The Ziegler Test is Reliable and Valid for Measuring Freezing of Gait in People With Parkinson Disease

Ziegler Test: Reliability and Validity Study

Measurement

Original Research

Lina Goh, BAppSc\textsuperscript{1}; Serene S. Paul, PhD\textsuperscript{1}; Colleen G. Canning, PhD\textsuperscript{1}; Kaylena A. Ehgoetz Martens\textsuperscript{2}; Jooeun Song, PhD\textsuperscript{1}; Stephanie L. Campoy, BAppSc\textsuperscript{3}; Natalie E. Allen, PhD\textsuperscript{1}\textsuperscript{*}

The Author(s) 2022. Published by Oxford University Press on behalf of the American Physical Therapy Association.
1. School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, New South Wales, Australia.

2. Department of Kinesiology, Faculty of Health, The University of Waterloo, Ontario, Canada

3. No affiliations.

CORRESPONDING AUTHOR*

Natalie E. Allen*

All correspondence to natalie.allen@sydney.edu.au

KEYWORDS: Freezing of Gait; Parkinson Disease; Reliability; Validity; Ziegler Test
Abstract

Objective. The purpose of this study was to determine interrater and test-retest reliability of the Ziegler test to measure freezing of gait (FOG) severity in people with Parkinson disease. Secondary aims were to evaluate test validity and explore Ziegler test duration as a proxy FOG severity measure.

Methods. Physical therapists watched 36 videos of people with Parkinson disease and FOG perform the Ziegler test and rated FOG severity using the rating scale in real time. Two researchers rated 12 additional videos and repeated the ratings at least 1 week later. Interrater and test-retest reliability were calculated using intraclass correlation coefficients (ICCs). Bland-Altman plots were used to visualize agreement between the researchers for test-retest reliability. Correlations between the Ziegler scores, Ziegler test duration, and percentage of time frozen (based on video annotations) were determined using Pearson r.

Results. Twenty-four physical therapists participated. Overall, the Ziegler test showed good interrater (ICC2,1 = 0.80; 95% CI = 0.65 to 0.92) and excellent test-retest (ICC3,1 = 0.91; 95% CI = 0.82 to 0.96) reliability when used to measure FOG. It was also a valid measure, with a high correlation (r = 0.72) between the scores and percentage of time frozen. Ziegler test duration was moderately correlated (r = 0.67) to percentage of time frozen and may be considered a proxy FOG severity measure.

Conclusion. The Ziegler test is a reliable and valid tool to measure FOG when used by physical therapists in real time. Ziegler test duration may be used as a proxy for measuring FOG severity.
Impact. Despite FOG being a significant contributor to falls and poor mobility in people with Parkinson disease, current tools to assess FOG are either not suitably responsive or too resource-intensive for use in clinical settings. The Ziegler test is a reliable and valid measure of FOG, suitable for clinical use, and may be used by physical therapists regardless of their level of clinical experience.

[H1]Introduction

Approximately 50% of people with Parkinson disease will experience freezing of gait (FOG), which is defined as brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk.¹ The incidence of FOG increases as disease progresses,² with people with FOG likely to experience reduced mobility, increased falls³,⁴ and reduced quality of life.⁵

To determine FOG severity, and consequently identify suitable interventions to reduce freezing, clinicians and researchers need reliable and valid measures. Tools commonly used to assess FOG severity remain largely inadequate. For instance, the New Freezing of Gait Questionnaire (NFOGQ)⁶ has a minimal detectable change of 9.95 points on a 28 points scale and may not be suitably responsive to reflect a clinically meaningful change.⁷ The use of video annotations to measure FOG by calculating the percentage of time spent frozen while people with Parkinson disease perform FOG-provoking tasks (ie, total time spent frozen during FOG episodes divided by total time spent performing the tasks),⁸ or the use of wearable technologies and machine learning to detect FOG⁹ have potential to accurately capture FOG. However, both these methods require information to be recorded and then analyzed, are time and resource-intensive, and hence are not yet feasible in a clinical setting.
The Ziegler test may be used clinically to assess FOG.\textsuperscript{10} It involves walking a course with 4 situations that commonly trigger freezing (start, clockwise turn, counterclockwise turn, and walking through a doorway) and is repeated under 2 dual tasks conditions (carrying a tray with a cup of water and carrying a tray with a cup of water plus counting backwards). The performance of each situation under each condition is rated using a 4-point rating scale which is based on the type of freezing observed. It is easy to set up, fast to use, and suitable for use in clinics and at home. Although there was high interrater and test-retest reliability between the researchers who designed and tested the Ziegler test,\textsuperscript{10} the reliability and validity of the Ziegler test when used by physical therapists are unknown. There was also recent evidence that suggests the time taken to complete the Ziegler test was an independent predictor of FOG severity\textsuperscript{11} and may be used as a proxy FOG severity measure. However, the utility of the Ziegler test duration as a proxy FOG severity measure when used by physical therapists is unknown.

The primary aim of this study was to determine the interrater and test-retest reliability of the Ziegler test when used by physical therapists in real time to assess the severity of FOG in people with Parkinson disease. The secondary aims were to evaluate the validity of the Ziegler test by determining correlations between the outcomes from the rating scale and the percentage of time spent frozen measured using video annotations; to explore whether the Ziegler test duration is a valid proxy measure of FOG severity; to evaluate test usability; and to evaluate any differences in reliability based on level of clinical experience.
[H1] Methods

[H2] Participants

Physical therapists were recruited from public and private health care sectors and academic institutions in New South Wales and the Australian Capital Territory, Australia. Physical therapists were eligible to participate in this study if they held general registration to practice in Australia and had a minimum of 6 months of clinical experience.

[H2] Design

Ten people (9 males, 1 female) with Parkinson disease and FOG completed the Ziegler test as part of a previous study. They had a mean age of 70.6 years (SD = 7.7), moderate to severe Parkinson disease with a mean Movement Disorder Society - Unified Parkinson’s Disease Rating Scale Section III (motor) score of 37.3 (SD = 13.3), and moderate to severe FOG with a mean NFOGQ score of 20.0 (SD = 4.3).

For the Ziegler test, they were asked to stand up from a chair positioned 3.4 meters in front of a closed door, walk 1 meter forward to a square outlined with tape on the ground (40 x 40 cm) in which they completed two 360-degree turns (clockwise then counterclockwise), walk a further 2 meters forward towards the door, open the door and walk through the doorway. Their performance at each test situation (ie, start, clockwise turn, counterclockwise turn, and doorway) was rated: 0 points when no festination and no FOG was detected; 1 point when festination or any hastening steps (“shuffling”) was observed but the person was able to continue moving forwards; 2 points when FOG (trembling-in-place or total akinesia) occurred in which the person was unable...
to move forwards but could overcome that freezing episode themselves; and 3 points for any abortion of the task or any need of interference by the examiner (e.g., acoustic or visual 1-off cue to overcome the freezing episode). People with Parkinson disease completed this test under 3 conditions in this order: (1) no additional task, (2) with an additional manual task (i.e., holding a tray with a full cup of water), and (3) with an additional cognitive and manual task (i.e., counting backwards by 7 seconds from 100 and holding a tray with a full cup of water). The total score ranged from 0 to 36. The performance of each test condition was captured on video, with a total of 3 videos generated per Ziegler test.

People with Parkinson disease performed the Ziegler test during their “on” phase (i.e., when their Parkinson medications were working optimally) in the clinic and “off” phase (i.e., after their Parkinson medication were withdrawn for at least 12 hours overnight) in their homes. Assessment at home during the “off” phase was for pragmatic reasons to ensure patient safety. A subset of videos was chosen across “on” and “off” phases to include a broad representation of FOG types and severity, as more severe forms of FOG such as trembling-in-place and akinesia are more likely to be triggered in the “off” phase. Videos representing a range of Ziegler scores reflecting different FOG types and severity were chosen to maximize the robustness of the calculations for intraclass correlations. The subset of videos was chosen by 2 researchers (L.G. and S.S.P.) not involved in the ratings used in this study.

[H2] Outcome
Interrater reliability was determined based on our sample of physical therapists who were recruited into the study, plus 2 physical therapists from the research team. All physical therapists viewed the videos remotely on a secure web platform (REDCap) and rated 12 sets of 3 videos (ie, a total of 36 videos), with the order of each set of videos randomized. Instructions on how to rate the performance were provided as outlined in Ziegler et al (2010)\textsuperscript{10} (Suppl. Material 1). No additional instructions were provided to the physical therapists, except to watch the videos only once at normal speed to best mimic use of the rating scale in a clinical setting in real time. For each test condition (ie, with or without additional tasks), physical therapists were asked to simultaneously determine the Ziegler test duration (ie, time taken to complete the Ziegler test from start to finish) using a stopwatch. At the end of the ratings, the physical therapists completed a closed-end questions survey to explore usability of the rating scale. At the end of the closed-end questions, they were given an opportunity to provide additional feedback via free text.

Two physical therapists from the research team, 1 Parkinson disease expert (N.E.A.) and 1 novice (S.L.C.), rated an additional 4 sets of 3 videos (ie, a total of 48 videos). To determine the test-retest reliability, both researchers re-rated the 48 videos after at least 1 week. Two other members of the research team determined the percentage of time spent frozen (JS and KAEM) via video annotations using a pre-determined protocol.\textsuperscript{12}

\textbf{Statistical Analyses}

The interrater and test-retest reliability of the Ziegler rating scale were calculated using intraclass correlation coefficients (ICC\textsubscript{2,1} and ICC\textsubscript{3,1}, respectively). An ICC > 0.90 was considered excellent;
Overall interrater and test-retest reliability were calculated, as well as reliability for individual test condition (ie, with or without additional tasks) and individual test situations (ie, starts, turns, doorways). Data from the clockwise and counterclockwise turns were pooled as “turns” due to their similar movement patterns. Bland-Altman plots were used to visualize agreement between the 2 researchers for test-retest reliability. Correlations between outcomes of the Ziegler rating scale, time taken to complete the test, and percentage of time frozen were determined using Pearson correlation coefficients ($r$). An $r > 0.90$ was considered very high; $0.70–0.90$, high; $0.50–0.70$, moderate; $0.30–0.50$, low; and $< 0.30$ negligible. Closed responses from the survey were analyzed using descriptive statistics, and open responses were analyzed thematically. Statistical analyses of the data were conducted using the SPSS Statistics for Windows, version 26.0 (IBM Corporation, Armonk, New York, USA).

[H2] Role of the Funding Source

The funding bodies played no role in the design, conduct, or reporting of this study.

[H1] Results

[H2] Participants

Twenty-four physical therapists (22 recruited, 2 researchers) participated in this study. They practiced in diverse healthcare settings, from public ($n = 13$) and private ($n = 9$) clinics to academic institutions ($n = 2$), across acute ($n = 7$), rehabilitation ($n=6$) and community care ($n = 4$). Seven physical therapists had between 0.5 and 2 years of postgraduate clinical experience, 9 had
between 3 and 10 years, and 8 had more than 10 years. Ten physical therapists reported seeing between 0 and 10 people with Parkinson disease in the past year, 11 reported seeing between 11 and 30, and 3 reported seeing more than 30. All physical therapists, except the Parkinson disease expert (N.E.A.), were unfamiliar with the Ziegler test. The 2 researchers were not involved in conducting the data analyses.

[H2] Interrater Reliability

The total score from the Ziegler test showed good interrater reliability (ICC$_{2,1}$ = 0.80; 95% CI = 0.65–0.92). There was also good interrater reliability for individual test conditions with no additional tasks (ICC$_{2,1}$ = 0.76; 95% CI = 0.59–0.90), and an additional manual task (ICC$_{2,1}$ = 0.76; 95% CI = 0.60–0.90). However, there was only moderate interrater reliability for the test condition with an additional cognitive and manual task (ICC$_{2,1}$ = 0.69; 95% CI = 0.51–0.87), and when the start, turns and doorway test situations were rated in isolation across the 3 test conditions (start: ICC$_{2,1}$ = 0.51, 95% CI = 0.40–0.65; clockwise and counterclockwise turns: ICC$_{2,1}$ = 0.64, 95% CI = 0.56–0.72; doorway: ICC$_{2,1}$ = 0.61, 95% CI = 0.50–0.73) (Tab. 1).

Physical therapists with high and low levels of overall or Parkinson disease-specific clinical experience showed moderate to good interrater reliability (Tab. 2).

[H2] Test-Retest Reliability

Overall, the total score from the Ziegler test showed excellent test-retest reliability (ICC$_{3,1}$ = 0.91; 95% CI = 0.82–0.96). There was also excellent test-retest reliability for the test condition with an
additional cognitive and manual task (ICC₃,₁ = 0.92; 95% CI = 0.85–0.96), and good test-retest reliability for test conditions with no additional task (ICC₃,₁ = 0.88; 95% CI = 0.76–0.94) and with an additional manual task (ICC₃,₁ = 0.88; 95% CI = 0.77–0.94). When test situations were rated in isolation, the test-retest reliability was good (start: ICC₃,₁ = 0.85, 95% CI = 0.77–0.90; clockwise and counterclockwise turns: ICC₃,₁ = 0.81, 95% CI = 0.75–0.85; doorway: ICC₃,₁ = 0.86, 95% CI = 0.80–0.91) (Tab. 3). The Bland-Altman plots did not show any systematic bias between the Ziegler scores at the 2 timepoints (Fig. 1).

Both the expert and the novice physical therapists showed good to excellent test-retest reliability when rating the overall Ziegler test or individual test conditions. Both therapists also showed good reliability when rating individual test situations, except for the novice physical therapist who showed moderate test-retest reliability when rating the turns (Tab. 3).

[H2] Validity of the Ziegler Test

There was a high correlation between scores from the overall Ziegler test and percentage of time spent frozen ($r = 0.72$), and moderate to high correlations between scores from individual test conditions and their corresponding percentages of time spent frozen (no additional task: $r = 0.79$; additional manual task: $r = 0.64$; additional cognitive and manual task: $r = 0.64$) (Tab. 4).

When considering the time taken to complete the Ziegler test and percentage of time spent frozen, there was a moderate correlation between the overall Ziegler test duration and total percentage of time spent frozen ($r = 0.67$) and moderate to high correlations between duration of
individual test conditions and their corresponding percentages of time spent frozen (no additional task: \( r = 0.77 \); additional manual task: \( r = 0.66 \); additional cognitive and manual task: \( r = 0.57 \) (Tab. 4).

[H2] Survey Responses

Of the 24 physical therapists, 21 (88%) strongly agreed or agreed that the rating scale was easy to use, 2 (8%) were neutral, and 1 (4%) disagreed. Twenty-one (88%) physical therapists strongly agreed or agreed that the instructions for using the rating scale were clear and 3 (13%) were neutral. Sixteen (67%) physical therapists strongly agreed or agreed it was easy to distinguish between the different scores of the rating scale, 5 (21%) were neutral, and 3 (13%) disagreed. Fifteen (63%) physical therapists strongly agreed, or agreed, the scores of the rating scale reflected the severity of FOG observed in the videos, 3 (13%) were neutral, and 6 (25%) disagreed. Of the 6 physical therapists who disagreed, 5 (83%) felt that the score underestimated freezing severity. Fifteen physical therapists (63%) reported they will use the Ziegler rating scale in their clinical practice (Fig. 2).

There was variability in how the physical therapists interpreted the start and end of test situations. Fifteen (63%) physical therapists rated any freezing that occurred during “start” from the start of walking up to the moment before the clockwise turn occurred, while 9 (38%) rated freezing occurring at the start of walking only. Fourteen (58%) physical therapists rated any freezing that occurred during “doorway” from the end of the counterclockwise turn up to the
moment the person with Parkinson disease walked through the doorway; 4 (17%) rated while the person was approaching the door, opening the door, and walking through the doorway; 3 (13%) rated when the person was walking through the doorway only; 2 (8%) rated while the person was opening the door and walking through the doorway; and 1 (4%) rated when the person approached the door only.

Fifteen physical therapists provided additional feedback on the Ziegler scale (Suppl. Material 2). A few physical therapists expressed difficulties identifying the presence or type of FOG. For example, they reported it was difficult to distinguish if the person with Parkinson disease was walking/turning with small steps or experiencing festination, or if a stoppage in movement was due to dual tasking or FOG. Several therapists also reported difficulties interpreting test instructions as they felt it was unclear when test situations began and ended (eg, start and doorway) and if prompts by the examiners to perform the task was considered interference. Two therapists felt the rating scale did not reflect the severity of FOG observed as it was possible to have significant differences in FOG severity within the 1 score level. For instance, a single festination episode of a short duration and multiple festination episodes of long durations were both scored as 1-point. One therapist suggested using a combination of Ziegler scores and test duration to get a better indication of FOG severity.

[H1]Discussion

This study aimed to explore the clinical utility of the Ziegler test to assess FOG severity in people with Parkinson disease. Our results suggest the Ziegler test had good interrater and excellent test-
retest reliability when used by physical therapists in real time. The Ziegler test was also found to be a valid measure, with a high correlation to the percentage of time spent frozen, which is currently considered “gold standard” for measuring FOG.15

The Ziegler test showed good interrater reliability when used in its entirety, and moderate to good reliability when test conditions were rated individually. Further exploration to examine the effect of clinical experience on interrater reliability showed that physical therapists with low or high levels of overall or Parkinson disease-specific clinical experience demonstrated moderate to good interrater reliability. The Ziegler test also showed excellent test-retest reliability for the overall score and good to excellent reliability for the individual test conditions, indicating a high level of reproducibility over time. Both the expert and the novice physical therapists demonstrated good to excellent test-retest reliability. Furthermore, the Ziegler test was shown to be a valid tool to measure FOG, with a high correlation between the overall Ziegler score and the percentage of time spent frozen. Correlations remain moderate to high when scores from individual test conditions were compared to corresponding percentages of time spent frozen, and when the Ziegler test duration was compared to percentage of time spent frozen, indicating time taken to complete the Ziegler FOG provoking course may be a valid proxy measure of FOG.

These results suggest the Ziegler test is a reliable and valid measure of FOG, and its use should be considered in clinical practice. The test provides several additional benefits over current FOG measures. It provides an empirical measure of FOG and presents further insights to FOG severity not captured by self-report questionnaires.15 The Ziegler test is cheap, easy to set up, and fast to implement, and thus highly accessible to time and resource-poor clinicians. The use of wearable
technologies and automated algorithms show potential in capturing FOG throughout the day and in the home, but it may be years before it can be successfully implemented in clinical practice due to the need to optimize testing procedures and fine-tune algorithms to accurately detect FOG, as well as the associated high cost, lack of resources and supporting infrastructure.\textsuperscript{9,16-18} In contrast, the Ziegler test can be implemented in current clinical practice and offers an immediate solution to gaps in objective FOG measurements.

Since it is often difficult to trigger FOG in the clinic,\textsuperscript{15} the Ziegler test situations and conditions provide a useful framework for triggering FOG. To implement the Ziegler test in clinical practice, physical therapists have a number of options. They may choose to complete the entire test, or for those who are time-poor, it may be possible to reliably measure FOG severity by only completing 1 or 2 of the test conditions (ie, with or without additional tasks). Physical therapists may choose to assess people with mild disease using test conditions with additional tasks which are necessary to trigger FOG and assess people with moderate to severe disease using the test condition without an additional task to avoid placing people with Parkinson disease under undue stress. In addition to the numerical rating, observations made by the physical therapist during test performance may inform potential intervention strategies.\textsuperscript{12}

All physical therapists can consider using the Ziegler test even if they have minimal overall or Parkinson disease-specific clinical experience, as their ratings agree with therapists with more experience. A second therapist may be required to ensure safety of the person with Parkinson disease performing the test given the high risks of falls in people with FOG. Our results also support outcomes from a previous study by Herman et al (2020), who reported a moderate
correlation between the Ziegler test duration and Ziegler score.\textsuperscript{11} Therefore, physical therapists who work alone may choose to time the Ziegler test as a proxy FOG severity measure as it may be easier to operate a stopwatch than rate the performance in real time while supervising a person at high risk of falls during the test.

Despite the overall good to excellent reliability and validity, several limitations with using the Ziegler test to assess FOG remain. The Ziegler test is designed to measure FOG by identifying the type of freezing observed, ie, festination, trembling-in-place, or akinesia, with the assumption that trembling-in-place and akinesia are more severe than festination. However, it is unclear if this assumption is valid as individuals experience FOG and its impact differently and different FOG types may not reflect increases in FOG severity. Whether the presence of different types of freezing reflect different freezing severity when measured by number and duration of FOG episodes, and how individuals with FOG perceive the relation between types of freezing and freezing severity, remain to be determined.

While identifying the type of freezing to measure FOG severity may be appropriate as a simple and quick clinical tool, the Ziegler test does not quantify the frequency and duration of freezing, which may provide a more detailed assessment of freezing severity.\textsuperscript{15} Scores from the Ziegler test are likely underestimating freezing severity in people with worse FOG, as individuals experiencing many freezing episodes of long duration were scored similarly to individuals experiencing a single freezing episode of short duration. There are also limitations to using the Ziegler test duration as a proxy measure for FOG severity as increases in time taken to complete the test conditions with additional tasks may be attributed to FOG and/or dual or triple demands, and not FOG alone. This
is reflected in the lower correlations between Ziegler test durations and percentages of time spent frozen for test conditions with additional tasks compared to no additional task.

The Ziegler test situations were not well defined in the instructions provided by the test developers\textsuperscript{10} (Suppl. Material 1), with the physical therapists having varying interpretations. While it was apparent when the “turns” occurred, how physical therapists interpreted “starts” and “doorway” were highly variable. For example, approximately half the physical therapists defined “start” as start of walking only, while the others defined “start” as start of walking up to the start of the clockwise turn. Hence, any FOG that occurred during forward walking just after the start of walking would be scored by some and not by others.\textsuperscript{15} Likewise, approximately half the physical therapists defined “doorway” as the period from the end of the counterclockwise turn to after the person steps across the doorway, while others only rated part of the task, such as approaching the door only, walking through the doorway only, or opening the door and walking through the doorway only. FOG may again be scored variably. This may explain the lower reliability when physical therapists rated “starts”, “turns” and “doorways” movement types, which suggest ratings of Ziegler test situations (ie, starts, turns, doorways) in isolation are currently not recommended.

The reliability and validity of the Ziegler test may be improved by developing a more detailed rating scale that clearly outlines when test situations begin and end, and accounts for frequency and duration of freezing episodes. The reliability between physical therapists when rating “turns” may also be improved by defining what constitutes a FOG episode during turning, as movements during turning do not show the same patterns as straight-line walking tasks.\textsuperscript{19} This is supported by emerging evidence that showed excellent reliability between raters assessing FOG using a set of
established criteria describing FOG episodes triggered by alternating 360-degree turns. Future studies should examine if Ziegler test reliability and validity vary depending on duration of FOG episodes, with recent studies showing poor agreement between raters for very short (<1 s) FOG episodes and better agreement for short (2 to 5 s) and long (>5 s) FOG episodes measured via video analyses. Future studies should also explore the utility of combining the Ziegler score and duration as a FOG measure and investigate the responsiveness of the Ziegler test as its ability to detect changes in FOG over time is unknown.

[H2] Limitations

Due to the global Covid-19 pandemic, physical therapists were asked to rate videos remotely. Therapists were instructed to view videos only once to best mimic clinical practice. Although video files were compressed to make them suitable for online streaming and instructions were given to the physical therapists to load the videos prior to viewing, 10% of videos were viewed more than once due to lagging or other technical issues and 3% of test durations were unable to be measured. The different sites used for testing during “on” and “off” phases testing may also have influenced outcomes, with videos captured at home less likely to be optimized for viewing due to lighting or space constraints. To mitigate any possible differences in raters’ ability to view FOG, the camera in both the clinic and home settings was set up similarly with adequate lighting such that the performance of the person with Parkinson disease performing the Ziegler test was clearly visible for rating in all videos. For some videos, it was hard to determine if there was forward advancement during a FOG episode, and consequently hard to distinguish between scores on the rating scale. The reliability and validity of the Ziegler test may also differ if scored by physical therapists in-person. However, differences in scores are likely to be negligible as prior evidence
showed good agreement between video and in-person ratings of Parkinson disease impairments. The low representation of women with Parkinson disease and FOG in this study may also limit the generalizability of our results in relation to the reliability and validity of the Ziegler test to measure FOG across different genders.

**Conclusion**

The Ziegler test is a reliable and valid tool to measure FOG when used by physical therapists in real time. Scores from individual test conditions and time taken to complete the test may also be used to measure FOG. Further clarification of Ziegler test instructions and a rating scale that accounts for frequency and duration of freezing episodes may improve its clinical utility and suitability to use as an outcome measure.
Author Contributions

Concept/idea/research design: L. Goh, S.S. Paul, C.G. Canning, K.A.E. Martens, N.E. Allen

Writing: L. Goh, N.E. Allen


Project management: L. Goh

Fund procurement: L. Goh

Consultation (including review of manuscript before submitting): L. Goh, S.S. Paul, C.G. Canning, S.L. Campoy, K.A.E. Martens, J. Song

Acknowledgments

The authors wish to thank all physical therapists who participated in this study.

Ethical Approval
The authors received ethical approval from The University of Sydney Human Research Ethics Committee (project number 2021/174). All data have been deidentified.

Funding

Author Lina Goh is supported by an Australian Government Research Training Program Scholarship. The University of Sydney Charles Perkins Centre Active Ageing Seeding Grant funded the original study, which generated data used in this study.

Disclosure

The authors completed the ICJME Form for Disclosure of Potential Conflicts of Interest and reported no conflicts of interest.
References

Table 1. Interrater Reliability (ICC<sub>2,1</sub>) With 95% CI of the Ziegler Test and When Rating Individual Test Conditions and Situations

<table>
<thead>
<tr>
<th>Ziegler Test</th>
<th>n</th>
<th>ICC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>24</td>
<td>0.80</td>
<td>0.65–0.92</td>
</tr>
<tr>
<td><strong>Test Condition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No additional task</td>
<td>24</td>
<td>0.76</td>
<td>0.59–0.90</td>
</tr>
<tr>
<td>Additional manual task</td>
<td>24</td>
<td>0.76</td>
<td>0.60–0.90</td>
</tr>
<tr>
<td>Additional manual and cognitive tasks</td>
<td>24</td>
<td>0.69</td>
<td>0.51–0.87</td>
</tr>
<tr>
<td><strong>Test Situation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Starts</td>
<td>24</td>
<td>0.51</td>
<td>0.40–0.65</td>
</tr>
<tr>
<td>Turns</td>
<td>24</td>
<td>0.64</td>
<td>0.56–0.72</td>
</tr>
<tr>
<td>Doorways</td>
<td>24</td>
<td>0.61</td>
<td>0.50–0.73</td>
</tr>
</tbody>
</table>

Table 2. Interrater Reliability (ICC<sub>2,1</sub>) With 95% CI of the Ziegler Test Based on Physical Therapists’ Clinical Experience<sup>a</sup>

<table>
<thead>
<tr>
<th>Categories of Physical Therapists</th>
<th>Ziegler Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall clinical experience (years)</td>
<td>n</td>
</tr>
<tr>
<td>0.5–2 y</td>
<td>7</td>
</tr>
<tr>
<td>3–10 y</td>
<td>9</td>
</tr>
<tr>
<td>&gt;10 y</td>
<td>8</td>
</tr>
<tr>
<td><strong>Number of people with PD seen in the past year</strong></td>
<td></td>
</tr>
<tr>
<td>0–10</td>
<td>10</td>
</tr>
<tr>
<td>11–30</td>
<td>11</td>
</tr>
<tr>
<td>&gt;30</td>
<td>3</td>
</tr>
</tbody>
</table>

<sup>a</sup>PD = Parkinson disease.

Table 3. Test-Retest Reliability (ICC<sub>3,1</sub>) With 95% CI of the Ziegler Test Rated at Least 1 Week Apart, and When Rating Individual Test Conditions and Situations

<table>
<thead>
<tr>
<th>Ziegler Test</th>
<th>Overall (n = 2)</th>
<th>Expert (n = 1)</th>
<th>Novice (n = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>95% CI</td>
<td>ICC</td>
</tr>
<tr>
<td>Overall</td>
<td>0.91</td>
<td>0.82–0.96</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>Test Condition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No additional task</td>
<td>0.88</td>
<td>0.76–0.94</td>
<td>0.87</td>
</tr>
<tr>
<td>Additional manual task</td>
<td>0.88</td>
<td>0.77–0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Additional manual and cognitive tasks</td>
<td>0.92</td>
<td>0.85–0.96</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Test Situation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 4. Correlations (r) Between the Ziegler Test Scores, Test Duration, and Percentage of Time Spent Frozen

<table>
<thead>
<tr>
<th>Ziegler Test</th>
<th>Percentage of Time Spent Frozen</th>
<th>Overall</th>
<th>No additional task</th>
<th>Additional manual task</th>
<th>Additional manual and cognitive tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.72</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No additional task</td>
<td>-</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Additional manual task</td>
<td>-</td>
<td>-</td>
<td>0.64</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Additional manual and cognitive tasks</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.64</td>
<td></td>
</tr>
</tbody>
</table>

### Ziegler Duration

| Overall           | 0.67                            | -       | -                | -                    | -                                    |
| No additional task| -                               | 0.77    | -                | -                    | -                                    |
| Additional manual task | -                               | -       | 0.66             | -                    | -                                    |
| Additional manual and cognitive tasks | -                               | -       | -                | 0.57                 |                                      |

**Figure Captions**
Figure 1. Bland-Altman plots showing agreement on the Ziegler test rated at 2 timepoints at least 1 week apart. The solid line indicates the mean difference between the 2 scores, and the dashed lines indicate the limits of agreement (1.96 SDs of the mean difference).
Figure 2. Physical therapists’ experience of using the Ziegler test. The solid vertical line represents neutral, that is, the midway point on a 5-point Likert scale that ranged from strongly disagree to strongly agree.