 scored during “real time” the ability scores, performance time scores, and anxiety scores at t1 and at t2. The quality scores of a random selection of 20 participants’ videos taken at t1 and at t2 were assessed after the test by rater number 2 (see fig 1). Stratification was performed to ensure that a full range of lesion levels was represented in the sample (ie, individuals with tetraplegia, and with high and low paraplegia).

Intrarater reliability was assessed by rater number 2 who scored twice (repeatedly, 1wk apart) the same participants’ videotapes (taken at t1). On the second evaluation, the rater was not permitted to review the results of his or her initial evaluation.

Interrater reliability was determined using the quality scores obtained at t1 by 2 independent raters (rater number 2 and rater number 3) (see fig 1).

Statistical procedures

Descriptive statistic and reliability analysis were performed using SPSS (version 16.0) and Microsoft Excel 2010. For measuring test-retest reliability, Wilcoxon signed-rank test was used to verify the absence of significant differences (systematic errors) between the measures at the 2 different times (t1 and t2). The interclass
The correlation coefficient (ICC) with a 95% confidence interval (CI) was used as a measure of reliability for both test-retest reliability and inter- and intrarater reliability. A priori an ICC of .80 or higher was defined as an indication of good reliability.20

In addition to measuring the reliability of instruments and raters, the consistency or stability of repeated participants’ responses was assessed (response stability). Response stability is basic to establish all other types of reliability, because if the response variable varies from measurement to measurement, it will not be possible to separate out errors due to the rater or the instrument. Response stability was expressed in terms of method error (ME), standard error of measurement, and coefficient of variation of ME (CVME).23

ME is a measure of(292,97),(358,118) discrepancy between 2 sets of repeated scores, or their difference scores. Large difference scores reflect greater measurement error. ME is often used as an adjunct to test-retest correlation statistics, as it reflects the percentage of variation from trial to trial, which the correlation coefficient does not. In addition, unlike the correlation coefficient, ME is not affected by the lack of variation in raw scores. ME was calculated using the SD of the difference scores between test and retest:

\[ ME = \frac{SD}{\sqrt{2}} \]

ME should be interpreted relative to the size of the mean differences. Therefore, it was converted to percentages using the CV:

\[ CV_{ME} = \frac{2ME}{X_1 + X_2} \times 100 \]

The standard error of measurement provides a value for measurement error in the same units as the measurement itself; that is, it is an indication of absolute reliability.23 This type of reliability is more clinically applicable on a day-to-day basis, rather than a relative reliability coefficient value, such as an ICC, which is more difficult to interpret for clinical decision making. The standard error of measurement was also calculated using the ICC values as reliability coefficient for the data24:

\[ \text{Standard error of measurement} = SD\sqrt{(1 - ICC)} \]

The standard error of measurement allows the calculation of the minimal detectable change (MDC), which is an estimate of the smallest change in score that can be detected objectively for a client (ie, the amount by which a patient’s score needs to change to be sure the change is greater than measurement error).24 The MDC was calculated to a 95% degree of confidence (MDC95):

\[ \text{MDC95} = \text{standard error of measurement} \times 1.96 \]

The accuracy of the measurements was analyzed by means of the TEM. The TEM index allows verifying the accuracy degree when performing and repeating measurements by the same appraiser (intrarater accuracy) and when comparing the
measurement with measurements from different appraisers (intrarater accuracy). TEM was calculated using the following formula:

\[ \text{TEM} = \sqrt{\frac{\sum D^2}{2N}} \]

where \( D \) is the difference between measurements and \( N \) is the total number of measurements taken. The TEM presents the same measurement unit (cm, mm, points), and at least 20 measurements are required.26

### Results

The mean, SD, and range scores of the TOWM and the Wheelie test are presented in Table 1. Because of various task difficulties, not all persons were able to perform all the tasks of the TOWM and the Wheelie test. If a participant was not able to perform a given task after 2 trials, an ability score of “0” was given and no time score was available. As a result, the group composition of participants differs slightly for the 4 scales.

The mean age of the participants was 38.8±8.0 years, and the time since injury was 12.4±10.5 years. The high values of these SDs demonstrate the large variability in the population group. SCI lesion level ranged from C5 to L1, (tetraplegia \( n = 6 \), high paraplegia \( n = 16 \), low paraplegia \( n = 7 \)). The mean body mass index (kg/m\(^2\)) was 24.2±3.9.

The mean age of the 20 participants who were qualitatively assessed was 39.9±7.12 years (range 26–52y) and time since injury was 13.2±11.5 years (range 1–35y). Mean and SD of the TOWM ability scores of that group was 19.9±4.8 (range 13–28) and of the Wheelie test ability scores 5.6±2.4 (range 0–8). No significant differences were found between the characteristics of the 2 subsample groups (ie, in age, time since injury, and the lesion level).

The center of gravity did not differ between \( t_1 \) and \( t_2 \), ensuring the use of the same wheelchair setup at both tests’ occasions, and ICCs (consistency of the center of gravity examination done by the examiners at \( t_1 \) and \( t_2 \) for axis \( x \) and \( y \)) were high (.92–.98).

### Test-retest reliability based on total scores

No significant differences were found between \( t_1 \) and \( t_2 \) in the ability, quality, and performance time total scores. Only the anxiety scores showed significant differences between \( t_1 \) and \( t_2 \) (the visual analog scale anxiety score was lower at the retest).

For the TOWM, the ICCs associated with the test-retest reliability varied from .91 to .99 and the 95% CIs from .80 to .99 (see Table 1). For the Wheelie test, the ICCs associated with the test-retest reliability varied from .94 to .99 and the 95% CIs from .87 to .99.

Response stability indexes are presented in Table 1. In general, response stability was higher for the ability and quality scales than for the time scale. The anxiety scale had very low response stability, as expected because of the significant differences found between \( t_1 \) and \( t_2 \) in the anxiety total score.

For the TOWM total scores, standard error of measurement values were .63 points for the total ability score and 1.29 points for the total quality score. Time and anxiety total scores showed much higher standard error of measurement. Similarly, ME and CV\(_{MBE}\) values were low for both ability and quality total score and much higher for the time and anxiety total score. MDC\(_{95}\) values were 1.74 points for the total ability score and 3.59 points for the total quality score and much higher for the time and anxiety total score.

### Table 1

<table>
<thead>
<tr>
<th>Scale</th>
<th>n</th>
<th>( t_1 ) (Mean±SD (range))</th>
<th>( t_2 ) (Mean±SD (range))</th>
<th>Difference</th>
<th>Significance*</th>
<th>ICC</th>
<th>95% CI</th>
<th>SEM</th>
<th>ME</th>
<th>( CV_{MBE} ) (%)</th>
<th>MDC(_{95} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability score (scale 0–30)</td>
<td>29</td>
<td>20.6±4.5 (11.5–28)</td>
<td>20.8±4.5 (11.5–29)</td>
<td>0.2</td>
<td>0.34</td>
<td>0.98</td>
<td>.96–.99</td>
<td>.63</td>
<td>.88</td>
<td>4</td>
<td>1.74</td>
</tr>
<tr>
<td>Quality score (scale 0–50)</td>
<td>20</td>
<td>28.8±13.5 (7–48)</td>
<td>28.6±12.7 (7–44)</td>
<td>0.2</td>
<td>0.81</td>
<td>0.99</td>
<td>.99–.99</td>
<td>1.29</td>
<td>1.16</td>
<td>4</td>
<td>3.59</td>
</tr>
<tr>
<td>Time score (s)</td>
<td>15</td>
<td>17.6±8.3 (10.9–41)</td>
<td>17.5±8.3 (9.6–39)</td>
<td>−0.1</td>
<td>0.82</td>
<td>0.94</td>
<td>.88–.97</td>
<td>5.73</td>
<td>9.08</td>
<td>26</td>
<td>15.87</td>
</tr>
<tr>
<td>Anxiety score (VAS 0–10×30 items)</td>
<td>29</td>
<td>14.6±21.5 (0–66)</td>
<td>8.7±16.4 (0–71)</td>
<td>−5.9</td>
<td>0.01</td>
<td>0.91</td>
<td>.80–.95</td>
<td>5.75</td>
<td>7.93</td>
<td>68</td>
<td>15.93</td>
</tr>
<tr>
<td>Wheelie test</td>
<td>Ability score (scale 0–8)</td>
<td>29</td>
<td>5.2±2.3 (0–8)</td>
<td>5.5±2.2 (0–8)</td>
<td>0.3</td>
<td>0.17</td>
<td>0.96</td>
<td>.91–.98</td>
<td>0.44</td>
<td>0.62</td>
<td>12</td>
</tr>
<tr>
<td>Quality score (scale 0–40)</td>
<td>20</td>
<td>17±11.9 (0–35)</td>
<td>16.5±11.5 (0–37)</td>
<td>−0.5</td>
<td>0.15</td>
<td>0.99</td>
<td>.99–.99</td>
<td>1.15</td>
<td>1.38</td>
<td>8</td>
<td>3.20</td>
</tr>
<tr>
<td>Time score (s)</td>
<td>15</td>
<td>13.2±5.5 (6.6–23.6)</td>
<td>12.6±5.1 (6.9–23.9)</td>
<td>0.7</td>
<td>0.28</td>
<td>0.97</td>
<td>.92–.99</td>
<td>3.63</td>
<td>4.81</td>
<td>9</td>
<td>10.07</td>
</tr>
<tr>
<td>Anxiety score (VAS 0–10×8 items)</td>
<td>29</td>
<td>6.2±10.7 (0–32)</td>
<td>4.7±9 (0–36)</td>
<td>−1.5</td>
<td>0.08</td>
<td>0.94</td>
<td>.87–.97</td>
<td>2.41</td>
<td>3.27</td>
<td>62</td>
<td>6.68</td>
</tr>
</tbody>
</table>

NOTE. Wilcoxon signed-rank test was used to verify the absence of significant differences (systematic errors) between the measures at 2 different times (\( t_1 \) and \( t_2 \)).

Abbreviations: SEM, standard error of measurement; \( t_1 \), test 1; \( t_2 \), test 2 (retest); VAS, visual analog scale.

* Significance level, \( P < .05 \).

† Statistically significant differences.
For the Wheelie test total scores, standard error of measurement values were .44 points for the total ability score and 1.15 points for the total quality score. Time and anxiety total scores showed higher standard error of measurement. ME and CVME values were low for the ability, quality, and time total score and much higher for the anxiety total score. MDC95 values were 1.23 points for the total ability scores and 3.20 points for the total quality score and much higher for the time and anxiety total scores.

**Test-retest reliability based on quality scores per task**

The results of the test-retest reliability for the quality scores, per task, are presented in table 2. Statistically significant differences between t1 and t2 were found only for the quality scores of the “wheelie forward” task, but “ascend sidewalk with a run up 10cm,” “up a slope with a run up,” and “stationary wheelie” were nearly significant. The ICCs associated with the test-retest reliability for the TOWM and Wheelie test quality scores, per task, varied from .88 to .99 and the 95% CIs from .73 to .99, except for “level propulsion forward,” which had a lower ICC (.43).

Examining the per task response stability indexes showed lower stability while performing “level propulsion forward” and “one-handed propulsion” of the TOWM and “one-handed wheelie” of the Wheelie test. MDC95 was higher for the first 2 tasks of the TOWM and Wheelie forward 10 meters of the Wheelie test.

**Intrarater reliability**

The results of the intrarater reliability for the quality scores per task are presented in table 3. No significant differences were found between the first and the second evaluation done by the same rater in the total quality scores of both tests. The ICC associated with the intrarater reliability of both the TOWM and Wheelie test total quality score was .99.

Intrarater TEM for the TOWM quality total score was 1.07 and that for the Wheelie test was 1.06 (see table 3), which corresponds to a relative TEM of 3.7% and 6.3%, respectively.

The analysis of each task separately showed significant differences between the assessments of the “uneven surface” and “accelerate and stop in a wheelie” tasks. For all tasks, ICCs were above .95 except for the “level propulsion forward” task, which had a lower ICC (.49). CI values varied from .88 to 1.0, except for “level propulsion forward” (.27—.80). Per task analysis of the TEM index revealed task’s average TEM of .30 for the TOWM and .38 for the Wheelie test.

**Interrater reliability**

The results of the interrater reliability are presented in table 4.

No significant differences were found between the 2 raters on the basis of total quality scores of both tests. The ICC associated with the interrater reliability of both the TOWM and Wheelie test total quality score was .99.

Interrater TEM of the TOWM quality total score was 1.23 and that for the Wheelie test was 1.04, which corresponds to a relative TEM of 4.3% and 6.1%, respectively.

The analysis of each task separately showed a significant difference between the 2 raters for the “descend 10 cm sidewalk” task. ICCs per task varied from .89 to .99 in both tests, except for “level propulsion forward” (.44). Per task analysis of the TEM index revealed task’s average TEM of .44 for the TOWM and .39 for the Wheelie test.

<table>
<thead>
<tr>
<th>Task</th>
<th>Mean ± SD</th>
<th>Significance*</th>
<th>ICC</th>
<th>95% CI</th>
<th>SEM</th>
<th>ME</th>
<th>CVME (%)</th>
<th>MDC95</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOWM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level propulsion forward 4 × 4</td>
<td>4.3 ± 0.7</td>
<td>0.56</td>
<td>0.43</td>
<td>−.43 to .77</td>
<td>0.48</td>
<td>0.53</td>
<td>12</td>
<td>1.34</td>
</tr>
<tr>
<td>One-handed propulsion (10m)</td>
<td>2.8 ± 1.7</td>
<td>0.41</td>
<td>0.88</td>
<td>.73 to .95</td>
<td>0.57</td>
<td>0.45</td>
<td>17</td>
<td>1.57</td>
</tr>
<tr>
<td>Ascend sidewalk 10 cm</td>
<td>0.8 ± 0.8</td>
<td>1.00</td>
<td>1.00</td>
<td>0.73 to .95</td>
<td>0.30</td>
<td>0.29</td>
<td>10</td>
<td>0.67</td>
</tr>
<tr>
<td>Descend sidewalk 10 cm</td>
<td>3.6 ± 1.7</td>
<td>0.41</td>
<td>0.97</td>
<td>.93 to .98</td>
<td>0.29</td>
<td>0.39</td>
<td>11</td>
<td>0.80</td>
</tr>
<tr>
<td>Ascend sidewalk run up 10 cm</td>
<td>2.8 ± 2.4</td>
<td>0.06</td>
<td>0.99</td>
<td>.98 to .99</td>
<td>0.24</td>
<td>0.29</td>
<td>10</td>
<td>0.67</td>
</tr>
<tr>
<td>Up a slope 15%</td>
<td>3.8 ± 1.6</td>
<td>0.18</td>
<td>0.97</td>
<td>.93 to .99</td>
<td>0.26</td>
<td>0.29</td>
<td>7</td>
<td>0.73</td>
</tr>
<tr>
<td>Down a slope 15%</td>
<td>3.5 ± 1.6</td>
<td>0.32</td>
<td>0.99</td>
<td>.98 to .99</td>
<td>0.16</td>
<td>0.16</td>
<td>4</td>
<td>0.43</td>
</tr>
<tr>
<td>Down a slope, stop in wheelie</td>
<td>2.1 ± 2.4</td>
<td>1.00</td>
<td>1.00</td>
<td>0.73 to .95</td>
<td>0.29</td>
<td>0.39</td>
<td>11</td>
<td>0.80</td>
</tr>
<tr>
<td>Up a slope with a run up</td>
<td>1.6 ± 2.1</td>
<td>0.08</td>
<td>0.99</td>
<td>.97 to .99</td>
<td>0.2</td>
<td>0.26</td>
<td>16</td>
<td>0.55</td>
</tr>
<tr>
<td>Chair transfer stable</td>
<td>3.1 ± 1.8</td>
<td>0.10</td>
<td>0.97</td>
<td>.94 to .99</td>
<td>0.30</td>
<td>0.37</td>
<td>12</td>
<td>0.84</td>
</tr>
<tr>
<td>Wheelie test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stationary wheelie</td>
<td>3.4 ± 1.3</td>
<td>0.08</td>
<td>0.98</td>
<td>.95 to .99</td>
<td>0.19</td>
<td>0.40</td>
<td>12</td>
<td>0.52</td>
</tr>
<tr>
<td>One-handed wheelie</td>
<td>1.2 ± 1.8</td>
<td>0.18</td>
<td>0.98</td>
<td>.95 to .99</td>
<td>0.27</td>
<td>0.53</td>
<td>40</td>
<td>0.74</td>
</tr>
<tr>
<td>Wheelie forward 10m</td>
<td>2.9 ± 1.9</td>
<td>0.02†</td>
<td>0.94</td>
<td>.86 to .97</td>
<td>0.46</td>
<td>0.57</td>
<td>20</td>
<td>1.27</td>
</tr>
<tr>
<td>Wheelie backward 10m</td>
<td>1.9 ± 1.5</td>
<td>0.56</td>
<td>0.98</td>
<td>.96 to .99</td>
<td>0.22</td>
<td>0.36</td>
<td>19</td>
<td>0.60</td>
</tr>
<tr>
<td>Circle forward</td>
<td>2.9 ± 2.0</td>
<td>1.00</td>
<td>0.98</td>
<td>.95 to .99</td>
<td>0.29</td>
<td>0.45</td>
<td>15</td>
<td>0.81</td>
</tr>
<tr>
<td>Uneven surface</td>
<td>1.5 ± 1.7</td>
<td>0.10</td>
<td>0.96</td>
<td>.88 to .98</td>
<td>0.32</td>
<td>0.42</td>
<td>31</td>
<td>0.87</td>
</tr>
<tr>
<td>Accelerate and stop in wheelie</td>
<td>1.7 ± 2.0</td>
<td>0.32</td>
<td>0.98</td>
<td>.96 to .99</td>
<td>0.26</td>
<td>0.42</td>
<td>28</td>
<td>0.71</td>
</tr>
<tr>
<td>Backward over curb</td>
<td>1.4 ± 1.8</td>
<td>0.56</td>
<td>0.98</td>
<td>.96 to .99</td>
<td>0.26</td>
<td>0.39</td>
<td>27</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Abbreviations: SEM, standard error of measurement; t1, test 1; t2, test 2 (retest).
* Significance level, P < .05.
† Statistically significant differences.
Discussion

The study’s objectives to determine the reliability, response stability, and MDC of the TOWM and the Wheelie test were accomplished.

Test-retest reliability based on total scores

According to standards suggested by Eliasziw et al., the test-retest reliability ICC values for the total ability, time, and quality scores of both tests were excellent. It indicates that both tests were able to measure WM with consistency and no learning or training effects were demonstrated. The only significant difference between t1 and t2 was confirmed for the anxiety scale, which had a significant difference (systematic bias) between t1 and t2.

The MDC95 values found in this study indicate that in a future intervention based on repeated measurements of the TOWM and the Wheelie test, the total ability score of both tests will need to exceed about 1.5 points (on a scale range 0–40) to be confident (in 95%) that the difference is due to a real WM ability change and not due to a measurement error. Because MDC95 values for WM test-retest assessment were not reported in previous studies, it is recommended to establish MDC95 values for the TOWM and the Wheelie test for specific SCI characteristics (eg, for people with paraplegia, people with tetraplegia, complete, incomplete, upon discharge from rehabilitation, experienced wheelchair users). ICC values found in the current study were compared with values reported in 3 previously published WM tests. The test-retest ICC values for the TOWM and the Wheelie test were similar to those reported for the overall score of the Obstacle Course Assessment of Wheelchair Users Performance test and for the Wheelchair Physical Functional Performance test and higher than the test-retest ICCs reported for the Wheelchair Skill Test (version 2.4).

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Intrarater reliability of the quality scores (n=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Mean ± SD</td>
</tr>
<tr>
<td>TOWM</td>
<td></td>
</tr>
<tr>
<td>Level propulsion forward 4×4</td>
<td>4.3±0.7</td>
</tr>
<tr>
<td>One-handed propulsion (10m)</td>
<td>2.8±1.7</td>
</tr>
<tr>
<td>Ascend sidewalk 10cm</td>
<td>0.8±1.8</td>
</tr>
<tr>
<td>Descend sidewalk 10cm</td>
<td>3.6±1.7</td>
</tr>
<tr>
<td>Ascend sidewalk run up 10cm</td>
<td>2.8±2.4</td>
</tr>
<tr>
<td>Up a slope 15%</td>
<td>3.8±1.6</td>
</tr>
<tr>
<td>Down a slope 15%</td>
<td>3.5±1.6</td>
</tr>
<tr>
<td>Up a slope with a run up</td>
<td>1.6±2.1</td>
</tr>
<tr>
<td>Down a slope, stop in wheelchair</td>
<td>2.1±2.4</td>
</tr>
<tr>
<td>Chair transfer stable</td>
<td>3.1±1.8</td>
</tr>
<tr>
<td>Total quality score TOWM</td>
<td>28.7±13.5</td>
</tr>
<tr>
<td>Wheelie test</td>
<td></td>
</tr>
<tr>
<td>Stationary wheelie</td>
<td>3.4±1.3</td>
</tr>
<tr>
<td>One-handed wheelie</td>
<td>1.2±1.8</td>
</tr>
<tr>
<td>Wheelie forward 10m</td>
<td>2.9±1.9</td>
</tr>
<tr>
<td>Wheelie backward 10m</td>
<td>1.9±1.5</td>
</tr>
<tr>
<td>Circle forward</td>
<td>2.9±2.1</td>
</tr>
<tr>
<td>Uneven surface</td>
<td>1.5±1.8</td>
</tr>
<tr>
<td>Accelerate and stop in wheelchair</td>
<td>1.7±2.0</td>
</tr>
<tr>
<td>Backward over curb</td>
<td>1.4±1.8</td>
</tr>
<tr>
<td>Total quality score Wheelie test</td>
<td>17.0±11.9</td>
</tr>
</tbody>
</table>

Abbreviations: t1 (1st), test occasion 1, first assessment; t1 (2nd), test occasion 1, second assessment done by the same rater 1wk later.
* Significance level, P<.05.
1 Statistically significant differences.
None of the previous WM tests published in the international literature included reliability assessment on measurement error parameters such as response stability indexes or MDC\textsuperscript{17}; therefore, the parameters reported in this study could not be compared with previous findings.

**Test-retest reliability based on quality scores per task**

The quality assessment per task showed that “level propulsion forward” had low ICC values. This can be explained by the low variance in the scores, because all participants obtained a high score on this particular task.

The Wheelchair Skills Test (version 2.4) showed a success rate of 100% for level propulsion forward,\textsuperscript{7} from which it can be concluded that this task is not challenging enough. However, level propulsion forward is the task most frequently tested in the available tests evaluating wheelchair skill performance\textsuperscript{17} and was graded “an extremely essential WM skill” in an international survey among wheelchair users.\textsuperscript{19}

The current study observation may suggest omitting this task from the TOWM because of a ceiling effect. However, the sample included only postrehabilitation wheelchair users; therefore, the decision to omit this task should be made only after testing the TOWM with participants during their early stage of rehabilitation.

The significant differences between t1 and t2 in the quality scores of the “wheelie forward” task may suggest a learning effect for this particular skill. To overcome this learning effect, it is recommended to have a habituation session before t1.

Response stability analysis per task showed that “one-handed propulsion” of the TOWM and “one-handed wheelie” of the Wheelie test had the highest response variability, pointing out that it was difficult for the participants to reproduce the same results in the repeated experiment. A possible explanation may be that one-handed tasks were not familiar to the participants, as they are less required in daily life situations, leading to performance variability from trial to trial.

MDC\textsubscript{95} analysis per task indicated that for the first 2 tasks of the TOWM (level propulsion and one-handed propulsion), and for wheelie forward on a 10-meter line, at least 2 points change between t1 and t2 is needed (quality scale range per task 0–5) to claim a real performance change for these tasks. It should be noted that a common characteristic of these 3 tasks is that the camera-shooting angle is in constant change (following the participant) while in the other task the camera is fixed. This may be an explanation for a higher measurement error while assessing the video of these tasks. A field test “on-spot” assessment analysis is recommended to verify whether the inconsistency is indeed due to the video analysis difficulties or whether the task criteria are not clear enough and need refining.

**Intrarater reliability**

The intrarater ICC of the total quality scores of .99 for both the TOWM and the Wheelie test is a higher value compared with the intrarater reliability reported in 3 previous WM tests.\textsuperscript{7,9,11} Relative TEMs of 3.7% for the intrarater assessment of the TOWM and 6.3% for the Wheelie test were found.
There are no acceptable ranges for TEM for either interrater or intrarater WM skill assessment (unlike other domains, such as anthropometry, where the relative TEM for a beginner anthropometrist for intrarater skin folds measurement is known to be 7.5% and 5% for a skillful anthropometrist). Future studies should aim to establish relative TEM acceptable ranges in WM assessment.

Interrater analysis per task showed that ICC values for all tasks, except for “level propulsion forward,” were excellent. The low ICC value for “level propulsion forward” (.50) was a result of the small variance between participants’ scores (variability among subjects’ scores must be large to demonstrate reliability).

The significant differences that were found between the 2 quality measurements of the “uneven surface” and “accelerate and stop in a wheelie” might indicate that the skill maturity criteria in these cases were not as clearly formulated as for the other skills. “The major protection against tester bias (ie, the observer can be influenced by his memory of the first score), is to develop grading criteria that are as objective as possible and to train the tester in the use of the instrument.” On that account, it is advisable to review the quality criteria of these specific skills and to make adjustments as necessary.

**Interrater reliability**

Similar to the intrarater reliability, the interrater reliability value for the total quality scores, and for all tasks except “level propulsion forward,” was excellent. The interrater ICC values of total quality scores (.99) of both tests are higher than the ICC values reported for the overall scores of 4 previously published WM tests.

The significant difference between the 2 raters that was found for “descend 10cm sidewalk” indicates the need to review the maturity criteria to maximize the reliability when qualitatively assessing this skill.

**Study limitations and future work**

Only 29 wheelchair users participated in the study in total. This is a rather small sample, but it is comparable with sample sizes used in other studies assessing the reliability of new wheelchair skills tests. Despite the small sample, representation of different lesion levels and time passed since injury was a goal; therefore, it represents a wide range of SCI levels and it includes participants shortly after clinical rehabilitation next to participants who have been using a wheelchair for a long time. Consequently, the result may be quite representative to the wider population of manual wheelchair users with SCI. The 1-week interval between the 2 test occasions was set to avoid learning or training effects during the testing period. This seems to have been quite adequate, although the anxiety scores showed a “learning” trend.

As suggested by Routhier et al., the procedure of quality assessment by video, which was used in this study, might be a limitation, because the examiners cannot see everything done by the wheelchair user as a result of a 2-dimensional film and restricted view produced by the video, compared with human observation of real-life conditions and actions. In addition, Kirby et al. agreed that scoring from a silent videotape is more challenging than doing so in person and that it might underestimate the true reliability. It is suggested that in future research, the quality assessment will be carried out on field, to allow comparison between both evaluation methods and to reassess the reliability of the quality scale.

The interrater reliability in this study was based on 2 raters’ assessments and for the quality criteria only. It is recommended that in a future study, interrater reliability will be assessed by 3 raters and will also be tested for the ability scores.

The time scale was found as less sensitive, and the anxiety scale showed a learning effect; it is recommended to reassess WM with the TOWM and the Wheelie test after a habituation session, and to perceive whether response stability findings will change. If instability recurred, it would be suggested not to include these scales in future studies.

In the intrarater and interrater analysis, TEMs demonstrated more errors when assessing the Wheelie test tasks compared to the TOWM. This reflects not only the need for refining the quality criteria of the Wheelie test tasks but also the necessity of technical training of evaluators on the quality assessment of the Wheelie test to minimize the variability verified.

**Conclusions**

Results suggest an excellent reliability of the total ability and quality scores of the TOWM and the Wheelie test when assessing WM of manual wheelchair users with SCI. It may be considered omitting the “level propulsion” task from the TOWM if a “ceiling effect” will also be detected among participants during the early rehabilitation phase. The quality criteria of “descend 10cm sidewalk” and “one-handed propulsion” (from the TOWM) and “wheelie forward,” “uneven surface,” and “accelerate and stop in a wheelie” (from the Wheelie test) need to be reviewed and possibly adjusted. Reducing the number of tasks may also be considered.

**Suppliers**

a. Canon Basler 100Hz camera; Canon, Inc, Tokyo, Japan; Local supplier: Canon België NV/SA, Berkenlaan 3 831 Diegem (Machelen).

b. IBM SPSS, Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.

c. Microsoft Excel 2010, Microsoft Corporation, Redmond WA 98052.

**Keywords**

Rehabilitation; Reliability; Spinal cord injuries; Wheelchair skills; Wheeled mobility

**Corresponding author**

Osnat Fliess-Douer, PhD, Ha’Dagan St, 22nd Binyamina, Israel 30500. E-mail address: douer@netvision.net.il.

**Acknowledgments**

We thank Joeri Verellen, PhD, Maarten Abeel, MA, Elien Baeten, MA, Melina Boeckmans, MA, and Thomas Witzduckand, MA for their assistance and contribution for the development and assessment of the new instruments.
References


www.archives-pmr.org