Disability and Rehabilitation

ISSN: 0963-8288 (Print) 1464-5165 (Online) Journal homepage: https://www.tandfonline.com/loi/idre20

The reliability and validity of the Timed Up and Go as a clinical tool in individuals with and without disabilities across a lifespan: a systematic review

Psychometric properties of the Timed Up and Go

Ashley Christopher, Emily Kraft, Hannah Olenick, Riley Kiesling & Antonette Doty

To cite this article: Ashley Christopher, Emily Kraft, Hannah Olenick, Riley Kiesling & Antonette Doty (2019): The reliability and validity of the Timed Up and Go as a clinical tool in individuals with and without disabilities across a lifespan: a systematic review, Disability and Rehabilitation, DOI: 10.1080/09638288.2019.1682066

To link to this article: https://doi.org/10.1080/09638288.2019.1682066

DISABILITY AND REHABILITATION
https://doi.org/10.1080/09638288.2019.1682066

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

The reliability and validity of the Timed Up and Go as a clinical tool in individuals with and without disabilities across a lifespan: a systematic review

Psychometric properties of the Timed Up and Go

Ashley Christopher, Emily Kraft, Hannah Olenick, Riley Kiesling and Antonette Doty

Physical Therapy Department, Walsh University, North Canton, OH, USA

ABSTRACT

Purpose: To summarize the available literature related to reliability and validity of the Timed Up and Go in typical adults and children, and individuals diagnosed with the following pathologies: Huntington's disease, stroke, multiple sclerosis, Parkinson's disease, spinal cord injury, Down syndrome, or cerebral palsy.

Materials and methods: A search was conducted using MeSH terms and keywords through a variety of databases. Data regarding reliability and validity were synthesized.

Results: This review included 77 articles. Results were variable depending on the studied population. The Timed Up and Go showed excellent reliability in typical adults, in individuals with cerebral palsy, in individuals with multiple sclerosis, in individuals with Huntington's disease, individuals with a stroke, and individuals with a spinal cord injury. The TUG demonstrated strong concurrent validity for individuals with stroke and spinal cord injury. Predictive validity data was limited.

Conclusions: Based on the literature assessed, the Timed Up and Go is clinically applicable and reliable across multiple populations. The Timed Up and Go has a wide variety of clinical use making it a diverse measure that should be considered when choosing an outcome an activity based outcome measure. However, there are some limitations in the validity of the utilization of the Timed Up and Go to some populations due to a lack of data and/or poor choice of comparison outcome measures when assessing validity. Additional research is needed for young to middle aged adults.

➤ IMPLICATIONS FOR REHABILITATION

- Outcome measures are a vital component of clinical practice across all populations.
- The Timed Up and Go is a highly studied outcome measure in the geriatric population, but lacks research of its applicability to other populations.
- This study was able to highlight the clinical utility of the Timed Up and Go in populations that under utilize this outcome measure.

ARTICLE HISTORY
Received 1 June 2018
Revised 14 October 2019
Accepted 15 October 2019

KEYWORDS
Timed Up and go; psychometric properties; reliability; validity; neurologic conditions; developmental disabilities

Introduction

There are over 400 outcome measures used in physical therapy, creating challenges for therapists when choosing the appropriate measure for patients [1]. Many people with and without disabilities experience difficulty with mobility and functional activities throughout their lifespan. These difficulties with mobility and functional activities need to be quantified in order to assess patients' success with therapy.

Outcome measures are designed to measure patient strengths and limitations, provide measurement of functions, predict future outcomes, and direct interventions for patients plan of care over time [2]. Measures may be structured as self-reports or performance-based measures, and may be disease specific or applied to the general population [3]. In order to obtain a full evaluation of patient function, multiple outcome measures should be utilized, targeting the different dimensions of the International Classification of Functioning, Disability, and Health (ICF) model [3,4]. In addition, clinicians must choose outcome measures based on strong psychometric properties.

The ICF, proposed as a framework for patient management, defines an individual by his or her participation roles rather than disease or illness. This model illustrates the interplay between health condition, community activities, and contextual factors to optimize patient function and participation within their own environment. It proposes a shift from solely treating impairments to a focus on increasing participation in life roles that are meaningful to the patient [4]. Figure 1 is an illustration of the ICF applied to a community dwelling geriatric patient.

The Timed Up and Go (TUG) is a commonly used outcome measure that can assess activity limitations in the ICF model by examining the patient's ability to ambulate and perform transfers. If a patient has a deficit in these activities, it may impact life participation if they are unable to participate in their societal roles. The TUG was originally created to predict fall risk in geriatric patients [5]. The focus of the test has been shown to be a relevant outcome measure when assessing balance and general mobility for patients with various disabilities across the lifespan. The psychometric properties of the TUG have been examined individually through clinical studies in the following populations: stroke [6],

CONTACT Antonette Doty adoty@walsh.edu Physical Therapy Department, Walsh University, 2020 East Maple St., North Canton, OH 44720, USA

Supplemental data for this article can be accessed here.

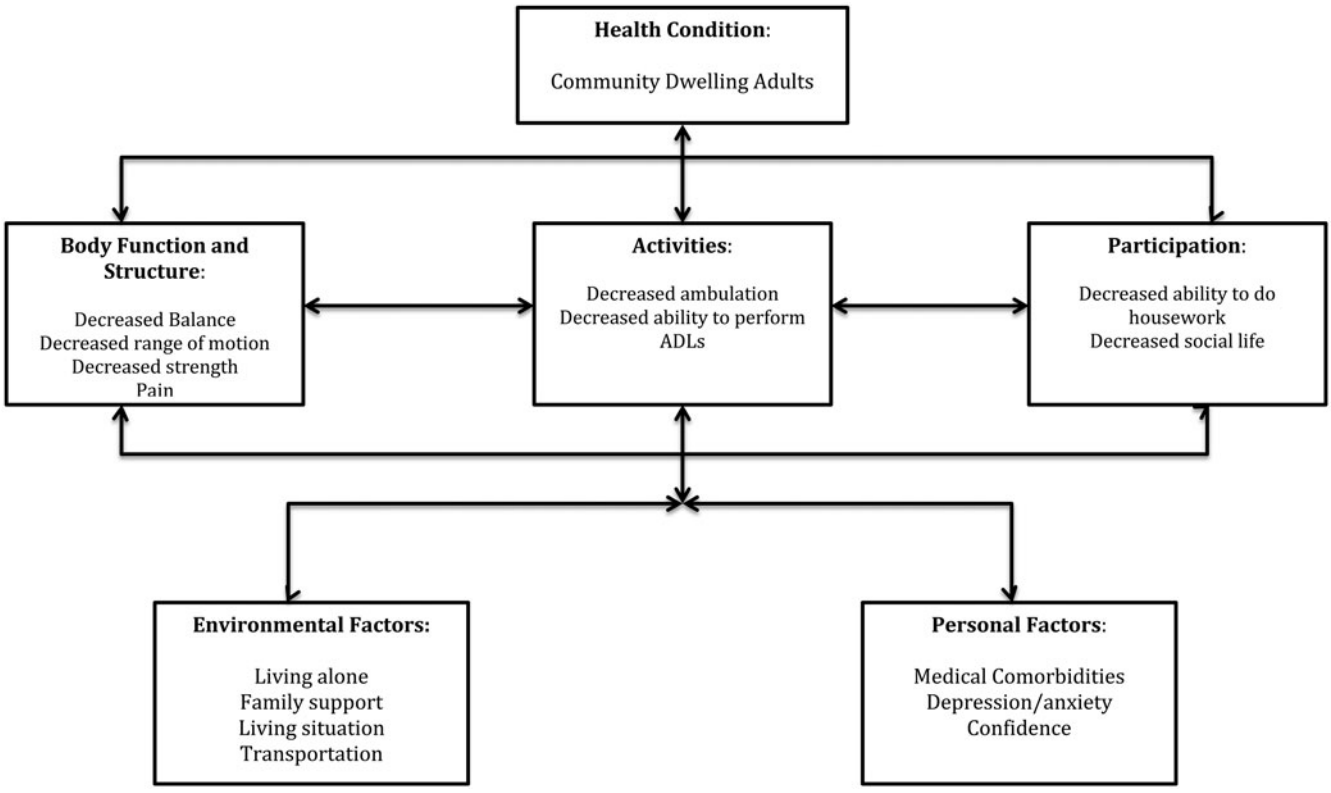

Figure 1. ICF Model applied to aging community dwelling adults.

cerebral palsy (CP) [7], Parkinson's disease (PD) [8], and spinal cord injury (SCI) [9].

A single evidence-based resource dedicated to the reliability and validity of the TUG across prevalent populations would streamline the decision-making process for clinicians. Inter-population comparisons illustrate the clinical utility of the TUG in a variety of clinical settings. Finally, providing an inclusive resource will highlight the necessity for further research in additional populations.

Since previous reviews only analyzed singular populations, the purpose of the current review is to assemble the current literature available and synthesize the clinical utility, reliability, and validity in a variety of populations including: (1) individuals with the adult diagnoses of Huntington's disease (HD), individuals with multiple sclerosis (MS), individuals with PD, individuals with SCI, individuals with a stroke; (2) pediatric patients with the diagnoses of Down syndrome (DS) and CP; and (3) typical adults and children.

Methods

This systematic review has been registered through PROSPERO, which combines prospectively registered systematic reviews into an international database. The aim of PROSPERO is to ensure that there are no duplications of systematic reviews [10]. The PROSPERO registration number associated with this systematic review is CRD42018085354. The systematic review also followed the *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) guidelines for publication, which is a checklist that insures consistency and quality of methodology behind systematic reviews [11].

Eligibility criteria

Populations available for inclusion examined neurological or developmental pathologies. Studies included are scholarly, peer-reviewed articles investigating the reliability and validity of the TUG. Studies were excluded if they were not published in English, did not examine neurological or developmental pathologies, and solely discussed responsiveness. Modified versions of the TUG were excluded, apart from the modified version for children, due to the preservation of measurement criteria and task conditions. The adaptations to the TUG for children were focused on altering the directions, yet kept the underlying methodology the same. This was done to ensure clear understanding for the child-based population.

Search strategy

Four electronic databases were searched (PUBMED, CINAHL, ProQuest, and Scopus) to locate applicable content. An example of the search strategy implemented with PUBMED is included in Supplemental Material. The original search concepts were age, disability, the TUG, and psychometric properties. The age concept was divided into the subgroups of child, adult, and elderly. The use of these subgroups and MeSH terms were determined to be the optimal way to ensure a complete search of the lifespan [12]. Following database searches, a hand search was performed of systematic reviews addressing the psychometric properties of the TUG to acquire additional articles that fit the criteria. A grey literature search was also completed in databases including Google Scholar, PROSPERO, Center Watch, and Grey Literature Report. The date of the last complete search was January of 2018.

Study selection

Following the initial search, all the articles found between databases were combined and duplicates were removed. Two reviewers then screened titles and abstracts for possible inclusion criteria. All discrepancies were discussed between the reviewers

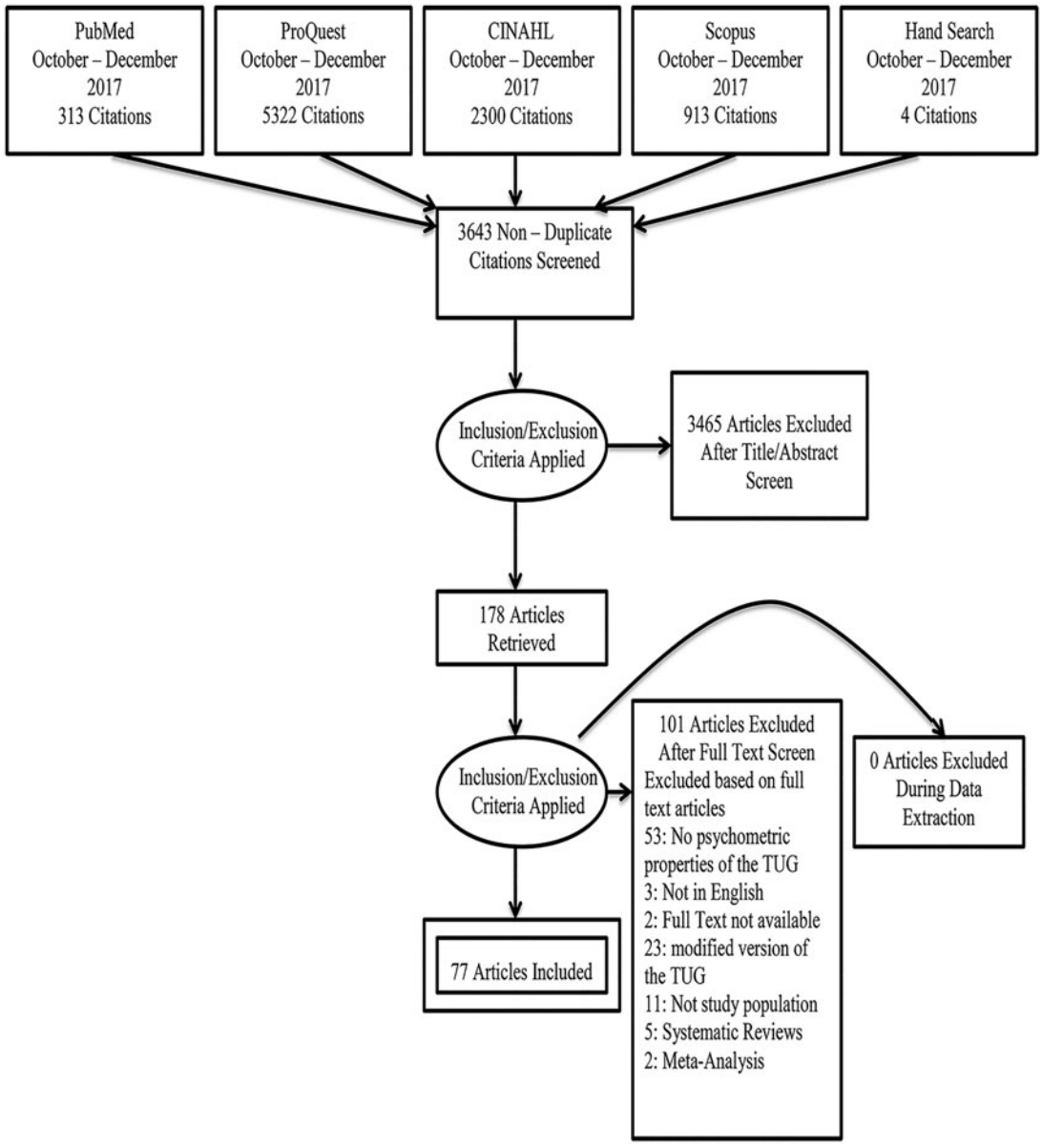

Figure 2. PRISMA flow diagram.

and a final decision was agreed upon. Two different reviewers screened the remaining full text articles. Disagreements were discussed, and a final decision was made through consensus between the two reviewers. Cohen's unweighted kappa values were calculated for each step of the screening process to quantify interrater reliability between reviewers [13]. The reference scale used for calculated kappa scores is as follows: a kappa score less than 0.00 is considered poor, 0.00–0.20 is considered slight, 0.21–0.40 is considered fair, 0.41–0.60 is considered moderate, 0.61–0.80 is considered substantial, and 0.81–1.00 is considered almost perfect [14].

Risk of bias in included studies

The McGill Mixed Methods Assessment Tool (MMAT) was used to assess the quality of the included studies [15]. This tool is applicable to a variety of study designs including qualitative, quantitative randomized control trials, quantitative non-randomized, quantitative descriptive, and mixed methods [15]. The MMAT has two screening questions for every included study followed by a category dictated by the study design. Each question is given the response of "Yes", "No", or "Can't tell." Dependent upon the category of the MMAT used, there are three to four questions to answer and score based on the percentage of questions answered yes. Scores will range from 0% (0/4 yes) to 100% (4/4 yes). Studies that scored 100% demonstrate least risk of bias as determined by sound methodology. The reliability and validity of the MMAT pilot study has been tested and determined to be adequate [15]. The risk of bias was first independently assessed by two reviewers; following individual assessment, the reviewers compared scoring and discussed any discrepancies to create a final score for each article. Kappa was calculated based on the agreement of the two individual reviewers.

Outcomes/summary measures

The outcomes of this study examined the reliability and validity of the TUG. The TUG is an outcome measure that is used to assess functional mobility in a variety of populations. The TUG is performed by having the patient seated in a chair and with the

Table 1. Study characteristics.

Author	Disease or disability	Number of subjects	Age population (mean age (SD); range in years)	Clinical use of the TUG
Aigner et al [21]·	Spinal Cord injury	2854		Walking assessment
Alexandre et al. [22]	Typical	63	Fallers: 66.68 (5.57); 60–82 Non-fallers: 66.36 (4.60); 60–75	Fall risk
Asano et al. [23]	Typical	132	Sample 1: 75.6 (7.4) Sample 2: 77	Basic mobility
Balasubramanian et al. [24]	Typical	39	73.3 (6.9)	Balance, likelihood of falls
Bandong et al. [25]	Cerebral palsy (CP), Down syndrome (DS)	CP: 16 DS: 14	CP: 8 (23); 5–12 DS: 8.7 (1.9); 6–12	Dynamic balance
Belgen et al. [26]	Stroke	50	Validity group: 54 (20) Repeatability group: 52 (20)	Fall risk
Bergström et al. [27]	Parkinson's disease, stroke	PD: 8 Stroke: 8	PD: 60.3; 46–90 Stroke: 78.4; 66–90	Fall risk
Besios et al. [28]	Cerebral palsy	20	4.85 (2.49)	Mobility
Bohannon et al. [29]	Typical	58	80.8 (7.2); 65–94	Fall risk
Bower et al. [30]	Stroke	30	68.3 (15.1)	Balance
Busse et al. [31]	Huntington's disease	75	52.12 (11.82)	Function
Carey et al. [32]	Cerebral palsy	51	5.8 (2.1); 3–10	Fall risk
Chrysagis et al. [33]	Cerebral palsy	35	14.97 (2.03); 2–18	Functional mobility
Creel et al. [34]	Typical	30	77.5; 65–92	Mobility
Dal Bello-Haas et al. [35]	Parkinson's disease	24	64.9 (8.0); 40–80	Fall risk
De Campos et al. [36]	Cerebral palsy	6	8.75 (2.95); 5–12	Functional mobility
Desai et al. [37]	Parkinson's disease	30		Mobility, risk of falls
Desai et al. [38]	Typical	72	Fallers: 81.5 (6.87) Nonfallers: 79.4 (5.48)	Physical performance
Engberg et al. [39]	Stroke	60	44–84	Functional mobility
Faria et al. [40]	Stroke	16	52 (17.9); 26–81	Performance based
Flansbjer et al. [41]	Stroke	50	Men: 59 (7); 46–72 Women: 58 (5); 50–66	Gait performance
Forsberg et al. [42]	Multiple sclerosis	81	49 (11.0); 19–73	Functional mobility
Forsberg et al. [43]	Stroke	67	68.1 (11.2); 39–92	Fall risk/balance
Fritz et al. [44]	Multiple sclerosis	29	49.1 (11.4); 20–70	Walking performance
Gan et al. [45]	Cerebral palsy	26	8.5 (2.0); 5- 11.8 y.o.	
Goh et al. [46]	Stroke, typical	Chronic stroke: 15, healthy: 15	Stroke: 57.70 (8.20) Healthy: 57.30 (3.60)	Fall risk
Hatch et al. [47]	Typical	50	81.7 (6.7); 65–95	Functional mobility
Herman et al. [48]	Typical	265	76.4 (4.3); 70–90	Fall risk
Hienkaew et al. [49]	Stroke	61	63.5 (10.0)	Postural control/balance
Huang et al. [50]	Parkinson's disease	72	67.5 (11.6)	Mobility
Iatridou et al. [51]	Cerebral palsy	20	6–14	Balance
Jette et al. [52]	Frail - muscle weakness at knees, functional limitation	105	78 (7); 65–94	Time-based performance
Johnston et al. [53]	parkinson's disease	102	72.4 (8.3)	Mobility
Jonsdottir et al. [54]	Stroke	25	61.6 (13.1)	Fall risk
Kalron et al. [55]	Multiple sclerosis	218	43.2 (13.5)	Functional mobility
Katz-Leurer et al. [56]	Cerebral palsy, post-traumatic brain injury, typical	60	9 (2); 7–13	Gait variability, balance performance
Katz-Leurer et al. [57]	Traumatic brain injury, typical	48	TBI: 8.7 (3.5) Healthy: 8.5 (3.0)	Functional balance
Kim et al. [58]	Stroke	33	52.4 (11.2); 41–64	Functional mobility
Kobayashi et al. [59]	Parkinson's disease	24	72.3 (7.4); 55–86	Mobility
Kumban et al. [60]	Cerebral palsy	33	13.8 (2.8); 6–18	Basic mobility
Landers et al. [61]	Typical, Parkinson's disease, stroke	64	72.2 (7.2)	Balance, fall risk
Learmonth et al. [62]	Multiple sclerosis	24	51.8 (7.9)	Fall risk
Lemay et al. [63]	Spinal cord injury	32	47.9 (12.8); 20–75	Standing balance
Lyders Johansen et al. [64]	Stroke	62	71.6 (13.6); 40–91	Functional performance
Mangione et al. [65]	Typical	62	78 (8)	Predictive of falls
Martin et al. [66]	Down syndrome	12	9.5 (4.4); 3–17	Mobility
Morris et al. [67]	Parkinson's disease	24	65.5 (10.5); 50–81	Detect differences in motor performance
Nair et al. [68]	Stroke	33	68.2; 47–86	Mobility
Ng et al. [69]	Stroke, typical	72	Stroke: 62.0 (6.2) Control: 64.3 (7.8)	Balance
Ng et al. [70]	Stroke, typical	Stroke: 11, healthy: 10	Stroke: 61.7 (7.2) Healthy: 63.5 (6.1)	Functional mobility
Nicolini-Panisson et al. [71]	Down syndrome, typical	499	Down Syndrome 10.5 (4.3); 3–18 Healthy 10.7 (4.3); 3–18	Functional mobility
Nilsagard et al. [72]	Multiple sclerosis	43	52 (9)	Ability to detect change in function

(*continued*)

Table 1. Continued.

Author	Disease or disability	Number of subjects	Age population (mean age (SD); range in years)	Clinical use of the TUG
Nilsagard et al. [73]	Stroke	37	79 (67–86)	Balance
Pires-Oliveria et al. [74]	Typical	84	68 (5.3)	Fall risk
Podsiadlo et al. [5]	Typical	60	Elderly: 79.5; 60–90 Control: 75; 70–84	Basic functional mobility
Poncumhak et al. [75]	Spinal cord injury	66	FIM-L 6: 50.9 (13.4) FIM-L 7: 50.2 (9.5)	Fall risk
Quinn et al. [76]	Huntington's disease	75	52.12 (11.82)	Physical performance
Rennie et al. [77]	Parkinson's disease	100	73.2 (5.6)	Fall risk
Russell et al. [78]	Typical	344	75.9	Fall risk
Salbach et al. [79]	Typical	51	72 (11); 38–91	Fall risk/quality of life
Sanjivani et al. [80]	Cerebral palsy	30	8.16 (2.76); 2–12	Mobility/function
Schlenstedt et al. [81]	Parkinson's disease	85	67.2 (9.8); 40–82	Postural control/balance
Shumway-Cook et al. [82]	Typical	30	Patients w/history of 2 or more falls in the past 6mo: 86.2 (6); 76–95 Patients w/no history of falls: 78 (6); 65–85	Fall risk
Sosnoff et al. [83]	Multiple sclerosis	13	51.5 (11.3)	Fall risk
Spagnuolo et al. [84]	Typical	64	57 (10.0); 40–84	Balance
Steffen et al. [85]	Parkinson's disease	37	71 (12)	Balance/ambulation
Steffen et al. [86]	Typical	96	Male: 73 (8); 61–89 Female: 73 (8); 60–88	Basic mobility skills
Stretton et al. [87]	Typical	243	79; 74–84	Performance
Tanji et al. [88]	Parkinson's disease	79	65.5; 43–85	Sensitivity to disability
Tsang et al. [89]	Stroke, typical	Stroke: 106, control: 48	Stroke: 57.1 (11.0) Control group: 60.2 (9.3)	Fall risk
Ursin et al. [90]	Stroke	183	72.1 (12.2); 25–94	Balance, mobility
Van Hedel et al. [91]	Spinal cord injury	75	4.85 (2.49)	Balance/fall risk
Verheyden et al. [92]	Parkinson's disease	38	Parkinson's disease: 69 (8); 47–88 Control group: 68 (0); 52–85	Fall risk
Villamonte et al. [93]	Down syndrome	21	16.8 (9.1); 5–31 y.o.	Balance/locomotor tasks
Williams et al. [94]	Cerebral palsy, spina bifida, typical	217	Disability: 8.9 (4.3); 3–19 Healthy: 5.8 (1.67); 3–9	Functional mobility
Winser et al.[95]	Multiple sclerosis	60	50 (10.76); 21–65	Fall risk
Wong et al. [96]	Stroke, typical	Stroke: 35 typical adult: 29	Stroke 57.26 (7.19) Healthy 57.76 (5.77)	Walking performance
Wrisley et al. [97]	Typical	35	72.9 (7.8); 60–90	Falls

command "go", rise from the chair, walk 3 meters, turn around, return to chair and sit. The trial is timed from when the patient's back leaves the backrest to when the patient returns to the seated position, and the patient is allowed one practice trial [5]. A shorter time represents better mobility [5]. The review looked at the reliability and validity of this outcome measure within each subgroup of the populations discussed.

Reliability measures how consistently and free from error an outcome measure captures results [16]. The sub-types of reliability that were used for the systematic review include test-retest, inter-rater, and intra-rater. Test-retest reliability examines reliability of a test between test trials. Inter-rater reliability is how reliable an outcome measure is between different test administrators, while intra-rater reliability demonstrates how reliable an outcome measure is when the same administrator conducts a test multiple times [16]. Intraclass correlation coefficient (ICC) measures reliability. Values less than 0.5 were considered to be poor, between 0.5 and 0.75 indicated moderate, values between 0.75 and 0.9 indicated good, and values larger than 0.9 were determined to be excellent [17].

Validity is defined by ability of a measure to capture the intended data [16]. The types of validity included were concurrent and predictive. Concurrent validity is the extent to which two measurements agree when measuring the same construct as represented by Spearman Rho (rho) and Pearson (*r*) [16,18]. The following scale was used to determine the strength of r values: greater than or equal to 0.60 was considered a strong correlation, 0.30–0.59 was a moderate correlation, and less than 0.30 was a weak correlation [19]. Predictive validity measures whether a finding was a predictor of a future outcome and is reported by area under the curve and odds ratio [20].

Data extraction

The reviewers extracted data regarding reliability and validity of the TUG across multiple ages and populations, as well as the demographics of participants utilizing the TUG. When available, the clinical use of the TUG and normative data were obtained from the articles. Two reviewers performed the data extraction. Both reviewers extracted data from the included articles and then data was cross-checked for accuracy.

Results

Study selection

The results of the search strategy produced 8848 articles, duplicates were removed to produce 3643 articles. Titles and abstracts were screened together to produce 178 articles to be reviewed for full text screen, with a kappa value of 0.59 (CI 95%, 0.53–0.66: moderate agreement) [13,14]. Upon completion of the full text screen, 77 articles were included in this systematic review, with a kappa value of 0.28 (CI 95%, 0.13–0.44: fair agreement) [13,14]. A hand search through relevant, previously published systematic reviews provided four additional articles that fit the inclusion criteria. The grey literature search did not produce any relevant

Table 2. McGill mixed methods (MMAT) quality assessment results of included studies.

	Category 4: Quantitative descriptive				
Author	4.1	4.2	4.3	4.4	Score
Aigner et al. [21]	C	Y	Y	C	50%
Alexandre et al. [22]	Y	Y	Y	Y	100%
Asano et al. [23]	Y	Y	Y	Y	100%
Balasubramanian et al. [24]	Y	Y	Y	Y	100%
Bandong et al. [25]	Y	Y	Y	Y	100%
Belgen et al. [26]	Y	Y	Y	Y	100%
Bergström et al. [27]	Y	Y	Y	Y	100%
Besios et al. [28]	N	Y	Y	C	50%
BohanNn et al. [29]	C	C	Y	Y	50%
Bower et al. [30]	Y	Y	Y	Y	100%
Busse et al. [31]	C	Y	Y	C	50%
Carey et al. [32]	Y	Y	Y	Y	100%
Chrysagis et al. [33]	Y	Y	Y	Y	100%
Creel et al. [34]	Y	Y	Y	Y	100%
Dal Bello-Haas et al. [35]	Y	Y	Y	Y	100%
De Campos et al. [36]	C	Y	Y	Y	75%
Desai et al. [37]	Y	Y	C	Y	75%
Desai et al. [38]	Y	Y	Y	Y	100%
Engberg et al. [39]	Y	Y	Y	Y	100%
Faria et al. [40]	Y	Y	Y	Y	100%
Flansbjer et al. [41]	Y	Y	Y	Y	100%
Forsberg et al. [42]	Y	Y	Y	Y	100%
Forsberg et al. [43]	Y	Y	Y	Y	100%
Fritz et al. [44]	C	Y	Y	Y	75%
Gan et al. [45]	Y	Y	Y	Y	100%
Goh et al. [46]	Y	Y	Y	Y	100%
Hatch et al. [47]	Y	Y	Y	Y	100%
Heinkaew et al. [48]	C	Y	Y	C	50%
Herman et al. [49]	Y	Y	Y	Y	100%
Huang S. et al. [50]	Y	Y	Y	Y	100%
Iatridou et al. [51]	Y	Y	Y	Y	100%
Jette et al. [52]	Y	Y	Y	Y	100%
Johnston et al. [53]	Y	Y	Y	Y	100%
Jonsdottir et al. [54]	Y	Y	Y	Y	100%
Kalron et al. [55]	Y	Y	Y	Y	100%
Katz-Leurer et al. [56]	Y	Y	Y	Y	100%
Katz-Leurer et al. [57]	Y	Y	Y	Y	100%
Kim et al. [58]	Y	Y	Y	Y	100%
Kobayashi et al. [59]	Y	Y	Y	Y	100%
Kumban et al. [60]	Y	Y	Y	Y	100%
Landers et al. [61]	Y	Y	Y	Y	100%
Learmonth et al. [62]	Y	Y	Y	Y	100%
Lemay et al. [63]	Y	Y	Y	Y	100%
Lyders et al. [64]	Y	Y	Y	Y	100%
Mangione et al. [65]	Y	Y	Y	Y	100%
Martin et al. [66]	C	Y	Y	Y	75%
Morris et al. [67]	Y	Y	Y	Y	100%
Nair et al. [68]	Y	Y	Y	C	75%
Ng et al. [69]	Y	Y	Y	C	75%
Ng et al. [70]	Y	Y	Y	Y	100%
Nicolini-Panisson et al. [71]	Y	Y	Y	Y	100%
Nilsagard et al. [72]	Y	Y	Y	Y	100%
Nilsagard et al. [73]	Y	Y	Y	Y	100%
Pires-Oliveria et al. [74]	Y	Y	Y	Y	100%
Podsiadlo et al. [5]	Y	Y	Y	Y	100%
Poncumhak et al. [75]	Y	Y	Y	Y	100%
Quinn et al. [76]	Y	Y	Y	Y	100%
Rennie et al. [77]	Y	Y	Y	Y	100%
Russell et al. [78]	Y	Y	Y	Y	100%
Salbach et al. [79]	Y	Y	Y	Y	100%
Sanjivani et al. [80]	Y	Y	Y	Y	100%
Schlenstedt et al. [81]	C	Y	Y	Y	75%
Shumway-Cook et al. [82]	Y	Y	Y	Y	100%
Sosnoff et al. [83]	C	C	Y	Y	50%
Spagnuolo et al. [84]	C	Y	Y	Y	75%
Steffen et al. [85]	Y	Y	Y	Y	100%
Steffen et al. [86]	Y	Y	Y	Y	100%
Stretton et al. [87]	Y	Y	Y	C	75%
Tanji H et al. [88]	Y	C	Y	Y	75%
Tsang et al. [89]	Y	Y	Y	Y	100%
Ursin et al. [90]	Y	Y	Y	Y	100%
Van Hedel et al. [91]	Y	Y	N	Y	75%
Verheyden et al. [92]	Y	Y	Y	Y	100%
Villamonte et al. [93]	Y	Y	Y	Y	100%
Williams et al. [94]	Y	C	Y	Y	75%
Winser et al. [95]	Y	Y	Y	Y	100%
Wong et al. [96]	Y	Y	Y	Y	100%
Wrisley et al. [97]	Y	Y	Y	Y	100%

Y: Yes; N: No; C: Can't tell.

Screening Questions: 1. Are there clear qualitative and quantitative research questions (or objectives), or a clear mixed methods question (or objective)? 2. Do the collected data allow address the research question (objective)? Category 4: Quantitative descriptive: 4.1. Is the sampling strategy relevant to address the quantitative representative of the population under study? 4.2. Is the sample representative of the population under study? 4.3. Are measurements appropriate (clear origin, or validity known, or standard instrument)? 4.4. Is there an acceptable response rate (60% of above)?

prepublications. The PRISMA flow diagram with detailed explanation of the selection process is presented in Figure 2.

Study characteristics

The study designs of the included full text articles were cross-sectional, quantitative-descriptive. The sample size in each study ranged from 6 to 2854, with the ages ranging from 3 to 94 years old. Studies examining a specific disease or disability were noted; otherwise, the study population was noted to be "without disability", "healthy", or "typical". Specific study characteristic data can be referenced in Table 1.

Risk of bias in included studies

The MMAT quality assessment scores can be viewed in Table 2. Most studies score a 75%-100% on the MMAT, with only six falling below 75%. This means that a majority of the studies had a low risk of bias. Many studies proved to be methodologically sound; however, a commonly missed criterion was acceptable response rate (question 4.4). The response rate was not able to be calculated because many studies failed to report the number of individuals originally recruited or the number of patients who completed the test-retest portion of the study. Another commonly missed criterion, was not reporting the source of the study sample (question 4.1). The kappa calculated for the agreement between reviewers for quality assessment was 0.56 (CI 95%, 0.38–0.73: moderate) [13,14].

Results of individual studies

Reliability

Reliability was divided based on test-retest, inter-rater and intra rater reliability. Reliability for all included studies can be found in Table 3. A majority of populations reported excellent test-retest reliability. The populations that support this include: pediatric individuals with CP [28,32,45,51,80], individuals with HD [76], individuals with a stroke[30,49,68,69], typical adults [65,70,85] and individuals with MS [62,72,95]. A range of test-retest reliability was found for the remainder of the populations that assessed test-retest reliability. Moderate to excellent values were reported for typical children [71,94]. Moderate to good values were reported for individuals with PD [35,50,85]. Good to excellent test-retest reliability was reported for pediatric individuals with DS

Table 3. Reliability (test-retest, inter-rater, and intra-rater) of the included studies.

Disease/disability	Author	Test-retest (ICC unless noted by *r*)	Inter-rater (ICC unless noted by *r*)	Intra-rater (ICC)
Cerebral palsy				
	Besios et al. [28]	0.994		
	Carey et al. [32]	GMFCS I: 0.97 GMFCS II: 0.981 GMFCS III: 0.995 Young (3–5 y.o.): 0.965 Old (6–10): 0.911		
	Gan et al. [45]	0.99 (95% CI, 0.98–0.99)		
	Iatridou et al. [51]	1st to 2nd trial: 0.999 1st to 3rd trial: 0.998 2nd to 3rd trial: 0.997 1st,2nd,3rd: 0.998		
	Sanjivani et al. [80]	0.99*		
	Williams et al. [94]			Trial 1: 0.98 (CI 95%, .97–0.99) Trial 2: 0.98, (CI 95%, 0.88–0.99) Same day retest time 1 and 2: 0.99, (CI 95%, 0.91–0.99)
Down syndrome				
	Martin et al. [66]	$r = 0.923$		
	Nicolini-Panisson et al. [71]	0.82		
	Villamonte et al. [93]	Boys: 0.22 Girls: 0.24		
Huntington's disease				
	Quinn et al. [76]	Pre-manifested stage of Huntington's: 0.93 Manifested stage of Huntington's: 0.96 Early Stage of Huntington's: 0.94 Middle Stage of Huntington's: 0.95 Late Stage of Huntington's: 0.97		
Multiple sclerosis				
	Learmonth et al. [62]	0.97 (CI 95%, 0.93–0.99)		
	Nilsagard et al. [72]	0.91 (CI 95 %, 0.83–0.95)		
	Winser et al. [95]	0.99* (CI 95%, 0.99–1.00)	0.99* (CI 95%, 0.99–1.00)	
Parkinson's disease				
	Dal Bello-Haas et al. [35]	0.69 (CI 95%, 0.41–0.850)		
	Huang et al. [50]	0.80 (CI 95%, 0.70–0.87)		
	Morris et al. [67]			Patients off medication and experienced clinician: 0.99 Patients off medication and inexperienced clinician: 0.87 Patients on medication and experienced clinician:0.99 Patients on medication and inexperienced clinician: 0.99
	Steffen et al. [85]	0.85		
	Verheyden et al. [92]		0.99 (CI 95%, 0.99–0.99)	0.99 (CI 95%, 0.99–1.00)
Spinal cord injury				
	Poncumhak et al. [75]		FIM Locomotor 6 (with device): 0.999 FIM Locomotor 7 (without device): 1.00	
	van Hedel et al. [91]		$r = 0.973$*	$r = 0.979$*
Stroke				
	Bower et al. [30]	0.99 (CI 95%, 0.98, 0.99)		
	Faria et al. [40]		0.96*	0.85*
	Flansbjer et al. [41]		0.96 (CI 95%, 0.93–0.98)	
	Heingkaew et al. [49]	All participants: 0.97 (CI 95%, 0.94–0.99) MAS score = 0: 0.97 (CI 95%, 0.88–0.99) MAS score = 1–1+: 0.98 (CI 95%, 0.96–0.99) MAS score = 2: 0.92 (CI 95%, 0.65–0.97)		
	Lyders Johansen et al. [64]		Test Session 1: 0.97 (CI 95%, 0.96–0.98) Test Session 2: 0.99 (CI 95%, 0.98–0.99)	Rater 1: 0.96 (CI 95%, 0.93–0.97) Rater 2: 0.95 (CI 95%, 0.91–0.97)

(*continued*)

Table 3. Continued.

Disease/disability	Author	Test-retest (ICC unless noted by *r*)	Inter-rater (ICC unless noted by *r*)	Intra-rater (ICC)
	Nair et al. [68]	0.96		
	Ng et al. [69]	0.95		
Typical adults				
	Creel et al. [34]		0.99	0.98
	Mangione et al. [65]	0.96 (CI 95%, 0.91–0.98)		
	Ng et al. [70]	0.97		
	Podsiadlo et al. [5]		0.99	0.99
	Shumway-Cook et al. [82]		$r = 0.98$	
	Spagnuolo et al. [84]			0.936 (CI 95%, 0.895–0.961)
	Steffen et al. [85]	0.97 (CI 95%)		
Typical children				
	Katz-Leurer et al. [56]			0.85 (CI 95%, 0.74–0.92)
	Nicolini-Panisson et al. [71]	0.95		0.93, 0.94, 0.95
	Williams et al. [94]	1 week between testing 0.83 (CI 95%, 0.77–0.88) 1 week between testing preschool, 0.61 (CI 95%, 0.39–0.75) 1 week between testing primary 0.83 (CI 95%, 0.73–0.89)		Trial 1, 0.80 (CI 95%, 0.75–0.84) Trial 2, 0.89 (CI 95%, 0.86–0.92) Trial 3, 0.85 (CI 95%, 0.81–0.89) Same day retest 0.89 (CI 95%, 0.86–0.92) Same day retest preschool 0.82 (CI 95%, .72–0.88) Same day retest primary 0.71 (CI 95%, 0.61–0.85)

*Significance of $p < 0.01$.
ICC: Intraclass Correlation Coefficient; GMFCS: Gross Motor Functional Classification Scale; MAS: Modified Ashworth Scale; r = Pearson's Correlation.

[66,71] in two studies; however, one study reported poor values for this population [93].

Intra-rater reliability reported as excellent in individuals that were diagnosed with CP [94], PD [67,92], and SCI [91], as well as typical adults [5,34,84]. Individuals with a stroke [40,64] and typical children [56,71,94] reported intra-rater reliability as a range of good to excellent when assessed. Inter-rater reliability was found to be excellent when studied in individuals with PD [92], individuals with SCI [75,91], individuals with a stroke[40,41,64], typical adults [5,34,82]and individuals with MS [95].

Validity

Validity was divided into two subgroups: concurrent and predictive. Within these categories, populations were then grouped based on strength of correlation. The TUG proved to have strong correlations with outcomes utilized in the populations including individuals diagnosed with SCI, stroke, CP, and DS. In patients with SCI the TUG demonstrated a strong correlation [63,75,91] when compared with Berg Balance Scale (BBS) and timed walking tests. The BBS, Dynamic Gait Index (DGI), Figure 8 test, and timed walking tests also have strong correlations with the TUG when assessed in patients with a stroke [39,41,54,69,70,90].

There was a moderate to strong correlation between the TUG and outcomes used to assess children with CP and DS, and individuals with MS. Correlation in concurrent validity in children with CP was shown between the TUG and outcome measures such as the Five Time Sit to Stand (FTSSS), BBS, and Four-Square Step Test (FSST) [25,33,36,45,60,94]. The children with DS that were assessed show moderate to strong correlation for concurrent validity with outcome measures assessing similar activities [7,25,71]. Moderate to strong correlations to the TUG occurred in individuals with MS when compared to timed walking tests, DGI, and other mobility measures, demonstrating acceptable clinical utility to the TUG in individuals with MS [43,72,83]. The concurrent validity values vary from weak to strong correlation in outcome measures used to evaluate individuals with PD [58,80,91]. In typical adults, depending on the outcome measure that the TUG was being compared, the correlations ranged from weak to strong as well [5,23,24,29,34,37,47,48,74,78,79,84,97].

Correlations between disease specific outcome measures demonstrated weak correlation. The Unified Huntington's Disease Rating Scale (UHDRS) was the outcome measure used to assess concurrent validity of the TUG and demonstrated weak correlation [31]. The TUG had a weak concurrent validity correlation when compared with the Lower Extremity Motor Score (LEMS) when assessed in patients with spinal cord injury [21]. The outcome measures used to assess typical children were found to have weak correlations with the TUG [57]. Table 4 expresses concurrent validity in all populations.

Predictive validity was found with area under the curve (AUC) and odds ratio in a limited number of studies, only pertaining to the typical adult population. In the first study the odds ratio data were results of multivariate logistic regression on functional mobility and balance outcomes discriminating fallers and recurrent fallers. The odds ratio reported for one or more falls was 0.77 and for two or more falls was 1.21 [24]. The second study determined if a patient takes longer than 11 s to complete the TUG, then a odds ratio of fall risk was determined to be 13.1 [97]. However, based on likelihood ratios, the optimal cut of time for determining fall risk is 12.3 s to optimize both sensitivity and specificity [97]. Lastly, AUC was reported in one study as 0.63 (95% CI; 0.57–0.69) for patients who fell in the twelve month period that the study included. The AUC slightly increased to 0.64 (95% CI, 0.58–0.70) when patients were not receiving conventional therapy. These results indicate the TUG is a potential predictor of risk for recurrent falling [78].

Discussion

The goal of the present review was to present a comprehensive overview of the ability of the TUG to be used for multiple pathologies and typical individuals. The discrepancies between

Table 4. Validity (concurrent) of the included studies.

Pathology	Author	Concurrent validity (Pearson's correlation: *r*, Spearman rho: rho)	Compared outcome measures
Cerebral palsy			
	Bandong et al. [25]	$r = 0.70$	FSST
	Chrysagis et al. [33]	$r = -0.600$*	GMFM-88
	De Campos et al. [36]	$r = -0.47$ P ≤ 0.12	GMFM D
		$r = -0.71$*	GMFM E
	Gan et al. [45]	$r = -0.88$*	BBS
		$r = -0.77$*	FRT
		$r = -0.80$*	Sit-to-Stand
		$r = -0.89$*	GMFM-88
	Kumban et al. [60]	$r = 0.552$*	FSST
		$r = -0.719$*	BBS
	Williams et al. [94]	rho = −0.52*	GMFM
Down syndrome			
	Bandong et al. [25]	rho = 0.71	FSST
	Martin et al. [66]	$r = -0.525$, P = 0.08	TUDS
	Nicolini-Panisson et al. [71]	$r = -0.55$*	GMFM
Huntington's disease			
	Busse et al. [31]	$r = 0.16$	UHDRS motor
		$r = -0.33$*	UHDRS functional assessment scale
		$r = -0.25$	UHDRS total functional capacity score
Multiple sclerosis			
	Fritz et al. [44]	$r = 0.805$*	SSST
	Forsberg et al. [43]	$r = -0.762$*	DGI
	Kalron et al. [55]	$r = 0.652$*	FSST
	Nilsagard et al. [72]	$r = 0.85$ (CI 95%, 0.74–0.92)	30meter walk test
		$r = 0.83$ (CI 95%,0.71–0.91)	10meter walk test
	Sosnoff et al. [83]	rho = −0.88*	FAP
	Winser et al. [95]	$r = -0.78$	BBS
		$r = 0.79$	PG of ICARS
		$r = 0.54$	ICARS
		$r = 0.58$	SARA
		$r = 0.79$	SARABal
		$r = 0.71$	EDSS
		$r = -0.39$	Barthel Index
Parkinson's disease			
	Bergström et al. [27]	$r = -0.81$*	Part of Mini-BESTest
	Dal Bello-Haas et al. [35]	$r = -0.44$*	ABC
	Desai et al. [37]	$r = -0.28$*	ABC
		$r = -0.43$	MFES
	Johnston et al. [53]	$r = -0.57$ (CI 95%, −0.69 to −0.42)*	DEMMI
		rho= −0.12 (CI 95%, −0.33 to 0.10)	MMSE
		rho = 0.10 (−0.10 to 0.29)*	Charlson Co-morbidity Index
	Kobayashi et al. [59]	$r = -0.68$*	6MWT
		$r = 0.63$*	Age
		$r = 0.66$*	Hoehn and Yahr Stage
		$r = 0.23$	UPDRS
		$r = -0.4$*	BBS
	Morris et al. [67]	Trial 1–2 end of dose (EOD) $r = 0.96$	Matched PD patients and dosage
		Peak dose (pd) $r = 0.96$	
		Trial 2–3 EOD $r = 0.90$	
		pd = 0.73	
		Trial 3–4 EOD $r = 0.80$	
		pd = 0.82	
		Trial 4–5 EOD $r = 0.98$	
		pd = 0.99	
	Rennie et al. [77]	$r = -0.42$*	GVI
	Schlenstedt et al. [81]	$r = -0.83$*	FAB scale
		$r = -0.76$*	Mini-BESTest
		$r = -0.81$*	BBS
		$r = 0.66$*	PIGD
		$r = 0.43$*	VAS
		$r = 0.54$*	UPDRS total
		$r = 0.38$*	UPDRS motor
	Tanji et al. [88]	$r = 0.51$*	OARS
		$r = 0.44$*	OARS best total
		$r = 0.67$*	UPDRS
	Verheyden et al. [92]	rho = 0.51	Hoehn & Yahr
		rho = 0.61	Part III UPDRS
Spinal cord injury			
	Aigner et al. [21]	$r = 0.26$	S-LEMS Stage 1
		$r = 0.34$	Stage 2
		$r = 0.56$	Stage 3

(*continued*)

Table 4. Continued.

Pathology	Author	Concurrent validity (Pearson's correlation: *r*, Spearman rho: rho)	Compared outcome measures
		$r = 0.64$	Stage 4
		$r = 0.60$	Stage 5
		$r = 0.58$,	I-LEMS Stage 1
		$r = 0.45$	Stage 2
		$r = 0.60$	Stage 3
		$r = 0.68$	Stage 4
		$r = 0.66$	Stage 5
	Lemay et al. [63]	$r = -0.815$	BBS
		$r = -0.623$	2 MWT
		$r = -0.646$	
	Poncumhak et al. [75]	$r = -0.692$	10 meter walk test
	van Hedel et al. [91]	rho $= -0.76$	WISCI II
		rho $= -0.88$	10 meter walk test
Stroke			
	Bower et al. [30]	rho$= 0.44^*$	Eyes open center of pressure
		rho$= -0.13$	WBA
		rho$= -0.03$	Sit-to-stand peak force asymmetry
		rho$= -0.23$	Sit-to-stand peak rate of force development
		rho$= -0.57^*$	Mediolateral weight shifting
	Engberg et al. [39]	rho $= -0.70^*$	FES
		rho $= -0.68, -0.72^*$	BBS
	Flansbjer et al. [41]	$r = -0.89$	6 MWT
	Forsberg et al. [42]	$r = -0.484^*$	ABC
	Goh et al. [46]	$r = 0.59^*$	FSST
	Jonsdottir et al. [54]	$r = -0.77$	DGI
	Ng et al. [69]	$r = -0.960^*$	6 MWT
	Ng et al. [70]	rho $=$ (time) 0.792^* (score)$= 0.658^*$	PWT 20 cm
		rho $=$ (time) 0.813^* (score) $= 0.63^*$	PWT 30.5cm
		rho $=$ (time) 0.842^* score $=0.466^*$	PWT 38 cm
	Nilsagard et al. [71]	$r = -0.46^*$	ABC
	Ursin et al. [90]	$r = -0.74$	BBS
		$r = 0.64$	Fig 8 test
		$r = 0.88$	MWS
	Wong et al. [96]	$r = 0.886$	Fig 8 test
Typical adults			
	Asano et al. [23]	$r = -0.46$, (95% CI -0.61, -0.25)	ASCQ
	Bohannon et al. [29]	$r = -0.70$	Physical Functioning Scale
	Creel et al. [34]	$r = 0.92^*$	Fig 8
	Desai et al. [37]	$r = 0.72$	6 MWT
		$r = -0.62$	BBS
	Hatch et al. [47]	$r = 0.698$	ABC
		$r = 0.810^*$	BBS
	Herman et al. [48]	$r = -0.19$	MMSE
		$r = -0.400$	DGI
		$r = -0.509$	BBT
		$r = -0.430$	ABC
	Pires-Oliveria et al. [74]	$r = -0.362$ and 0.001 (CI 95%)	SF – 36 physical
		$r = -0.095$ and 0.389 (CI:95%)	SF-36 mental
	Podsiadlo et al. [5]	$r = -0.72$	BBS
		$r = -0.51$	Barthel Index
	Salbach et al. [79]	$r = -0.34$ (95%)	ABC
		$r = -0.52$ (95%)	ABC-CF
	Spagnuolo et al. [84]	$r = -0.65$	BBS
	Wrisley et al. [97]	$r = -0.84^*$	FGA
Typical children			
	Katz-Leurer et al. [57]	Preferred leg $r = -0.10$ (length), $r = -0.09$ (time)	FRT
		Non preferred leg $r = 0.05$ (length), $r = 0.04$ (time	Typical development preferred leg vs. non preferred let

$^* = p < 0.05$; P: p value; Mini-BESTest: Mini Balance Evaluations Systems Test; ABC: Activities-Specific Balance Confidence; MFES: Modified Falls Efficacy Scale; DEMMI: the de Morton Mobility Index; MMSE: Mini-Mental State Exam; FSST: Four Square Step Test; MWT: Minute Walk Test; UPDRS: Unified Parkinson's Disease Rating Scale; BBS: Berg Balance Scale; PD: Parkinson's Disease; GVI: Gait Variability Index; FAB: Fear Avoidance Belief Questionnaire; PIGD: Postural Instability and Gait Difficulty Scale; VAS: Visual Analog Scale; OARS: Older Americans Resource and Services Disability Subscale; SSST: Six Spot Step Test; FAP: Functional Ambulation Profile; PG: Postural and gait subcomponent; ICARS: International Co-operative Ataxia Rating Scale; SARA: Scale for Assessment and Rating of Ataxia; SARABal: Balance Component of the SARA; EDSS: Expanded Disability Status Scale; UHRDS: Unified Huntington's Disease Rating Scale; GMFM: Gross Motor Function Measure; FRT: Functional Reach Test; TUDS: Timed Up and Down Stairs; WBA: Weight Bearing Asymmetry; FES: Falls Efficacy Scale. PWT: Parallel Walk Test; MWS: Maximal Walking Speed; S-LEMS: Single Variable- Lower Extremity Motor Score; I-LEMS: Ten Variable LEMS; WISCI: Walking Index for Spinal Cord Injury; ASCQ: The Ambulatory Self-Confidence Questionnaire; SF-36: Short Form-36; ABC-CF: Canadian French Version of the ABC.

populations can frequently be attributed to differences in impairments associated with specific health conditions.

The cause of variance found in reliability and validity are due to common themes across an array of populations. For example, age impacted the performance of individuals; older individuals required increased time to complete the TUG compared to younger individuals [29,33,67,86,92]. Another common theme that impacted TUG scores was the use of multiple trials. Participants generally did better on the second trial than the first, as experience allowed for increased learning and improved performance [68]. However, if fatigue was a factor that impacted the individual, the score in latter trials may not be representative of function. Another common theme across a variety of populations is the minimal differences in inter-rater reliability. This could possibly be explained by the objectivity of the TUG. The TUG does not require advanced interpretation of results, as the only data recorded was the time taken to complete 3 meters.

Alterations in reliability and validity in the studies that examine children may have external factors that could impact the data. For example, as noted by Williams et al. when children perform the TUG they are more likely to add different movement patterns such as hopping or skipping [94]. It was recommended to redo the trial if the patient deviates from walking since this could mask the child's true ability. This data could be extrapolated to additional populations that assessed children such as DS and CP. Children may also have limited understanding of the exact directions which may lead to increased time to complete the outcome measure.

For many conditions, the level of an individual's function may affect reliability and validity. Most of these disorders they are categorized based on functional differences. The conditions that emphasize this idea include CP with Gross Motor Functional Classification Scale (GMFCS) levels, HD with United Huntington's Disease Rating Scale (UHDRS) levels, PD with Hoehn Yahr stages, SCI level of injury, and location of stroke. In addition to being classified based on functional level, there is a likely functional decline associated with these disorders. Therefore, consideration must be taken when interpreting the results of the TUG because it does not indicate the extent of injury and it is not sensitive to the individual body structure and function [91]. The TUG may have demonstrated better reliability and validity if it were consistently assessed in similar classifications of the disease within each population.

An individual may lack a specific body structure and function component due to their specific disease-causing difficulty and increased time to complete the TUG. For example, a patient may have limited functional strength to perform a sit to stand transfer but may able to ambulate once standing. The TUG is a multidimensional test; consequently, the therapist must consider the impact of the patients' overall performance because they may be challenged by one aspect while having success with another. However, this is not reflected in the final time since the TUG is a composite time of multiple tasks performed as a whole.

Concurrent validity was assessed by comparing the TUG to another outcome measure. Outcome measures used for comparison assessed constructs in multiple domains such as: activity, participation and body structure and function. Depending on the construct being assessed, correlations ranged from weak to strong. Validity proved to be strong when the TUG was associated with outcome measures that analyze the same construct of activity. This would be expected due to the similarity in tasks being measured. For example, in patients with SCI there was a strong correlation with BBS and the timed walking test. These outcome measures also study the same fundamental principles as the TUG such as balance and ambulation capabilities. This is also true for the outcomes used for concurrent validity with children diagnosed with CP. For this population they assessed dynamic balance and movement.

The TUG was not highly correlated with outcome measures that did not assess the same constructs as the TUG. Many studies proved this when using outcome measures that assessed alternative constructs like body structure function. In studies that assessed the individuals with HD, concurrent validity was assessed with the UHDRS. While there are components to this measure that are in line with the goals of the TUG many of the constructs fall outside the realm of activity and into body structure function categories. This is also true for the UPDRS in individuals with PD and the Lower Extremity Motor Scale (LEMS) in individuals with SCI. These differences are also noted in the studies assessing validity in typical adults. In addition, assessing specific characteristics of movement, as some studies did, would not correlate well with a test such as the TUG since it is focused on more global function.

Within each population there are population specific aspects that may account for differences in reliability and validity. First, for patients with PD there may be differences due to the on: off periods of their medication. These on:off periods of the medications can hinder or bolster their performance and should be taken into consideration when designing studies [88]. In the children diagnosed with DS there was a variety of results. According to Villamonte et al. [93], which recorded poor reliability of the TUG, the results of the study may not be able to be generalized to the population due to small sample size and the wide range of cognitive and motor abilities of the tested children.

There are various factors affecting individuals after surviving a stroke including residual effects [30,41,42,49,58,69,70,90], cognitive and psychologic changes [30,42,79], and the use of an assistive device [39] that can alter the correlation. The residual cognitive and physical changes are not the same for each person dependent upon severity and location of the stroke therefore the factors can create different barriers for each individual [30,41,42,49,58,69,70,79,90]. The use of an assistive device may alter the correlation because other outcome measures may not allow the use of an assistive device so they may demonstrate decreased performance on these measures. These functional barriers may or may not impact the mobility aspects that the TUG assesses so it is important to note the patients' specific limitations.

The construct validity could have been skewed due to decreased mental function impacting the ability to follow specific directions, maintain attention to complete a task, or ability to correctly sequence events [28,36]. Populations that typically have alterations in cognition that could cause this limitation include individuals with a stroke, children with DS or CP, and individuals with PD. Therefore, if the patient loses attention or forgets where they are going, time on the TUG may be increased and cause misrepresentation of their functional ability.

Overall, the results reflected data similar to individual systematic reviews on specific populations. The systematic reviews that assessed reliability and validity of the TUG in individuals with a stroke, SCI, or CP, all reported high values for the TUG, which is congruent with these findings [6,7,9]. The authors of the present review found conflicting evidence in the elderly population when comparing results assessing the utilization of the TUG as a fall risk predictor to previously published data [98]. The Rydwik et al. [98] review states that "… the TUG should not be used to discriminate between persons with high and low fall risk…" In comparison,

this systematic review found excellent reliability and validity in this population. The discrepancy may be due to the alternative focus on the predictive value of the TUG, while that was not the focal point of the current review. The focus of this review is to create a more comprehensive analysis for a variety of populations.

There were common strengths and weaknesses in the methodology of the included articles in this systematic review. A strength across many studies was that there were high sample sizes allowing for accurate representation of the specified population. The overall low risk of bias of each study indicated that they were sound in their methodological process. A weakness of the studies included could be attributed to study design due to many studies not using control groups to further validate their data.

There are multiple limitations to this systematic review. In relation to reliability limitations, minimal data were reported on inter-rater reliability compared to intra-rater and test-retest reliability. The TUG is an objective measure with specific instructions, and does not require advanced clinical experience, which could attribute to overall high reliability values. More limitations were noted in validity studies; one such limitation was in regard to the outcome measure of comparison. Many of the included studies used disease specific outcome measures that did not assess the same constructs. If they did not assess similar constructs, then validity data was misleading.

The predictive validity studies only assessed typical adults or children. A similar limitation was found in the lack of normative data on specific diagnoses and in typical individuals aged 9 to 38. An increase in research on normative data across all populations would allow for more research on predictive validity.

Another limitation included gaps within the literature for many populations. These gaps show a need for future research regarding the TUG in populations other than the individuals who are frail and elderly. This was evidenced by the lack of normative data for a variety of age groups such as 20–60 year olds who may also experience CNS insults or neurodegenerative diseases. In addition, this could be a useful tool for adolescence and adults in individuals with developmental disorders, such as CP and DS but no comparison of normative data exists. The increase in normative data would allow for more predictive validity, which could help develop the true clinical utility of the TUG across a patient's lifespan. Limitations in this systematic review's methodology include a risk for missing relevant articles or error in data extraction. Another limitation is a low kappa agreement between reviewers in regards to final inclusion. This low kappa value is due to being chance corrected, the present review examined many articles when determining inclusion and therefore had high agreement with certain categories that the kappa calculated determined was by chance.

Overall, the TUG is determined to have excellent reliability across all analyzed populations. In comparison, concurrent validity values had a broader range of results. This could be because the TUG shows stronger correlations with other dynamic balance tests rather than outcome measures assessing different aspects of function. When being compared to other activity focused outcome measures, in any population under study, the TUG proved to have high concurrent validity. The results of this systematic review indicate that the clinical utility of the TUG is useful across the continuum of care when assessing many populations, regardless of age.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Rehabilitation Measures. AbilityLab. 2018 [cited 2018 Feb 1]. Available from: https://www.sralab.org/rehabilitation-measures

[2] Wedge F, Braswell-Christy J, Brown C, et al. Factors influencing the use of outcome measures in physical therapy practice. Physiother Theory Pract. 2012;28(2):119–133.

[3] Alexandria V. 2018. Guide to physical therapist practice. Available from: http://guidetoptpractice.apta.org/

[4] World Health Organization. 2002. Towards a common language for functioning, disability and health ICF [PDF]. Geneva. Available from: https://www.who.int/classifications/icf/icfbeginnersguide.pdf

[5] Podsiadlo D, Richardson S. The Timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc. 1991;39(2):142–148.

[6] Pollock C, Eng J, Garland S. Clinical measurement of walking balance in people post stroke: a systematic review. Clin Rehabil. 2011;25(8):693–708.

[7] Nicolini-Panisson R, Donadio M. Timed "Up and Go" test in children and adolescents. Rev Paul Pediatr. 2013;31(3):377–383.

[8] Tyson S, Connell L. The psychometric properties and clinical utility of measures of walking and mobility in neurological conditions: a systematic review. Clin Rehabil. 2009;23(11):1018–1033.

[9] Lam T, Noonan V, Eng J. A systematic review of functional ambulation outcome measures in spinal cord injury. Spinal Cord. 2008; 46(4):246–254.

[10] About PROSPERO. PROSPERO. [cited 2018 Nov 13]. Available from: https://www.crd.york.ac.uk/PROSPERO/#aboutpage

[11] Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4:1.

[12] Kastner M, Wilczynski N, Walker-Dilks C, et al. Age-specific search strategies for Medline. J Med Internet Res. 2006;8(4):e25.

[13] Kappa as a Measure of Concordance in Categorical Sorting. Vassarstatsnet. 2018 [cited 2018 Feb 1]. Available from: http://vassarstats.net/kappa.html

[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159.

[15] Pluye P, Robert E, Cargo M, et al. 2011. Proposal: A mixed methods appraisal tool for systematic mixed studies reviews. [November 28]. Available from: http://mixedmethodsappraisaltoolpublic.pbworks.com. Archived by WebCite® at http://www.webcitation.org/5tTRTc9yJ

[16] Jewell D. Guide to evidence-based physical therapist practice. 4th ed. Burlington, MA: Jones & Bartlett Learning; 2017.

[17] Koo TK, Li MY. Erratum to "A guideline of selecting and reporting intraclass correlation coefficients for reliability research" [J Chiropr Med. 2016;15(2):155-163]. J Chiropr Med. 2016;16(4):346.

[18] Shong N. 2008. Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data [PDF]. Pittsburg: University of Pittsburg. Available from: http://d-scholarship.pitt.edu/8056/1/Chokns_etd2010.pdf

[19] Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 3rd ed. Philiadelphia (PA): F.A. Davis Company; 2015.

[20] Andresen E. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000;81:S15–S20.
[21] Aigner A, Curt A, Tanadini L, et al. Concurrent validity of single and groups of walking assessments following acute spinal cord injury. Spinal Cord. 2017;55(5):435–440.
[22] Alexandre T, Meira D, Rico N, et al. Accuracy of Timed Up and Go Test for screening risk of falls among community-dwelling elderly. Rev Bras Fisioter. 2012;16(5):381–388.
[23] Asano M, Miller W, Eng J. Development and psychometric properties of the ambulatory self-confidence questionnaire. Gerontology. 2007;53(6):373–381.
[24] Balasubramanian C, Boyette A, Wludyka P. How well do functional assessments of mobility and balance discriminate fallers and recurrent fallers from non-fallers among ambulatory older adults in the community? Physiother Canada. 2015;67(2):184–193.
[25] Bandong A, Madriaga G, Gorgon E. Reliability and validity of the four square step test for children with cerebral palsy and down syndrome. Res Dev Disabil. 2015;47:39–47.
[26] Belgen B, Beninato M, Sullivan P, et al. The association of balance capacity and falls self-efficacy with history of falling in community-dwelling people with chronic stroke. Arch Phys Med Rehabil. 2006;87(4):554–561.
[27] Bergström M, Lenholm E, Franzén E. Translation and validation of the Swedish version of the mini-BESTest in subjects with Parkinson's disease or stroke: a pilot study. Physiother Theory Pract. 2012;28(7):509–514.
[28] Besios T, Nikolaos A, Vassilios G, et al. Comparative reliability of the PEDI, GMFM and TUG Tests for children with cerebral palsy. J Phys Ther Sci. 2013;25(1):73–76.
[29] Bohannon R, DePasquale L. Physical Functioning Scale of the Short-Form (SF) 36: internal consistency and validity with older adults. J Geriatr Phys Ther. 2010;33(1):16–17.
[30] Bower K, McGinley J, Miller K, et al. Instrumented static and dynamic balance assessment after stroke using Wii Balance Boards: reliability and association with clinical tests. PLoS One. 2014;9(12):e115282.
[31] Busse M, Quinn L, Khalil H, et al. Optimising mobility outcome measures in Huntington's disease. J Huntington's Dis. 2014;3(2):175–188.
[32] Carey H, Martin K, Combs-Miller S, et al. Reliability and responsiveness of the Timed Up and Go Test in children with cerebral palsy. Pediatr Phys Ther. 2016;28(4):401–408.
[33] Chrysagis N, Skordilis E, Koutsouki D. Validity and clinical utility of functional assessments in children with cerebral palsy. Arch Phys Med Rehabil. 2014;95(2):369–374.
[34] Creel G, Light K, Thigpen M. Concurrent and construct validity of scores on the timed movement battery. Phys Ther. 2001;81(2):789–798.
[35] Dal Bello-Haas V, Klassen L, Sheppard M, et al. Psychometric properties of activity, self-efficacy, and quality-of-life measures in individuals with parkinson disease. Physiother Canada. 2011;63(1):47–57.
[36] de Campos A, Costa C, Rocha N. Measuring changes in functional mobility in children with mild cerebral palsy. Dev Neurorehabil. 2011;14(3):140–144.
[37] Desai A, Goodman V, Kapadia N, et al. Relationship between dynamic balance measures and functional performance in community-dwelling elderly people. Phys Ther. 2010;90(5):748–760.
[38] Desai A, Kulkarni V, Mishra N, et al. Correlation of balance performance and fear of fall in Parkinson's disease. Indian J Physiother Occup Ther. 2014;8(4):11–15.
[39] Engberg W, Lind A, Linder A, et al. Balance-related efficacy compared with balance function in patients with acute stroke. Physiother Theory Pract. 2008;24(2):105–111.
[40] Faria C, Teixeira-Salmela L, Neto M, et al. Performance-based tests in subjects with stroke: outcome scores, reliability and measurement errors. Clin Rehabil. 2012;26(5):460–469.
[41] Flansbjer U, Holmbäck A, Downham D, et al. Reliability of gait performance tests in men and women with hemiparesis after stroke. J Rehabil Med. 2005;37(2):75–82.
[42] Forsberg A, Nilsagård Y. Validity and reliability of the Swedish version of the activities-specific balance confidence scale in people with chronic stroke. Physiother Canada. 2013;62(2):141–147.
[43] Forsberg A, Andreasson M, Nilsagard Y. Validity of the dynamic gait index in people with multiple sclerosis. Phys Ther. 2013;93(10):1369–1376.
[44] Fritz N, Jiang A, Keller J, et al. Utility of the six-spot step test as a measure of walking performance in ambulatory individuals with multiple sclerosis. Arch Phys Med Rehabil. 2016;97(4):507–512.
[45] Gan S, Tung L, Tang Y, et al. Psychometric properties of functional balance assessment in children with cerebral palsy. Neurorehabil Neural Repair. 2008;22(6):745–753.
[46] Goh E, Chua S, Hong S, et al. Reliability and concurrent validity of four square step test scores in subjects with chronic stroke: a pilot study. Arch Phys Med Rehabil. 2013;94(7):1306–1311.
[47] Hatch J, Gill-Body K, Portney L. Determinants of balance confidence in community-dwelling elderly people. Phys Ther. 2003;83(12):1072–1079.
[48] Herman T, Giladi N, Hausdorff J. Properties of the 'Timed Up and Go' test: more than meets the eye. Gerontology. 2011;57(3):203–210.
[49] Hiengkaew V, Jitaree K, Chaiyawat P. Minimal detectable changes of the berg balance scale, Fugl-Meyer assessment scale, Timed "Up & Go" test, gait speeds, and 2-minute walk test in individuals with chronic stroke with different degrees of ankle plantarflexor tone. Arch Phys Med Rehabil. 2012;93(7):1201–1208.
[50] Huang S, Hsieh C, Wu R, et al. Minimal detectable change of the Timed "Up & Go" test and the dynamic gait index in people with Parkinson disease. Phys Ther. 2011;91(1):114–121.
[51] Iatridou G, Dionyssiotis Y. Reliability of balance evaluation in children with cerebral palsy. Hippokratia. 2013;17(4):303–306.
[52] Jette A, Jette D, Ng J, et al. Are performance-based measures sufficiently reliable for use in multicenter trials? J Gerontol A Biol Sci Med Sci. 1999;54(1):M3–M6.
[53] Johnston M, de Morton N, Harding K, et al. Measuring mobility in patients living in the community with Parkinson disease. NeuroRehabilitation. 2013;32(4):957–966.
[54] Jonsdottir J, Cattaneo D. Reliability and validity of the dynamic gait index in persons with chronic stroke. Arch Phys Med Rehabil. 2007;88(11):1410–1415.
[55] Kalron A, Givon U. Construct validity of the four square step test in multiple sclerosis. Arch Phys Med Rehabil. 2016;97(9):1496–1501.
[56] Katz-Leurer M, Rotem H, Keren O, et al. Balance abilities and gait characteristics in post-traumatic brain injury, cerebral palsy and typically developed children. Dev Neurorehabil. 2009;12(2):100–105.

[57] Katz-Leurer M, Rotem H, Lewitus H, et al. Functional balance tests for children with traumatic brain injury: within-session reliability. Pediatr Phys Ther. 2008;20(3):254–258.
[58] Kim J, Chu D, Jeon H. Reliability and validity of the L test in participants with chronic stroke. Physiotherapy. 2015; 101(2):161–165.
[59] Kobayashi E, Himuro N, Takahashi M. Clinical utility of the 6-min walk test for patients with moderate Parkinson's disease. Int J Rehabil Res. 2017;40(1):66–70.
[60] Kumban W, Amatachaya S, Emasithi A, et al. Five-times-sit-to-stand test in children with cerebral palsy: reliability and concurrent validity. NeuroRehabilitation. 2013;32(1):9–15.
[61] Landers M, Oscar S, Sasaoka J, et al. Balance confidence and fear of falling avoidance behavior are most predictive of falling in older adults: prospective analysis. Phys Ther. 2016;96(4):433–442.
[62] Learmonth Y, Paul L, McFadyen A, et al. Reliability and clinical significance of mobility and balance assessments in multiple sclerosis. Int J Rehabil Res. 2012;35(1):69–74.
[63] Lemay J, Nadeau S. Standing balance assessment in ASIA D paraplegic and tetraplegic participants: concurrent validity of the Berg Balance Scale. Spinal Cord. 2010;48(3): 245–250.
[64] Lyders Johansen K, Derby Stistrup R, Skibdal Schjøtt C, et al. Absolute and relative reliability of the Timed 'Up & Go' test and '30second chair-stand' test in hospitalised patients with stroke. PLoS One. 2016;11(10):e0165663.
[65] Mangione K, Craik R, McCormick AA, et al. Detectable changes in physical performance measures in elderly African Americans. Phys Ther. 2010;90(6):921–927.
[66] Martin K, Natarus M, Martin J, et al. Minimal detectable change for TUG and TUDS tests for children with Down syndrome. Pediatr Phys Ther. 2017;29(1):77–82.
[67] Morris S, Morris M, Iansek R. Reliability of measurements obtained with the Timed "Up & Go" test in people with Parkinson disease. Phys Ther. 2001;81(2):810–818.
[68] Nair B, Dobson A, O'Dea I, et al. Further validation of "Timed Up and Go" in stroke patients. Australas J Ageing. 1999;18(2):98–99.
[69] Ng S, Hui-Chan C. The Timed Up & Go test: its reliability and association with lower-limb impairments and locomotor capacities in people with chronic stroke. Arch Phys Med Rehabil. 2005;86(8):1641–1647.
[70] Ng S, Chan L, Chan CS, et al. Parallel walk test: its correlation with balance and motor functions in people with chronic stroke. Arch Phys Med Rehabil. 2015;96(5): 877–884.
[71] Nicolini-Panisson R, Donadio M. Normative values for the Timed 'Up and Go' test in children and adolescents and validation for individuals with Down syndrome. Dev Med Child Neurol. 2013;56(5):490–497.
[72] Nilsagard Y, Lundholm C, Gunnarsson L, et al. Clinical relevance using timed walk tests and 'Timed Up and Go' testing in persons with multiple sclerosis. Physiother Res Int. 2007;12(2):105–114.
[73] Nilsagård Y, Forsberg A. Psychometric properties of the Activities-Specific Balance Confidence Scale in persons 0–14 days and 3 months post stroke. Disabil Rehabil. 2011; 34(14):1186–1191.
[74] Pires-Oliveira D, Gil A, De Oliveira L, et al. Analysis of quality of life in elderly practitioners of physical activity and relationship with risk of falls. Man Ther Posturol Rehabil J. 2014;12:181.
[75] Poncumhak P, Saengsuwan J, Kamruecha W, et al. Reliability and validity of three functional tests in ambulatory patients with spinal cord injury. Spinal Cord. 2013; 51(3):214–217.
[76] Quinn L, Khalil H, et al. Reliability and minimal detectable change of physical performance measures in individuals with pre-manifest and manifest Huntington disease. Phys Ther. 2013;93(7):942–956.
[77] Rennie L, Dietrichs E, Moe-Nilssen R, et al. The validity of the Gait Variability Index for individuals with mild to moderate Parkinson's disease. Gait Posture. 2017;54:311–317.
[78] Russell M, Hill K, Blackberry I, et al. The reliability and predictive accuracy of the falls risk for older people in the community assessment (FROP-Com) tool. Age Ageing. 2008;37(6):634–639.
[79] Salbach N, Mayo N, Hanley J, et al. Psychometric evaluation of the original and Canadian French version of the activities-specific balance confidence scale among people with stroke. Arch Phys Med Rehabil. 2006;87(12): 1597–1604.
[80] Sanjivani N D, Prema A K. Intra-rater reliability of Timed 'Up and Go' test for children diagnosed with cerebral palsy. Int J Ther Rehabil. 2012;19(10):575–580.
[81] Schlenstedt C, Brombacher S, Hartwigsen G, et al. Comparing the Fullerton Advanced Balance Scale with the Mini-BESTest and Berg Balance Scale to assess postural control in patients with Parkinson disease. Arch Phys Med Rehabil. 2015;96(2):218–225.
[82] Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go test. Phys Ther. 2000;80(9): 896–903.
[83] Sosnoff J, Weikert M, Dlugonski D, et al. Quantifying gait impairment in multiple sclerosis using GAITRite^{TM} technology. Gait Posture. 2011;34(1):145–147.
[84] Spagnuolo D, Jürgensen S, Iwama Â, et al. Walking for the assessment of balance in healthy subjects older than 40 years. Gerontology. 2010;56(5):467–473.
[85] Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the Unified Parkinson Disease Rating Scale in people with Parkinsonism. Phys Ther. 2008;88(6):733–746.
[86] Steffen T, Hacker T, Mollinger L. Age- and gender-related test performance in community-dwelling elderly people: six-minute walk test, Berg Balance Scale, timed Up & Go test, and gait speeds. Phys Ther. 2002;82(2): 128–137.
[87] Stretton C, Latham N, Carter K, et al. Determinants of physical health in frail older people: the importance of self-efficacy. Clin Rehabil. 2006;20(4):357–366.
[88] Tanji H, Gruber-Baldini A, et al. A comparative study of physical performance measures in Parkinson's disease. Mov Disord. 2008;23(13):1897–1905.
[89] Tsang C, Liao L, Chung R, et al. Psychometric properties of the mini-balance evaluation systems test (Mini-BESTest) in community-dwelling individuals with chronic stroke. Phys Ther. 2013;93(8):1102–1115.
[90] Ursin M, Ihle-Hansen H, Fure B, et al. Balance and mobility in acute stroke: association with subgroups of stroke and socio-demographic characteristics. Eur J Physiother. 2014; 16(4):230–237.

[91] Van Hedel H, Wirz M, Dietz V. Assessing walking ability in subjects with spinal cord injury: validity and reliability of 3 walking tests. Arch Phys Med Rehabil. 2005;86(2): 190–196.

[92] Verheyden G, Kampshoff C, et al. Psychometric properties of 3 functional mobility tests for people with Parkinson disease. Phys Ther. 2014;94(2):230–239.

[93] Villamonte R, Vehrs P, Feland J, et al. Reliability of 16 balance tests in individuals with Down syndrome. Percept Mot Skills. 2010;111(2):530–542.

[94] Williams E, Carroll S, Reddihough D, et al. Investigation of the timed 'Up & Go' test in children. Dev Med Child Neurol. 2005;47(8):518–524.

[95] Winser S, Smith C, et al. Psychometric properties of a core set of measures of balance for people with cerebellar ataxia secondary to multiple sclerosis. Arch Phys Med Rehabil. 2017;98(2):270–276.

[96] Wong S, Yam M, Ng S. The Figure-of-Eight Walk test: reliability and associations with stroke-specific impairments. Disabil Rehabil. 2013;35(22):1896–1902.

[97] Wrisley D, Kumar N. Functional gait assessment: concurrent, discriminative, and predictive validity in community-dwelling older adults. Phys Ther. 2010;90(5):761–773.

[98] Rydwik E, Bergland A, Forsén L, et al. Psychometric properties of Timed Up and Go in elderly people: a systematic review. Phys Occup Ther Geriatr. 2011;29(2):102–125.