The Timed Up and Go Test in Children: Does Protocol Choice Matter? A Systematic Review

Evi Verbecque, PT, PhD; Kirsten Schepens, MS; Joke Theré, MS; Bénédicte Schepens, PT, PhD; Katrijn Klingels, PT, PhD; Ann Hallemans, PhD

Department of Rehabilitation Sciences and Physiotherapy (Drs Verbecque and Hallemans, Ms Schepens, and Mr Theré) and Multidisciplinary Motor Center Antwerp (Drs Verbecque and Hallemans), Faculty of Medicine and Health Sciences, University of Antwerp, Belgium; Laboratory of Physiology and Biomechanics of Locomotion (Dr Schepens), Institute of Neuroscience, Université catholique de Louvain, Louvain-la-Neuve, Belgium; Rehabilitation Research Center (Dr Klingels), Biomedical Research Institute, Hasselt University, Diepenbeek, Belgium; Department of Rehabilitation Sciences (Dr Klingels), KU Leuven, Leuven, Belgium.

Purpose: Results on reliability and normative data for the Timed Up and Go test (TUG) in children who are developing typically are systematically reviewed.
Summary of Key Points: Six different TUG protocols are presented for which normative data are available for ages 3 to 18 years. TUG time is consistent within and between raters and sessions and is influenced by age. The choice of protocol, self-selected versus fastest walking speed, and use of a motivational aspect and of the outcome calculation affect TUG time as well as its consistency within and between sessions.
Conclusions: A standard protocol for the TUG is lacking and should be developed with attention to reliability.
Recommendations for Clinical Practice: If the TUG is to be used as a screening tool for dynamic balance control, clinicians need to apply protocols that include fastest walking speed motivation. (Pediatr Phys Ther 2019;31:22–31)
Key words: children who are developing typically, reference values, reliability, “reproducibility of results” [mesh], TUG

INTRODUCTION

Balance control is a prerequisite for motor skills in children.[1-3] The identification of potentially underlying balance deficits is fundamental for therapy planning. After children have learned to maintain the upright standing position, they acquire motor skills such as walking, running, and jumping. These skills increase functional independence. These motor skills require dynamic balance control, referring to the child’s ability to maintain stability while moving from one base of support to the next.

The Timed Up and Go test (TUG) is a functional dynamic balance test. The TUG is a timed measure during which the child has to stand up from a chair, walk 3 m, turn around, walk back, and sit down. The TUG was developed to assess functional mobility and dynamic balance control in frail elderly people,[4] and used to screen for an increased risk of falling.[5] Because it is easily administered, practical, inexpensive, and does not require specific training, use of the TUG has been generalized to pediatrics to screen for dynamic balance control. In contrast to elderly people, the TUG in children can be used to assess the development of functional dynamic balance and to identify dynamic balance deficits that interfere with the acquisition of motor skills and may even induce motor delay. As the TUG addresses balance control during movements in sitting and bipedal postures, its task composition approximates a child’s daily tasks and therefore addresses a child’s developing functional independence.[6] However, if it is to be used as a screening tool, the TUG for children needs to be sensitive to age and related to the motor progression level of the child. Normative data are used to determine cutoff values. A review conducted in 2013 on the TUG in children suggested normative values for the test need to be established.[7] Since then, several authors have reported normative data for the TUG[6,8] but using different protocols and age groups.

Motor competence is influenced by age, sex, weight, socioeconomic status (SES), and ethnicity.[9,10] Balance control, similar to motor development, increases with increasing age. It can be

0898-5669/110/3101-0022
Pediatric Physical Therapy
Copyright © 2019 Academy of Pediatric Physical Therapy of the American Physical Therapy Association

Correspondence: Evi Verbecque, PT, PhD, Campus Drie Eiken, Universiteitsplein 1, 2610 Wilrijk (evi.verbecque@uantwerpen.be).

Kirsten Schepens and Joke Theré completed this work when they were master’s students in physical therapy in training at the University of Antwerp, Belgium; they have an academic bachelor’s degree in physical therapy obtained at the University of Antwerp, Belgium.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal’s Web site (www.pedpt.com).

The authors declare no conflicts of interest.

DOI: 10.1097/PEP.0000000000000558

hypothesized that TUG time is influenced by the same factors. Therefore, an overview of the available normative data and identification of the potential influence of age, sex, weight, SES, and ethnicity on these values is needed.

To determine whether a child deviates from the norm, *z* scores are used.[6,8,11] These scores include the number of standard deviations (SDs) the child's performance deviates from the normative mean and are based on the reliability interval of the data. This suggests that reliability analyses are crucial for establishing normative data. Investigators of the TUG have focused on assessing these properties in children with atypical development (eg, cerebral palsy,[12-14] traumatic brain injury,[15,16] and lower extremity sarcoma),[17] providing evidence for high test-retest, intrarater, and interrater reliability in children with various motor impairments (intraclass correlation coefficient [ICC] ≥ 0.85).[7,18] In children who are developing typically, test-retest, intra-, and interrater reliability varies between moderate and excellent (ICC ≥ 0.61).[6,12,19] The TUG's reliability data indicate that the standard error of measurement (SEM) is scant.[7,18] An update on reliability of the TUG for children who are developing typically could provide insights into the applicability and usefulness of reported normative data.

Several authors have adjusted the protocol for testing in a pediatric population, such as using a chair with or without[13] arm- and backrest,[12] barefoot walking,[14] walking with footwear,[15,16] or with orthotics.[15,16] In contrast to the original protocol by Podsiadlo and Richardson,[4] Williams et al[12] suggested that self-selected walking speed should be preferred over the fastest walking speed when assessing TUG in children. Moreover, to be sure that children understand the test instructions, most authors propose an explanation followed by a demonstration with verbal feedback during the test as necessary.[6,8,12] To improve the children's motivation, different tools are described in the literature such as a target on the wall the children need to touch or a Duplo brick they need to grab and transport.[6,8,12] Whether children are motivated may also influence the outcome. In the original protocol, the best of 3 trials was taken as the final result,[4] but research with the TUG has for example used an average of 2[15,16] or 3[12] trials. To screen for dynamic balance deficits, an overview of normative data is necessary and the protocol used for investigation is needed.

This study aims to provide an overview of the available normative TUG data for children. The following research questions guided this investigation:

- Which TUG protocols have been used in literature to establish normative data for children who are developing typically and are the protocols reliable?
- Which study sample characteristics influence TUG time in children who are developing typically?
- Does the applied protocol influence the available normative data?

METHODS

Protocol and Registration

This systematic review is written according to the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines.[19] The protocol is available at PROSPERO (registration number CRD42016053927) and is online (www.crd.york.ac.uk/prospero/).

Search

Relevant literature was extracted from the PubMed, Web of Science, and Science Direct databases, including Medline, Cochrane Database of Systematic Review, Cochrane Central Register of Controlled Trials, ISI Web of Knowledge, and Web of Science. The search was conducted on October 13, 2017, using the following keywords: (Children OR Minor OR Adolescents OR adolescence OR “Teens” OR “Teen” OR “Teenagers” OR “Teenager” OR “Youth” OR “Youths” OR Preschool Child OR Children, Preschool OR Preschool Children) AND (“Timed up and go” OR “Timed up & go” OR TUG OR TGUGT OR “Timed Get up and go” OR “timed get up & go” OR “Timed get up and go test” OR “Timed get up & go test” OR “Get Up and Go test” OR “get up & go test” OR “Get up and go” OR “get up & go”). The search details were used to define the query in Web of Science and Science Direct. Mesh terminology was used in PubMed. No limits or filters were used. The search query was defined by 4 investigators.

Study Selection

Relevant studies were identified using predefined selection criteria according to the Population Intervention Comparison Outcome Study Design method. Original studies (S), full and brief reports with transparent methods, that reported normative data (O) for the TUG (I) in children who are developing typically 18 years or younger (P) and were written in Dutch, French, English, and German were included. All types of reviews, meta-analyses, conference proceedings, abstract only, and unpublished studies were not included. The selection criteria were applied in the following sequence: population, intervention, outcome, study design, and language. Two investigators assessed these criteria independently in 2 phases: phase 1, title and abstract; phase 2, full text. In case of disagreement, a third investigator's opinion was decisive. References from the included articles were additional articles.[19]

Risk of Bias

Risk of bias in studies reporting reliability data was assessed using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN). The COSMIN assesses risk of bias of studies investigating the psychometric properties of assessment tools.[20] The COSMIN checklist contains 1 box for each of the 9 defined psychometric properties (eg, reliability and measurement error). Each box comprises questions that can be answered by “excellent,” “good,” “fair,” or “poor.” The final score of a box is determined by its lowest score on an individual question.

For this review, only box B (relative reliability, ie, consistency of values) and box C (absolute reliability, measurement error) were relevant. For both boxes, 2 items were omitted, as the TUG does not necessarily require independent measures to be reliable (item 5) and the study population comprises children who are developing typically indicating that motor abilities are stable within a short interval (item 7). Each article was assessed independently by 2 investigators, and after a consensus meeting, a final score was assigned. Interrater reliability was determined using the Cohen's κ measure agreement between 2 raters (k).

Risk of bias in studies reporting normative data was not assessed, as adequate tools are not available. To provide insights into how the sample was selected, the nonresponders' rate was acknowledged and typical development was ascertained, these characteristics were mapped, based on the "selection" category of the Newcastle-Ottawa Scale adapted for cross-sectional studies.[21]

Data Extraction

The following data were extracted where available:

- Population-specific characteristics: number of children, mean age (and SD), age range, and male-female ratio.
- Specifics on the applied TUG protocol: instructions given to the subject (self-selected walking speed vs walking as fast as possible), when timing started (child gets up, start/go cue), type of motivation (none, touch object, grab and transport object), footwear (barefoot, shoes), and TUG outcome (best performance vs averaging trials and the number of trials included for analysis).
- For reliability analyses: TUG values (mean and SD), ICCs and the applied model, SEM, and minimal detectable change (MDC) were extracted. The ICC values were interpreted as follows: poor (ICC < 0.5), moderate ($0.5 \leq$ ICC < 0.75), or good (ICC ≥ 0.75).[22] When the raw TUG values were provided, but the SEM was not reported, this was calculated with the following formula: $\text{SEM} = \text{SD}_{\text{test1}} \times \sqrt{(1 - \text{ICC})}$. Subsequently, the MDC_{95} was calculated: $\text{MDC}_{95} = \text{SEM} \times 1.96 \times \sqrt{2}$.[22]
- For normative data: Raw TUG time values (mean and SD) were extracted from available literature and classified according to the age under investigation and the applied protocol. Based on the SD of the mean, z scores were calculated and used as cutoff values. Because higher TUG values represent poorer balance control, $+1z$ can be interpreted as "at risk for deviant dynamic balance control" and $+2z$ as "highly likely to have deviant dynamic balance control". The *mean TUG*, *mean TUG + 1 SD*, and *mean TUG + 2 SD* were presented graphically as a function of age.

All data were extracted by 2 independent investigators and compared in a consensus meeting.

Level of Evidence

The level of evidence (strong, moderate, limited, unknown, and conflicting) for the TUG's reliability was based on the number of studies, the methodological quality (determined with the COSMIN checklist), and the consistency of findings. The level of evidence was determined according to the criteria of the Cochrane Collaboration Back Review Group[23] and Saether et al.[24] No level of evidence was assigned for the available normative data since there is no validated measure for assessing risk of bias in these studies.

RESULTS

Study Selection

The search query had 293 hits in PubMed, 230 hits in Web of Science, and 204 hits in Science Direct, of which 616 were unique. After screening, 5 studies[6,8,12,25,26] met the criteria. One study[27] was added after reference screening, resulting in 6 studies used for data extraction (Figure 1).

Risk of Bias

Relative reliability was assessed in 5 of the studies with method quality varying between poor and excellent (Table 1). The main reason for poor quality was a small sample size (<30 children). There was high agreement between the 2 raters ($k = 0.769$).

Methods for sample selection for studies to establish normative data are shown in Table 2. All studies used (partial[27]) convenience sampling, of which two[6,25] calculated the minimum sample size in advance. Three studies reported the nonresponders rate.[6,8,26] Typical development of the included children was ascertained mainly by investigating the (parent-reported) medical history.[6,8,12,25,26]

Population Characteristics

The TUG was administered to a total of 2626 children who are developing typically between ages 3 and 18 years, of which 1212 were boys (46%; range 42%[26]-54%[12]). The children were recruited in Australia (Melbourne),[12] Belgium,[8] South Brazil,[6] Pakistan,[27] and the United States (Connecticut[26] and New York[25]).

TUG

Protocols. Six different TUG protocols were used. Half of the protocols consisted of the specific instruction for the children to walk as fast as possible[6,8,27] (Figure 2), whereas the others allowed self-selected walking speed[12,25,26] (Figure 3). Additional motivation was provided using a star on the wall the children needed to touch[6,12,25] or a Duplo brick the children needed to grab and transport.[8] In 2 studies, timing was started when the child got up from the chair[12,25] whereas in the other 4 studies a specific cue was used (go/start).[6,8,26,27] Children were assessed either barefoot[6,8,27] or with shoes.[12,25,26] The TUG outcome varied between an average of 2 trials,[25-27] average of 3 trials,[12] or the best of 3 trials.[6,8]

Reliability of Protocols for Establishing Normative Data. Reliability results are sown in Table 3.

- *Intrarater (within session) reliability.* Intrarater reliability was good across studies, with mean ICC values varying between 0.80 and 0.998.[6,12,25,27] Williams et al[12] reported the SEM of 0.6 and 0.4, respectively, for the baseline assessment and the TUG retest 10 to 20 minutes after the first test session in 3- to 9-year-old children.
- *Interrater reliability.* Three studies investigated interrater reliability[12,26,27] and reported high ICC values (>0.9).[12,26] Habib et al[27] reported high percentages of

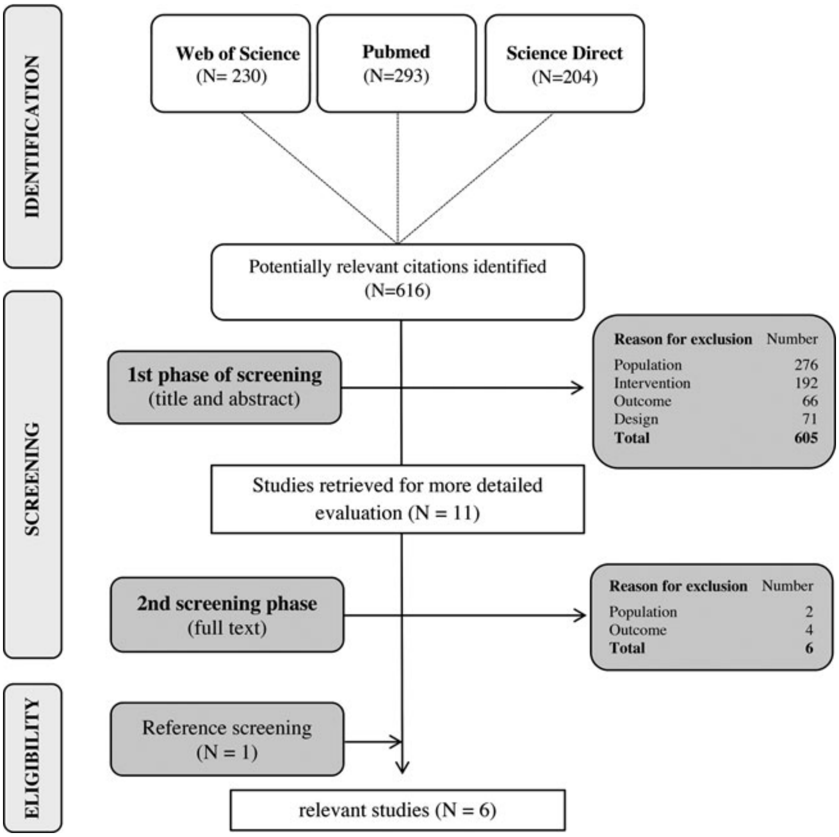
IDENTIFICATION
Web of Science (N= 230)
Pubmed (N=293)
Science Direct (N=204)
Potentially relevant citations identified (N=616)
SCREENING
1st phase of screening (title and abstract)
Reason for exclusion Number
Population 276
Intervention 192
Outcome 66
Design 71
Total 605
Studies retrieved for more detailed evaluation (N = 11)
2nd screening phase (full text)
Reason for exclusion Number
Population 2
Outcome 4
Total 6
ELIGIBILITY
Reference screening (N = 1)
relevant studies (N = 6)

Fig. 1. Flowchart of the selection process for studies.

agreement between raters (95%-100%). None of the studies reported SEM, nor was it calculable.

- *Test-retest reliability.* Mean ICC values between test and retest sessions were moderate to good, depending on the chronological age band under investigation. In a study sample of 3- to 18-year-olds[6] and 3- to 9-year-old children,[12] TUG time is consistent (ICC = 0.80-0.95), regardless of the test moment (1-2 hours after the first test or 1 week afterward).[6,12] But when children were divided into younger (3-5 years) and older groups (5-9 years), test-retest reliability was more variable (ICC = 0.61-0.83).[12] Younger children tended to have more reliable results compared with older children when performing the test 10 to 20 minutes after the first test (3-5 years: ICC = 0.82; 5-9 years: ICC = 0.76), whereas in older children test-retest reliability was better when assessed 1 week after the first test session (3-5 years: ICC = 0.61; 5-9 years: ICC = 0.83).[12] The SEM and the

TABLE 1

Risk of Bias

Author	COSMIN	Type of Reliability	Rater A	Rater B	Consensus Score	Reason for a Consensus Rating Less Than Excellent
Butz et al[25]	Box B[a]	Intrarater	Poor	Poor	Poor	Sample size <30
		Interrater	Poor	Poor	Poor	Sample size <30
Habib et al[27]	Box B	Intrarater	Poor	Poor	Poor	Sample size <30
		Interrater	Poor	Poor	Poor	Sample size <30, statistical method (percentage of agreement)
Itzkowitz et al[26]	Box B	Interrater	Poor	Poor	Poor	Sample size <30
Nicolini-Panisson and Donadio[6]	Box B	Intrarater	Excellent	Excellent	Excellent	
		Test-retest	Excellent	Excellent	Excellent	
	Box C[b]	Test-retest	Excellent	Poor	Excellent	
Williams et al[12]	Box B	Intrarater	Good	Good	Good	The applied ICC model was not reported
		Test-retest	Good	Good	Good	The applied ICC model was not reported
	Box C	Intrarater	Excellent	Excellent	Excellent	

Abbreviations: COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; ICC, interclass correlation coefficient.

[a]Box B: relative reliability.

[b]Box C: measurement error.

TABLE 2

Characteristics of Method for Sample Selection to Report Normative Data

Authors	Sample Selection	Sample Size Calculation	Nonresponders Rate	Ascertainment of Typical Development	Sample Size and Age Bands
Butz et al[25]	Sample of convenience (5-12 y old); elementary schools Connecticut, private schools in West Haven, outpatient rehabilitation Connecticut	Effect size between 0.3 and 0.4, power 0.8-0.99: 100 children		Medical history: (1) absence of neurological or orthopedic diagnoses, (2) no history of developmental delay or balance impairments, and (3) no orthopedic surgical procedures within the past 6 mo	160 children, 8 age bands according to chronological age
Habib et al[27]	Partial random and partial convenience sampling (5-13 y old); 2 private schools (random), 4 orphanages and 1 school from Malir (convenience) in Pakistan			Physical examination: (1) upper and lower extremity strength and flexibility, (2) spinal flexibility, and (3) coordination	180 children, 9 age bands according to chronological age
Itzkowitz et al[26]	Sample of convenience (5-17 y old); 20 public elementary and middle schools from 5 New York City boroughs		18 231 invitation letters; 1653 responders of which 1481 completed the TUG	Medical history: (1) no orthopedic surgical procedures or injuries within the past 6 m, (2) no history of neurological disorders, and (3) no individualized educational program	1481 children, 9 age bands according to chronological age
Nicolini-Panisson and Donadio[6]	Sample of convenience (3-18 y old); 5 schools in South Brazil	Sample size calculation based on 50 children for multiple regression analysis: power of 90%, minimum coefficient of determination of 0.22 and significance level of 0.05	598 questionnaires on health and consent forms delivered to the participating schools; 520 responders	Medical history through parental questionnaire: (1) no fracture or who had undergone surgery of the lower limbs less than 6 mo previously, (2) cardiorespiratory and neuromuscular diseases, or intellectual disability, and (3) incorrect performance of the test	459 children, 4 age groups: 3-5, 6-9, 10-13, and 14-18 y
Verbecque et al[8]	Sample of convenience (3-5 y old); 3 schools in Belgium		400 invitation letters; 192 responders	Medical history through parental questionnaire: (1) no developmental or neuromotor disorder, (2) no severe visual or hearing impairment, (3) no use of aids (except for glasses), (4) no cochlear implants, and (5) cooperative in performing 3 trials	172 children, 3 age bands according to chronological age
Williams et al[12]	Sample of convenience (3-9 y old); nearby schools, kindergartens, childcare centers in Melbourne				176 children; 2 age groups: 3-5 and 5-9 y

Abbreviation: TUG, Timed Up and Go.

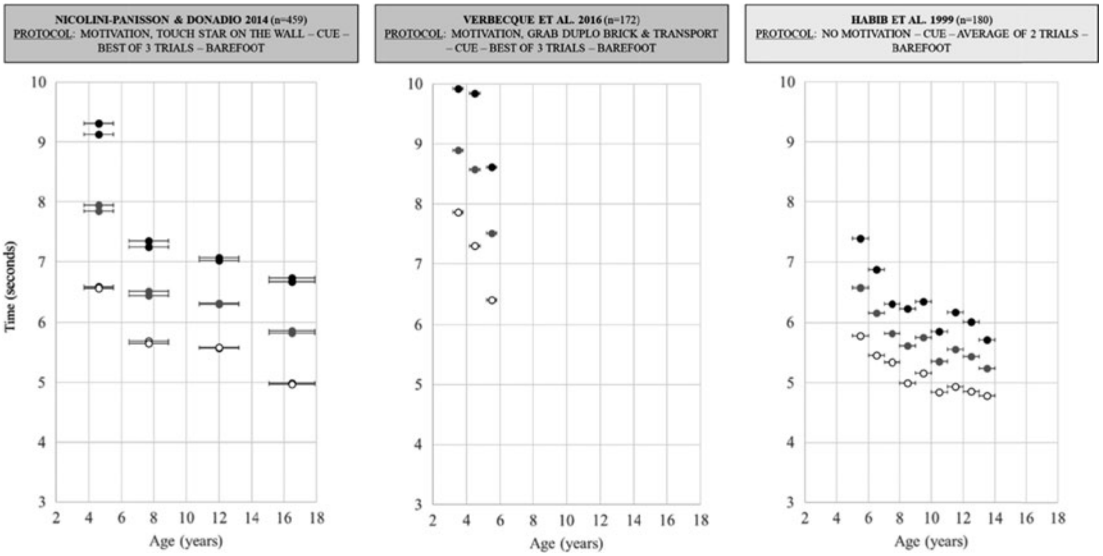

Fig. 2. Overview of the applied protocols requiring fastest walking performance with corresponding normative data as a function of chronological age (groups). Mean TUG values + 2 SDs (●); mean TUG values + 1 SD (●); mean TUG values (○); and horizontal error bars represent 1 SD from the mean age. For Nicolini-Panisson and Donadio,[6] TUG 1 and TUG 2 are in Table 3.

MDC were not reported, but were calculated. The SEM varied between 0.33 and 0.75 seconds depending on the age group under (Table 3).

Normative Data: Influence of Study Sample Characteristics and Protocols. In Figures 2 and 3, normative data for the TUG in children are presented as a function of age and applied protocol. Numeric values per protocol and age band are listed in Supplemental Digital Content 1 (available at: http://links.lww.com/PPT/A232). Four studies[8,25-27] reported numeric values for chronological age groups, whereas 2 studies[6,12] reported age bands combining several chronological ages.

In 2 studies, significant differences in TUG time between boys and girls are reported.[26,27] In these studies no motivational aspects were added to the protocol. Habib et al[27] found an overall better performance for boys compared with girls regardless of age, whereas Itzkowitz et al[26] showed that only 8-, 9- and 11-year-old boys performed better than girls. In all other studies when motivational aspects were added to the protocol, sex did not affect TUG time.

Several authors investigated predictors for TUG time based on study sample characteristics. Age accounted for 24.3%[25] to 49.0%[26] of the variance in TUG time in samples of children in the United States, when allowing self-selected walking speed. In children from south Brazil, age and weight accounted for 25%[6] of the variance in TUG time (fastest performance), whereas for children who were preschool age from Belgium, ethnicity

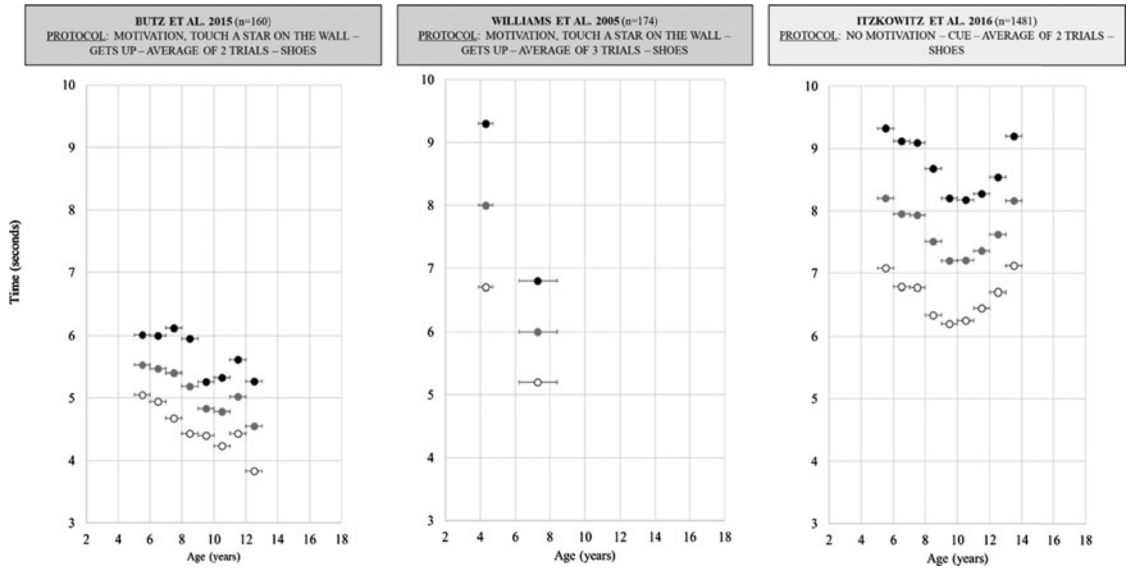

Fig. 3. Protocols using *self-selected walking performance* with corresponding normative data as a function of age groups. Mean TUG values + 2 SDs (●); mean TUG values + 1 SD (●); mean TUG values (○); and horizontal error bars represent 1 SD from the mean age. SD indicates standard deviation; TUG, Timed Up and Go.

TABLE 3

Intrarater and Test-Retest Reliability of the TUG Test Protocols Used for Normative Data

				TUG, s						
	Author	n	Age Range, y	Test Session	Trial 1 Mean (SD)	Trial 2 Mean (SD)	Trial 3 Mean (SD)	ICC Mean (95% CI)	SEM	MDC$_{95}$
Intrarater (within session) reliability	Butz et al[25]	10	5-12	Test 1				0.998[a]		
	Habib et al[27]	180	5-13	Test 1				0.81[b]		
	Nicolini-Panisson and Donadio[6]	459	3-18	Test 1				0.93[c]		
		459	3-18	Retest same day				0.94[c]		
		178	3-18	Retest 1 wk after test 1				0.95[c]		
	Verbecque et al[8]	172	3-5	Test 1	8.24[d] (1.97)	7.92[d] (1.72)	7.58[d] (1.60)			
	Williams et al[12]	176	3-9	Test 1 (baseline)	6.0 (1.5)	5.9 (1.3)	5.9 (1.3)	0.8[e] (0.75-0.84)	0.6	1.86[f]
		173	3-9	10-20 min after test 1	5.9 (1.5)	5.9 (1.5)	5.9 (1.5)	0.89[e] (0.86-0.92)	0.4	1.38[f]
		151	3-9	1 wk after test 1	5.7 (1.2)	5.8 (1.2)	5.7 (1.2)	0.85[e] (0.81-0.89)	0.46[f]	1.29[f]

				TUG, s					
	Author	n	Age Range, y	Test Session	Rater 1 Mean (SD)	Rater 2 Mean (SD)	ICC Mean (95% CI)	SEM	MDC$_{95}$
Interrater reliability	Butz et al[25]	10		Test 1			0.999[g]		
	Itzkowitz et al[26]	22		Test 1			0.988[b]		

				TUG, s							
				Measurement Characteristics							
	Author	n	Age Range, y	Live/video	Duration Between Tests	Trial Used for Analysis	Session 1 Mean (SD)	Session 2 Mean (SD)	ICC Mean (95% CI)	SEM	MDC$_{95}$
Test-retest reliability	Nicolini-Panisson and Donadio[6]	178	3-18	Live	1 wk	Best of 3			0.95[c]		
		459	3-18	Live	1-2 h	Best of 3			0.95[c]		
	Williams et al[12]	173	3-9	Live	1 wk	Average of 3	5.90 (1.3)	5.70 (1.1)	0.83[h] (0.77-0.88)	0.54[f]	1.49[f]
		83	3-5	Live	1 wk	Average of 3	6.7 (1.2)	6.50 (1.0)	0.61[h] (0.39-0.75)	0.75[f]	2.08[f]
		90	5-9	Live	1 wk	Average of 3	5.2 (0.8)	5.0 (0.8)	0.83[h] (0.73-0.89)	0.33[f]	0.91[f]
		173	3-9	Live	10-20 min	Average of 3	5.9 (1.3)	5.9 (1.5)	0.89[h] (0.86-0.92)	0.43[f]	1.20[f]
		83	3-5	Live	10-20 min	Average of 3	6.7 (1.2)	7.0 (1.3)	0.82[h] (0.72-0.88)	0.51[f]	1.41[f]
		90	5-9	Live	10-20 min	Average of 3	5.2 (0.8)	4.9 (0.8)	0.76[h] (0.61-0.85)	0.39[f]	1.09[f]

Abbreviation: CI, confidence interval; ICC, intraclass correlation coefficient; MDC$_{95}$, minimal detectable change = SEM × 1.96 × $2^{1/2}$; TUG, Timed Up and Go; SD, standard deviation; SEM, standard error of measurement = $SD_{\text{test 1}} \times (1 - ICC)^{1/2}$.

[a] Two-way mixed effects, consistency, multiple raters/measurements, ICC (3,2).

[b] Two-way mixed effects, consistency, single rater/measurement, ICC (3,1).

[c] Not reported.

[d] Values differ significantly.

[e] One-way random effects, absolute agreement, single rater, ICC (1,1).

[f] Values have been calculated.

[g] Two-way random effects, absolute agreement, single rater/measurement, ICC (2,1).

[h] One-way random effects, absolute agreement, multiple raters/measurements, ICC (1,3).

explained 28%[8] of the variance in TUG time (fastest performance). Several authors reported that body mass index[6,26] and body height[6,8,25] did not account for the variance in TUG time.

There are differences in the normative data depending on the protocol (Figures 2 and 3). Significant differences between age groups are reported. When no motivation was used, differences between age groups were dependent on the required walking speed, which resulted in the composition of different age bands. When performing the TUG as fast as possible (Figure 2), significant differences in TUG time were found between 3 age bands—5 to 7-year-olds, 8- to 10-year-olds, and 11- to 13-year-olds,[27] whereas, when using self-selected walking speed instruction, the ages in the bands changed into 5- to 7-year-olds, 8- to 11-year-olds, and 12- to 13-year-olds.[26] When motivation was used and the TUG was performed at self-selected walking speed, children of preschool age (3-years) performed the TUG significantly slower than children who were older (5-9 years).[12] When children of preschool age performed the TUG with motivation as fast as possible, significant differences between these 3 chronological age groups were identified.[8] When SES was taken into account, boys from Pakistan with low SES performed significantly better on the TUG than girls, but when compared with girls with high SES, girls with low SES performed poorer and boys with high SES performed poorer than boys with low SES.[27]

Level of Evidence

The level of evidence and how it was obtained for reliability of the TUG are shown in Supplemental Digital Content 2A (relative reliability) and 2B (absolute reliability) (available at: http://links.lww.com/PPT/A233). Strong evidence was found for relative and absolute intrarater (within session) and test-retest reliability of the TUG protocol by Williams et al,[12] consisting of self-selected walking speed with a motivational aspect and averaging 3 trials.[12] Moderate evidence was found for relative and strong evidence for absolute intrarater and test-retest reliability for the protocol by Nicolini-Panisson and Donadio,[6] consisting of fastest walking speed with motivational aspects and the best of 3 trials.

DISCUSSION

The aim of this systematic literature review was to provide an overview of the reliability and available normative data in children for the TUG, a screening tool for dynamic balance control. Six different protocols were identified. Consistency of TUG time is moderate to good, with a measurement error below 1 second. Age influences TUG performance, but other predictors such as the applied protocol also influence performance.

Reliability of the TUG Protocols

Reliability analyses on TUG protocols (used for reporting normative data) remain incomplete, especially when protocol differences are considered. Mainly intrarater (within session) reliability has been investigated.[6,12,25,27] Thus, the body of evidence regarding the reliability of the TUG should be interpreted with caution. All 6 articles included in this review report a different protocol, which limits the generalizability of results. Moreover, most studies were rated as poor due to small sample sizes[25-27] or the applied statistical technique to assess consistency between raters,[27] implying that due to method shortcomings, reliability results need to be interpreted with caution. Nevertheless, the reported ICC values were high (ICC $\geq$ 0.8) for all types of reliability, indicating that strong agreement exists between the administered trials, raters, and/or sessions.[6,12,25-27]

Younger children (3-5 years) tend to have more consistent results over a shorter interval and less consistency over a longer interval compared with older children (5-9 years).[12] All children were considered to be stable during a short interval (maximum 2 weeks), as no changes in their motor progression are to be expected. However, ICC values seem to be affected by age and the interval between the test sessions. A presumable explanation is that gait in children younger than 7 years is developing toward a mature gait pattern.[13] Because of large intravariability in their developing motor patterns, performances on the TUG are more likely to differ from each other, which can be reflected in lower ICC values. Cognitive functions such as attention and concentration may play a role, particularly in younger children. Especially when self-selected walking speed is allowed, these cognitive functions can interfere with the children's performance. Williams et al[12] did not provide instructions on walking speed, which might have induced more variance in the preschoolers' performances and thus in TUG time. Similar to research in adults and elderly people,[28] these findings suggest that fastest walking speed should be preferred over self-selected walking speed, but this needs to be confirmed in future research.

When a shorter interval was introduced between sessions (eg, 10-20 minutes),[12] less variance in ICC, SEM and MDC values was observed, suggesting practice/learning effects occur (eg, recall of task instructions). Such practice effects were found within a session for the fastest walking speed protocol shown by a decrease in TUG time in preschool children.[8] This may be the same for assessment between sessions with short intervals, especially in younger children.

Normative Data

To screen balance deficits in children, normative data and corresponding cutoff values are needed. Because of the ongoing development and maturation of balance control during walking, it is expected that increasing age results in better TUG performance in children who are developing typically, and thus a descending trend of TUG time as a function of age. Several authors suggest that variance in TUG time is explained by age[6,8,25,26] and that significant differences between specific age groups exist.[12,27] Normative values have been reported for different age bands. In 2 studies, reference values were reported by chronological age,[8,25] whereas most authors grouped several chronological ages into 1 age band (eg, age 5-9 years,[12] or 6-9 years,[6] or 5-7 years).[26,27] Only Habib et al[27] and Itzkowitz et al[26] found significant differences by chronological age. They found 1 identical age band, 5- to 7-year-old children,[26,27] whereas age bands in older children tended to differ, 8- to 10-year-olds[27] versus 8- to 11-year-olds[26] and 11- to 13-year-olds[27] versus 12- to 13-year-olds.[26] A potential explanation for these age band differences might be the samples. Habib et al[27] had a smaller sample size in each subgroup (approximately 20), but they were equally distributed over the chronological age bands, which was not the sample distribution in the study by Itzkowitz et al[26] (sample size varies between 45 and 244 per subgroup). The composition of the sample may have influenced TUG results (Table 2). Children were recruited from different countries, with the potential influence of cultural differences.[6,12] For example, poorer performance of girls from Pakistan compared with boys with a low SES was assigned to cultural influences, as the girls often wore a chador, limiting their mobility.[27] Itzkowitz et al[26] were the only other author to find sex-related differences and both these studies[26,27] lack the "motivational aspect" during TUG administration. Sex-related differences might be a result of a lack of motivational cues. Bardid et al[29] stated that gender differences before puberty have been associated with a child's perception of their appropriate gender role with regard to sports and games. Therefore, boys might be more motivated to perform gross motor skills through sports. This again suggests that the protocol influences performance. Based on the presentation of the normative data as a function of the protocol (Figures 2 and 3), the applied protocol interferes with normative data. The expected trend of decreasing TUG time with increasing age is seen with a protocol that demands fastest walking speed with motivation (touching/grabbing and transporting an object).[6,8] When no additional motivation is provided during fastest walking speed,[8] or when self-selected walking speed with[25] or without[26] motivation is allowed, TUG time becomes more variable. With these protocols, fluctuations in TUG time between chronological age groups are observed: 11-year-old children perform poorer than 12-year-old children but also poorer than 10-year-old children.[26,27] When the TUG is to be used as a screening tool for dynamic balance control, fluctuations should be limited.

The number of trials used also influences the normative data. When protocols use the best of 3 performances or an average of 3 trials, a decreasing trend of TUG time with increasing age is observed.[6,8,12] The best of 3 trials provides information on the best performance, whereas averaging trials has the advantage of taking the intraindividual variability of performances into account. In children of preschool age, walking as fast as possible, 3 trials within 1 session differed, highlighting the need for using best performance, but also the need to determine whether 3 TUG trials within 1 session are sufficient.[8] None of the 5 studies that investigated reliability reported within session differences. According to Podsiadlo and Richardson,[4] the best of 3 trials should be used as the final result. However, it remains to be determined how many trials are necessary for children, taking the children's developmental progression into account. Again, this highlights the need for more thorough reliability analyses of the TUG protocols with attention to age effects and number of trials.

Thus, both protocol differences and method characteristics used to select the sample may influence fluctuations in TUG time. Although differences in both were identified, they were not assessed on their potential risk of bias, resulting in a limited body of evidence on the best protocol for the TUG and the corresponding normative data to use in clinical practice.

Based on current knowledge[4] and our experience (unpublished observations), fastest walking speed, the use of an additional motivational aspect, and best performance, at least 3 trials should be preferred in pediatric rehabilitation. These protocol characteristics motivate the child to provide his/her best performance, thereby approximating real-life, self-induced movements driven by motivation and attention but assessed in a standardized and reliable manner.

Limitations of the Study

To identify risk of bias in individual studies addressing reliability, we used the COSMIN checklist, a validated tool. However, no such scales are currently available to address risk of bias in studies investigating normative data. Although the selection subscale of the Newcastle-Ottawa Scale has not been validated and was not designed to address risk of bias in studies investigating normative data, it provides valuable information on features of the sample selection process and it was therefore used in the present study. However, because of the lack of a risk of bias assessment, the body of evidence regarding normative data remains limited. The suggestion for a most suitable protocol, such as fastest walking speed, use of an additional motivational aspect, best performance, and use of at least 3 trials, needs to be tested on a larger sample.

Several authors suggested that cultural differences may affect TUG time.[6,12,27] The impact of cultural influences on normative data for the TUG remains unclear because sample characteristics such as weight,[6,25] body height,[6,8,25] leg length,[6] SES,[27] race, or ethnicity[6,8] were not consistently reported. Most studies used convenience samples, thereby increasing the risk of selection bias, which highlights the need for random sampling in future research.

Five of 6 relevant studies were retrieved using 3 main databases such as PubMed, Science Direct, and Web of Science, but hand searching was added after full-text screening, acknowledging the weakness of systematic search queries to possibly miss relevant literature.[19] Finally, only studies published in English, French, German, and Dutch were included. As none were excluded based on language, indicating that although language restrictions were defined prior to conducting the systematic review, multiple languages were included in the results.

CONCLUSIONS AND IMPLICATIONS FOR RESEARCH AND CLINICAL PRACTICE

Although widely used in clinical practice to assess dynamic balance control, large variety in TUG protocols exists, which influences validity of normative data. Investigators have changed the TUG protocol without investigating its impact on both reliability and normative data. This review suggests that the protocol may affect TUG time and variance in reliability measures. Future research needs to determine which protocol is most reliable and therefore most suitable to screen for deficits in dynamic balance control in clinical practice.

If the TUG is to be used as a screening tool for deficits in dynamic balance control, a standard protocol needs to be developed and its psychometric properties such as reliability, validity, responsiveness, sensitivity, and specificity need to be investigated. Based on the results of this review, we recommend fastest walking speed, the use of an additional motivational aspect, best performance, and administration of at least 3 trials within 1 session. However, the results in the present review are to be interpreted cautiously, as they are based on only 6 studies that used different protocols, included different sample sizes and sample compositions, which limits their generalizability. Moreover, when establishing normative data, validated developmental motor scales should be used.

REFERENCES

1. Huxham FE, Goldie PA, Patla AE. Theoretical considerations in balance assessment. *Aust J Physiother*. 2001;47(2):89-100.
2. Massion J. Postural control system. *Curr Opin Neurobiol*. 1994;4(6):877-887.
3. Shumway-Cook A, Woollacott MH. *Motor Control, Translating Research Into Clinical Practice*. 4th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2014:161-193.
4. 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.
5. Park SH. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. *Aging Clin Exp Res*. 2018;30(1):1-16.
6. Nicolini-Panisson RD, Donadio MV. Normative values for the Timed "Up and Go" test in children and adolescents and validation for individuals with Down syndrome. *Dev Med Child Neurol*. 2014;56(5):490-497.
7. Nicolini-Panisson RD, Donadio MV. Timed "Up & Go" test in children and adolescents. *Rev Paul Pediatr*. 2013;31(3):377-383.
8. Verbecque E, Vereeck L, Boudewyns A, et al. A modified version of the Timed Up and Go Test for children who are preschoolers. *Pediatr Phys Ther*. 2016;28(4):409-415.
9. Barnett LM, Lai SK, Veldman SLC, et al. Correlates of gross motor competence in children and adolescents: a systematic review and meta-analysis. *Sports Med*. 2016;46(11):1663-1688.
10. Mendonça B, Sargent B, Fetters L. Cross-cultural validity of standardized motor development screening and assessment tools: a systematic review. *Dev Med Child Neurol*. 2016;58(12):1213-1222.
11. Norris RA, Wilder E, Norton J. The functional reach test in 3- to 5-year-old children without disabilities. *Pediatr Phys Ther*. 2008;20(1):47-52.
12. Williams EN, Carroll SG, Reddihough DS, et al. Investigation of the Timed "Up & Go" Test in children. *Dev Med Child Neurol*. 2005;47(8):518-524.
13. Gan SM, Tung LC, Tang YH, et al. Psychometric properties of functional balance assessment in children with cerebral palsy. *Neurorehabil Neural Repair*. 2008;22(6):745-753.
14. Zaino CA, Marchese VG, Westcott SL. Timed up and down stairs test: preliminary reliability and validity of a new measure of functional mobility. *Pediatr Phys Ther*. 2004;16(2):90-98.
15. 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.
16. Katz-Leurer M, Rotem H, Lewitus H, et al. Relationship between balance abilities and gait characteristics in children with post-traumatic brain injury. *Brain Inj*. 2008;22(2):153-159.
17. Marchese VG, Spearing E, Callaway L, et al. Relationships among range of motion, functional mobility, and quality of life in children and adolescents after limb-sparing surgery for lower-extremity sarcoma. *Pediatr Phys Ther*. 2006;18(4):238-244.
18. Verbecque E, Lobo Da Costa PH, Vereeck L, et al. Psychometric properties of functional balance tests in children: a literature review. *Dev Med Child Neurol*. 2015;57(6):521-529.
19. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283(15):2008-2012.
20. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21(4):651-657.
21. McPheeters ML, Kripalani S, Peterson NB, et al. Closing the quality gap: revisiting the state of the science (vol. 3: quality improvement interventions to address health disparities). *Evid Rep Technol Assess (Full Rep)*. 2012;208(3):1-475.
22. Portney L, Watkins M. *Foundations of Clinical Research Applications to Practice*. 3rd ed. NJ, Upper Saddle River: Pearson Prentice Hall; 2009.
23. van Tulder M, Furlan A, Bombardier C, et al. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. *Spine (Phila Pa 1976)*. 2003;28(12):1290-1299.
24. Saether R, Helbostad J, Riphagen I, et al. Clinical tools to assess balance in children and adults with cerebral palsy: a systematic review. *Dev Med Child Neurol*. 2013;55(11):988-999.
25. Butz SM, Sweeney JK, Roberts PL, et al. Relationships among age, gender, anthropometric characteristics, and dynamic balance in children 5 to 12 years old. *Pediatr Phys Ther*. 2015;27(2):126-133.
26. Itzkowitz A, Kaplan S, Doyle M, et al. Timed Up and Go: reference data for children who are school age. *Pediatr Phys Ther*. 2016;28(2):239-246.
27. Habib Z, Westcott S, Valvano J. Assessment of balance abilities in Pakistani children: a cultural perspective. *Pediatr Phys Ther*. 1999;11:73-82.
28. Bergmann JH, Alexiou C, Smith IC. Procedural differences directly affect timed up and go times. *J Am Geriatr Soc*. 2009;57(11):2168-2169.
29. Bardid F, Huyben F, Lenoir M, et al. Assessing fundamental motor skills in Belgian children aged 3-8 years highlights differences to US reference sample. *Acta Paediatr*. 2016;105(6):e281-e290.