Made available by Hasselt University Library in https://documentserver.uhasselt.be

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

Peer-reviewed author version

VERBECQUE, Evi; Schepens, Kirsten; There, Joke; Schepens, Benedicte; KLINGELS, Katrijn & Hallemans, Ann (2018) The Timed Up and Go Test in Children: Does Protocol Choice Matter? A Systematic Review. In: Pediatric physical therapy, 31 (1), p. 22-31.

DOI: 10.1097/PEP.0000000000000558
Handle: http://hdl.handle.net/1942/28004

The Timed Up and Go test in children: does protocol choice matter? A systematic review.

Verbecque Evi[1,2], Schepens Kirsten[1], Theré Joke[1], Schepens Bénédicte[3], Klingels Katrijn[4-5], Hallemans Ann[1,2]

1. Department of Rehabilitation Sciences and Physiotherapy, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
2. Multidisciplinary Motor Center Antwerp (M²ocean), Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
3. Laboratory of Physiology and Biomechanics of Locomotion, Institute of Neuroscience, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
4. BIOMED, Rehabilitation Research Center (REVAL), Hasselt University, Diepenbeek, Belgium.
5. Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium.

Abstract

Purpose: Results on reliability and normative data for the Timed Up and Go test (TUG) in typically developing children are systematically reviewed.

Summary of key points: Six different TUG protocols are presented for which normative data are available (age 3 to 18). In general, TUG time is quite consistent within/between raters and sessions (ICC≥0.61) and is influenced by age (R^2=[24.3;49.0]). But the choice of the protocol (self-selected versus fastest walking speed, use of a motivational aspect) and of the outcome calculation (averaging trials versus best performance) affect TUG time as well as its consistency within and between sessions.

Conclusions: A standard protocol for the TUG is lacking and needs to be developed, with attention for its 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 comprise fastest walking speed and a motivational aspect.

Keywords: typically developing children, TUG, reference values, reliability, "reproducibility of results"[mesh]

Introduction

Balance control serves as an overall prerequisite for acquiring (all) motor skills in children.[1-3] Therefore, the identification of potentially underlying balance deficits is fundamental for therapy planning. After children have learned to maintain the upright standing position, they acquire (fundamental) motor skills such as walking, running and jumping, which increases their functional independence. These motor skills require dynamic balance control, referring to the child's ability to maintain stable while moving from one base of support to the next.

The Timed Up and Go test (TUG) is a functional dynamic balance test for which interest has been growing strongly in the past fifteen years. The TUG is a timed measure during which the child has to stand up from a chair, walk 3 meters, turn around, walk back and sit down again. Originally, 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 transferred to pediatrics to screen for deviating 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 in children needs to be sensitive to age, related to the motor progression level of the child under investigation. For this purpose, normative data are imperative as they are used to determine cut-off values. A review conducted in 2013 on the TUG in children suggested normative values for the test still needed to be established.[7] Since then, several authors have reported normative data

for the TUG.[6,8] but with different protocols and age groups.

In general, motor competence is influenced by age, sex, weight, socio-economic status (SES) and ethnicity.[9,10] Balance control, similarly to motor development, increases with increasing age, so it can be 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.

Traditionally, in pediatric rehabilitation, to determine whether a child deviates from the norm, z-scores are used.[6,8,11] Such z-scores provide insights into how many standard deviations 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 when it comes to establishing normative data. Literature on reliability of the TUG shows that researchers have mainly focused on assessing these properties in children with atypical development, e.g. cerebral palsy[12-14], traumatic brain injury[15,16] and lower extremity sarcoma[17], providing evidence for high test-retest, intra-rater and interrater reliability in children with various motor impairments (ICC $\geq$0.85)[7,18]. In typically developing children, test-retest, intra- and interrater reliability varies between moderate and excellent (ICC$\geq$0.61).[6,12,19] Literature reviews regarding the TUG's reliability showed that information on the standard error of measurement (SEM) was scarce at the time.[7,18] Therefore, an update on reliability of the TUG in typically developing children could provide new insights into the applicability and usefulness of reported normative data.

In addition several authors have made adjustments to the protocol for testing in a pediatric population, such as using a chair with or without[13] arm- and backrest[12] and allowing to walk barefoot[14], walking with footwear[15,16] or with orthotics[15,16]. In contrast to the original protocol by Podsiadlo and Richardson (1991)[4], Williams et al. (2005)[12] suggested that self-selected walking

speed should be preferred over fastest walking speed when assessing TUG in children. Moreover, to make sure the children understand the test instructions, most authors propose an explanation followed by a demonstration with verbal feedback during the test if necessary.[6,8,12] To improve the children's motivation, different tools are described in 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 additionally or not may also influence the outcome. Finally in the original protocol the best of three trials was taken as the final result[4], but TUG outcome measures reported since then present for example an average of 2[15,16] or 3[12] trials, which might also affect the outcome.

Therefore, to screen for dynamic balance deficits, not only an overview on currently available normative data is necessary, but the protocol under investigation may also play a crucial role in the nature of these normative data and thus the cut-off values. Knowledge of the sample- and protocol-based influences on the TUG time may facilitate clinicians in selecting the most suitable type of TUG protocol for the individual child.

In summary, this study aims to provide an overview of the available normative TUG data for children through a systematic literature review. Due to potential influences of study sample characteristics as well as the possible effect of the applied protocol on TUG time, answers will be sought to the following research questions:

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

Methodology

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 can be consulted online (www.crd.york.ac.uk/prospero/).

Search method

Relevant literature for this systematic review was extracted from the Pubmed, Web of Science and Science Direct databases, covering Medline, Cochrane Database of Systematic Review, Cochrane Central Register of Controlled Trials, ISI Web of Knowledge and Web of Science. The final search was conducted on October 13th 2017 containing 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 of this search strategy were used to define the query in Web of Science and Science Direct. Mesh terminology was only used in Pubmed. No limits or filters were used. The search query was defined by four researchers (two students physical therapy with a bachelor's degree (KS, JT), a PhD student (EV) and the principal investigator (AH)).

Selection process

Relevant studies were identified using predefined selection criteria according to the Population Intervention Comparison Outcome Study Design (PICOS) method. Original studies (S), full and brief reports with transparent methodology, that reported normative data (O) for the TUG (I) in

typically developing children ≤18 years old (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, language. Two researchers (both students physical therapy with an academic bachelor's degree) assessed these criteria independently in two phases: on title and abstract (phase 1) and on the full text (phase 2). In case of doubt or disagreement, a third researcher's opinion was decisive. Afterwards references from the included articles were screened to assure no relevant information would be missed with the systematic search query.[19]

Risk of Bias in individual studies

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 checklist was constructed with the specific purpose of assessing risk of bias of studies investigating the psychometric properties of assessment tools.[20] The COSMIN checklist contains one box for each of the nine defined psychometric properties, e.g. reliability and measurement error. Each box comprises questions which 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, i.e. consistency of values) and box C (absolute reliability, measurement error) were relevant. For both boxes, two items were omitted as the TUG does not necessarily require independent measures to be reliable (item 5) and the study population comprises typically developing children indicating that motor abilities are stable within a short time interval (item 7). Each article was assessed independently by two investigators (EV and

KS/JT) and after a consensus meeting, a final score was assigned. Interrater reliability was determined using the Cohen's Kappa measure agreement between two raters (*k*).

Risk of bias in studies reporting ***normative data*** was not assessed as adequate tools are still missing. Nevertheless, to provide insights into how the sample was selected, the non-responders rate was acknowledged and typical development was ascertained, these characteristics were mapped, based on the "selection" category of the Newcastle-Ottawa Scale (NOS) adapted for cross-sectional studies.[21]

Data extraction

The following data were extracted where applicable:

- **Population-specific characteristics**: number of children, mean age (and standard deviations (SD)), age range and male-female ratio.
- **Specifics on the applied TUG protocol**: *instructions given to the subject* (self-selected walking speed versus 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 versus averaging trials and the number of trials included for analysis).
- **For reliability analyses**: TUG values (mean and SD), intra-class correlation coefficients (ICC) and the applied model, standard error of measurement (SEM), and minimal detectable change (MDC) were extracted. The ICC values were interpreted as follows: poor (ICC < 0.5), moderate (0.5 ≤ 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: $SEM = SD_{test\,1} \times \sqrt{(1 - ICC)}$. Subsequently, the MDC_{95} was calculated: $MDC_{95} = SEM \times 1.96 \times \sqrt{2}$.[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 cut-off 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+1SD* and *mean TUG+2SD* were presented graphically as a function of age.

All data were extracted by two independent researchers (KS/JT and EV) and compared in a consensus meeting.

Level of evidence

The level of evidence (strong, moderate, limited, unknown, 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. Implementation of the level of evidence was done according to the criteria of the Cochrane Collaboration Back Review Group[23], and Saether et al (2014)[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 revealed 293 hits in Pubmed, 230 hits in Web of Science and 204 hits in Science Direct, of which 616 were unique. After screening, five studies[6,8,12,25,26] met the criteria. One study[27] was added after reference screening, resulting in six studies that were used for data-extraction. The selection process is presented in Appendix 1.

Risk of bias in individual studies

Relative reliability was assessed in 5 of the included studies with methodological quality varying between poor and excellent (Table 1). Main reason for poor quality assignment was the small sample size (<30 children). Strong agreement was found between the two raters (*k*=0.769).

Selection procedures to establish normative data are presented 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 non-responders 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 in a total of 2626 typically developing children between age 3 and 18, of which 1212 boys. Overall, 46% of the children were boys, varying between 42[26]-54[12] % in the individual studies. 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]).

The Timed Up and Go test

Protocols

Six different TUG protocols were identified in literature. Half of the protocols consisted of the specific instruction for the children to walk as fast as possible[6,8,27] (Figure 1), whereas the others allowed self-selected walking speed[12,25,26] (Figure 2). 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 two studies, timing was started when the child got up from the chair[12,25] whereas in the other four 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 average of 2 trials[25-27], average of 3 trials[12] and the best of 3 trials[6,8].

Reliability of the protocols used for establishing normative data

Specifics on reliability results are presented in Table 3.

Intra-rater (within session) reliability

Intra-rater reliability was good across studies, with mean ICC-values varying between 0.80 and 0.998.[6,12,25,27] Williams et al. (2005) reported the SEM of 0.6 and 0.4 for respectively the baseline assessment and the TUG retest 10-20 minutes after the first test session in 3- to 9-year-old children.[12]

Interrater reliability

Three studies investigated interrater reliability[12,26,27] and reported very high ICC values (>0.9).[12,26] Habib et al. (1999) reported high percentages of agreement between raters (95-100%).[27] None of the studies reported SEM, nor were they 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 very consistent (ICC = [0.80;0.95]), regardless of the test moment (1-2 hours after the first test or 1 week afterwards).[6,12] But when children were divided into younger (age 3 to 5) and older children (age 5 to 9), test-retest reliability became more variable (ICC=[0.61;0.83].[12] Younger children tended to have more reliable results compared to older children when performing the test 10-20 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 MDC were not reported, but were calculated. The SEM varies between 0.33 and 0.75 seconds depending on the age group under consideration (Table 3).

Normative data: influence of study sample characteristics and protocols

In Figure 1 and 2, 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 Appendix 2. Four studies[8,25-27] reported numeric values for chronological age groups, whereas two studies[6,12] reported age bands combining several chronological ages.

In two studies significant differences in TUG time between boys and girls have been reported.[26,27] In these studies no motivational aspects were added to the protocol. Habib et al (1999) found an overall better performance for boys compared to girls regardless of age[27], whereas Itzkowitz et al. (2016) showed that only 8-, 9- and 11-year-old boys performed better than girls[26]. In all other studies where 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 American children, when allowing self-selected walking speed. In south Brazilian children, age and weight accounted for 25%[6] of the variance in TUG time (fastest performance), whereas in Belgian preschool children age and ethnicity explained 28%[8] of the variance in TUG time (fastest performance). Several authors reported that BMI[6,26] and body height[6,8,25] did not account for the variance in TUG time.

Differences in the normative data are observed depending on the applied protocol (Figure 1 and 2). Overall, significant differences between age groups have been 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 1), significant differences in TUG time have been found between three age bands, 5-7 year-olds, 8-10 year-olds and 11-13 year-olds[27], whereas when using self-selected walking speed instruction, the ages in the bands changed into 5-7 year-olds, 8-11 year-olds and 12-13 year-olds[26].

When motivation was used and TUG was performed at self-selected walking speed, preschoolers (age 3 to 5) performed the TUG significantly slower than older children (age 5 to 9).[12] When preschoolers performed the TUG with motivation as fast as possible, significant differences between these three chronological age groups were identified.[8] When SES was taken into account, Pakistani boys with low SES performed significantly better on the TUG than girls, but when compared to high SES girls, low SES girls perform poorer and high SES boys poorer than low SES boys.[27]

Level of evidence

The level of evidence and how it was obtained for reliability of the TUG are shown in Appendix 3A (relative reliability)/3B (absolute reliability). Strong evidence was found for relative and absolute intra-rater (within session) and test-retest reliability of the TUG protocol by Williams et al. (2005)[12] consisting of self-selected walking speed with a motivational aspect and averaging three trials[12]. Moderate evidence was found for relative and strong evidence for absolute intra-rater and test-retest reliability for the protocol by Nicolini-Panisson and Donadio (2014)[6], consisting of fastest walking speed with a motivational aspects and the best of three trials. Evidence for the other protocols' reliability remains unknown for now.

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 is recognized to influence TUG performance, but other predictors such as the applied protocol have been identified as well. In the following paragraphs, these findings will be discussed in more detail.

Reliability of the TUG Protocols

Reliability analyses on TUG protocols (used for reporting normative data), remain incomplete, especially when protocol differences are taken into account. Mainly intra-rater (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 six papers 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 methodological shortcomings, reliability results need to be interpreted with caution. Nevertheless, the reported ICC values were high (ICC≥0.8) for all types of reliability, indicating strong agreement exists between the administered trials, raters and/or sessions.[6,12,25-27]

Interestingly, younger children (age 3 to 5) tend to have more consistent results over a shorter time interval and less consistency over a longer time interval compared to older children (age 5 to 9).[12] All children were considered to be stable during a short time interval (maximum 2 weeks), as no prominent changes in their motor progression are to be expected. However, ICC values seem to be affected by age and the time interval between the test sessions. A presumable explanation is that gait in children under age 7 is still developing towards a mature gait pattern.[13] Because of large intra-variability 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. Also, 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. (2005)[12] did not provide any 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 still needs to be confirmed in future research.

When a shorter time interval was introduced between sessions (e.g. 10-20 minutes)[12], less variance in ICC-, SEM- and MDC values were observed, suggesting practice/learning effects occur, e.g. 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] The same may hold up for assessment between sessions with short time intervals, especially in younger children.

Normative data

To screen balance deficits in children, normative data and corresponding cut-off values need to be available and were therefore mapped. Because of the ongoing development and maturation of balance control during walking, it is expected that increasing age results in better TUG performance in typically developing children, and thus a descending trend of TUG time as a function of age. Several authors indeed suggested that variance in TUG time is mainly 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 two studies reference values were reported by chronological age[8,25], whereas most authors grouped several chronological ages into one age band, e.g. age 5 to 9 years[12], or 6 to 9 years[6] or 5 to 7 years[26,27]. Only Habib et al (1999)[27] and Itzkowitz et al (2016)[26] provided evidence, i.e. the presence of significant differences, for grouping specific chronological age groups into age bands. They found one 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 different age bands might be the investigated samples. First, though Habib et al. (1999)[27]

had a smaller sample size in each subgroup (approximately 20), they were equally distributed over the chronological age bands, which was not the case in the study by Itzkowitz et al. (2016)[26] (sample size varies between 45 and 244 per subgroup). Not only sample size, but also the composition of the sample may have influenced TUG results (Table 2). Children were recruited from different countries all over the world, which emphasizes the potential influence of intercultural differences, as previously mentioned by the investigating authors.[6,12] For example, poorer performance of Pakistani girls compared to boys with a low SES, was assigned to cultural influences as these girls often wear a chador, limiting their mobility.[27] Interestingly, Itkowitz et al. (2016)[26] was the only other author to find sex-related differences and both these studies[26,27] lack the "motivational aspect" during TUG administration. Indeed sex-related differences might be a result of lacking motivation. Bardid et al. (2016) 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.[29] Therefore boys might be more stimulated in performing gross motor skills through sports as well as their competition thrive. By adding a motivational aspect to the TUG protocol, girls may be stimulated more, hence the lack of differences between sexes observed.[6,8,25] This again suggests that the protocol also influences performance. Indeed, based on the presentation of the normative data as a function of protocol (Figure 1 and 2), the applied protocol interferes with how normative data present. The expected trend of decreasing TUG time with increasing age is seen when using 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, strong 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, such fluctuations should be limited.

The applied TUG outcome also seems to influence the normative data. When protocols consist of using the best of three performances or an average of three trials, a decreasing trend of TUG time with increasing age is observed.[6,8,12] The best of three trials provides information on the best performance, whereas averaging trials has the advantage of taking the intra-individual variability of performances into account. In preschool children, walking as fast as possible, three trials within one session differed significantly, highlighting the need for using best performance, but also the need to determine whether three TUG trials within one session are enough to overcome practice effects.[8] None of the five studies that investigated reliability reported within session differences. According to Podsiadlo and Richardson (1991) the best of three trials should be used as the final result.[4] However, it remains to be determined how many trials are actually necessary in 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 towards both age effects and number of trials required allowing adequate assessment.

Thus, both protocol differences and methodological characteristics used to select the sample (Table 2) may have induced fluctuations in TUG time. Although differences in both were mapped, they were not assessed on their potential risk of bias, resulting in a limited body of evidence towards the most suited protocol for the TUG and the corresponding normative data to use in clinical practice.

Based on current knowledge[4] and our experience (unpublished observations), it seems that fastest walking speed, the use of an additional motivational aspect, best performance and at least three trials should be preferred in pediatric rehabilitation. These protocol characteristics will stimulate

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. However, further research into the most suited TUG protocol still needs to be performed.

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. So, 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 lacking 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 at least three trials, still needs to be tested on a larger sampling, as now it is based on reliability results of one study and the fact that such a protocol seems to reduce fluctuations in results between different age (groups).

Several authors have suggested that intercultural differences may affect TUG time as well.[6,12,27,] For now, the impact of intercultural 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 always reported. Next to sample descriptions, sample recruitment plays a role as well. Most studies used convenience samples, thereby increasing the risk of selection bias, which highlights the need for random sampling in future research.

Five out of six relevant studies were retrieved using three main databases such as Pubmed, Science Direct and Web of Science, but a 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, this did not influence the results.

Conclusion 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. Formerly, authors have been changing the TUG protocol without investigating its impact on both reliability and normative data. However, this review suggests that the protocol may affect TUG time and variance in reliability measures. Also age seems to play a role as a result of ongoing psychomotor development. Especially children under age six should be addressed separately. Thus, 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 three trials within one session. However, the results in the present review are to be interpreted cautiously, as they are based on only six studies that all investigated different protocols, included different sample sizes and sample compositions which limits their generalizability. Moreover, when establishing normative data, attention needs to be payed to objectifying typical motor development through validated developmental motor scales.

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 Dec;4(6):877-87.
3. Shumway-Cook A, Woollacott MH. Motor Control, Translating Research into Clinical Practice. 4th edition. Philadelphia, PA: Lippincott Williams and Wilkins, 2014: 161–93.
4. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc.* 1991 Feb;39(2):142-8.
5. Park SH. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. Aging Clin Exp Res. 2018 Jan;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 May;56(5):490-7.
7. Nicolini-Panisson RD, Donadio MV. Timed "Up & Go" test in children and adolescents. *Rev Paul Pediatr*. 2013 Sep;31(3):377-83.
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 winter;28(4):409-15.
9. Barnett LM, Lai SK, Veldman SLC, Hardy LL, Cliff DP, Morgan PJ, Zask A, Lubans DR, Shultz SP, Ridgers ND, Rush E, Brown HL, Okely AD. Correlates of Gross Motor Competence in Children and Adolescents: A Systematic Review and Meta-Analysis. Sports Med. 2016 Nov;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 Dec;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 Spring;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 Aug;47(8):518-24.
13. Gan SM, Tung LC, Tang YH, et al. Psychometric properties of functional balance assessment in children with cerebral palsy. *Neurorehabil Neural Repair*. 2008 Nov-Dec;22(6):745-53.
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 Summer;16(2):90-8.
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 Fall;20(3):254-8.
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 Feb;22(2):153-9.
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 Winter;18(4):238-44.

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 Jun;57(6):521-9.
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 Apr 19;283(15):2008-12.
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 May;21(4):651-7.
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 Aug;(208.3):1-475.
22. Portney L, Watkins M. Foundations of Clinical Research Applications to Practice, 3rd ed. Upper Saddle River, New Jersey: 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 Jun 15;28(12):1290-9.
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 Nov;55(11):988-99.
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 Summer;27(2):126-33.

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 Summer;28(2):239-46.
27. Habib Z, Westcott S, Valvano J. Assessment of Balance Abilities in Pakistani Chidren: 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 Nov;57(11):2168-9.
29. Bardid F, Huyben F, Lenoir M, Seghers J, De Martelaer K, Goodway JD, Deconinck FJ. Assessing fundamental motor skills in Belgian children aged 3-8 years highlights differences to US reference sample. *Acta Paediatr*. 2016 Jun;105(6):e281-90.

Figure legend

Figure 1: Overview of the applied protocols requiring fastest walking performance with corresponding normative data as a function of chronological age (groups).

Figure 2: Overview of the applied protocols requiring self-selected walking performance with corresponding normative data as a function of chronological age (groups).

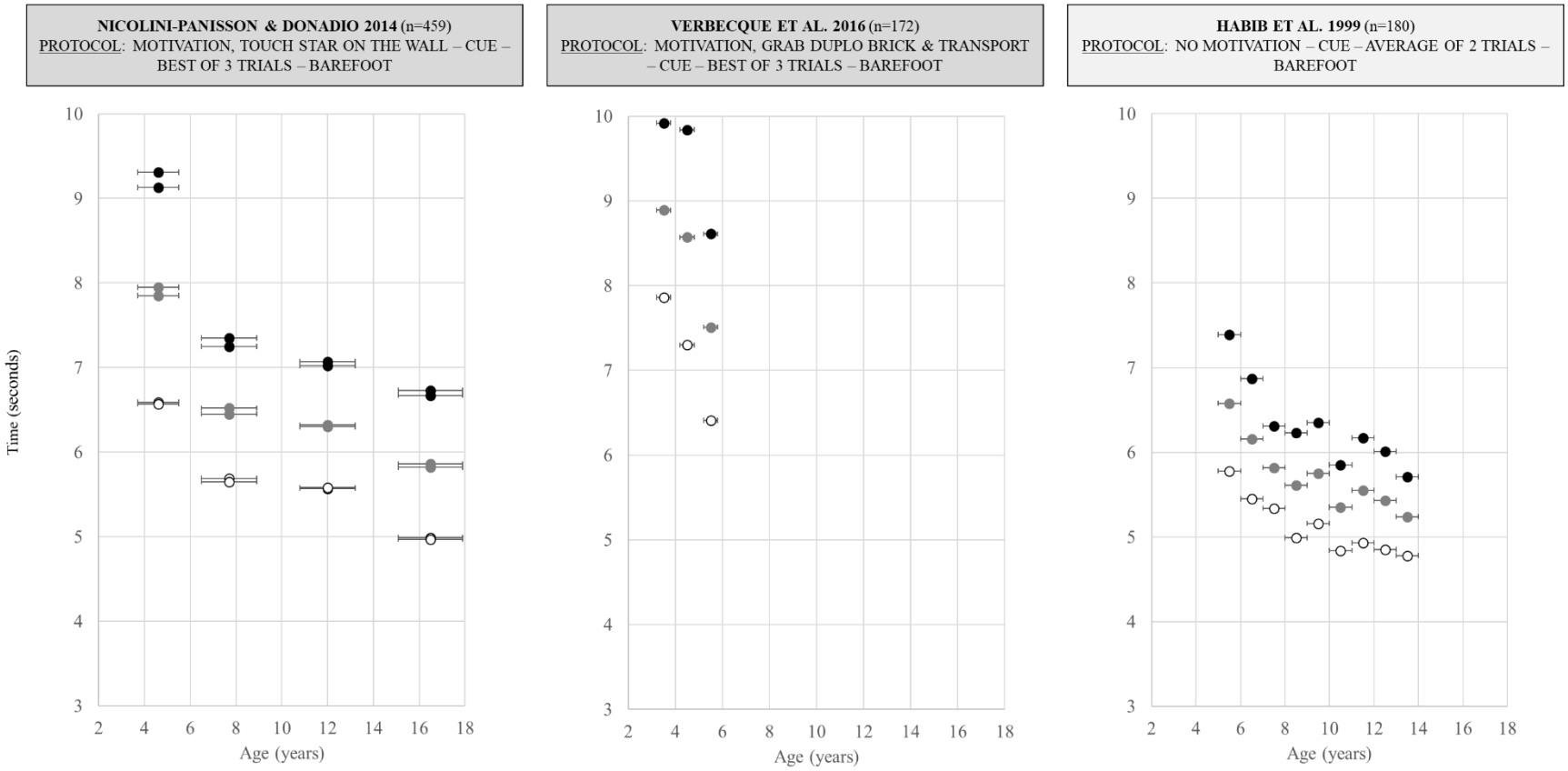
Table legend

Table 1: Risk of bias in individual studies regarding reliability using the Consensus-based Standards for the selection of health Measurement Instruments.

Table 2: Methodological characteristics of sample selection in studies reporting normative data.

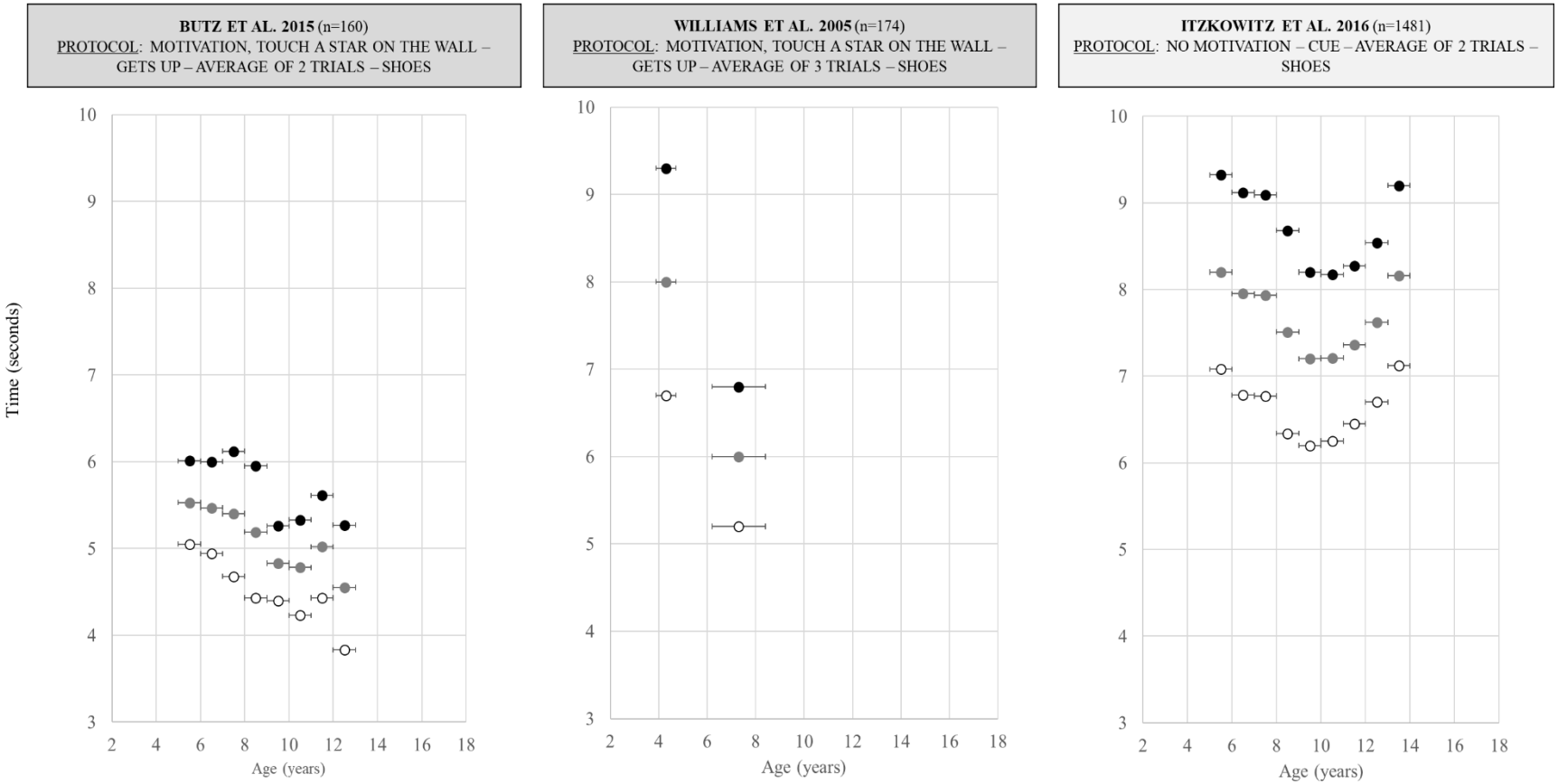
Table 3: Intra-rater (within session), interrater and test-retest reliability of the Timed Up and Go test protocols used for normative data.

Figure 1: Overview of the applied protocols requiring <u>fastest walking performance</u> with corresponding normative data as a function of chronological age (groups).



<u>Legend</u>: ● Mean TUG values + 2 standard deviations (SD); ● Mean TUG values + 1SD; ○ Mean TUG values; horizontal error bars represent 1SD from the mean age. For Nicolini-Panisson and Donadio 2014, TUG 1 and TUG 2 are presented (see Results Table 3).

Figure 2: Overview of the applied protocols requiring <u>self-selected walking performance</u> with corresponding normative data as a function of chronological age (groups).



<u>Legend</u>: ● Mean TUG values + 2 standard deviations (SD); ● Mean TUG values + 1SD; ○ Mean TUG values; horizontal error bars represent 1SD from the mean age.

Table 1: Risk of bias in individual studies regarding reliability using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN).

Author	COSMIN	Type of reliability	Rater A	Rater B	Consensus score	Reason for a consensus rating less than excellent
Butz et al. 2015	Box B	Intra-rater	Poor	Poor	Poor	Sample size < 30
		Inter-rater	Poor	Poor	Poor	Sample size < 30
Habib et al. 1999	Box B	Intra-rater	Poor	Poor	Poor	Sample size < 30
		Inter-rater	Poor	Poor	Poor	Sample size < 30, statistical method (percentage of agreement)
Itzkowitz et al. 2016	Box B	Inter-rater	Poor	Poor	Poor	Sample size < 30
Nicolini-Panisson & Donadio 2014	Box B	Intra-rater	Excellent	Excellent	Excellent	
		Test-retest	Excellent	Excellent	Excellent	
	Box C	Test-retest	Excellent	Poor	Excellent	
Williams et al. 2005	Box B	Intra-rater	Good	Good	Good	The applied ICC model was not reported
		Test-retest	Good	Good	Good	The applied ICC model was not reported
	Box C	Intra-rater	Excellent	Excellent	Excellent	

Legend: Box B concerns the relative reliability, box C the measurement error.

Table 2: Methodological characteristics of sample selection to report normative data.

Authors	Sample selection	Sample size calculation	Non-responders rate	Ascertainment of typical development	Sample size and age bands
Butz et al. 2015	Sample of convenience (5-12 years old); elementary schools Connecticut, private schools in West Haven, outpatient rehabilitation Connecticut.	Effect size between 0.3-0.4, 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 surgeries within the past 6 months.	160 children, 8 age bands according to chronological age.
Habib et al. 1999	Partial random and partial convenience sampling (5-13 years old); 2 private schools (random), 4 orphanages and 1 school from Malir (convenience) in Pakistan.			Physical exam: 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. 2016	Sample of convenience (5-17 years old); 20 public elementary and middle schools from 5 New York City boroughs		18231 invitation letters; 1653 responders of which 1481 completed the TUG.	Medical history: 1) no orthopedic surgeries or injuries within the past 6 months; 2) no history of neurological disorders; 3) no individualized educational program.	1481 children, 9 age bands according to chronological age.
Nicolini-Panisson & Donadio 2014	Sample of convenience (3-18 years 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 months previously, 2) cardiorespiratory and neuromuscular diseases, or intellectual disability, 3) incorrect performance of the test.	459 children, 4 age groups: age 3-5, age 6-9, age 10-13 and age 14-18.
Verbecque et al. 2016	Sample of convenience (3-5 years 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. 2005	Sample of convenience (3-9 years old); nearby schools, kindergartens, child-care centres in Melbourne.				176 children; 2 age groups: age 3-5 and age 5-9.

Table 3: Intra-rater (within session) and test-retest reliability of the Timed Up and Go test (TUG) protocols used for normative data.

	Author	n	Age range (years)	TUG (seconds)									
				Test session	Trial 1		Trial 2		Trial 3		ICC		SEM	MDC_{95}
					Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*95% CI*		
Intra-rater (within session) reliability	Butz et al. 2015	10	5-12	*Test 1*							*0.998*[E]			
	Habib et al. 1999	180	5-13	*Test 1*							*0.81*[D]			
	Nicolini-Panisson & Donadio 2014	459	3-18	*Test 1*							*0.93*[F]			
		459	3-18	*Retest same day*							*0.94*[F]			
		178	3-18	*Retest 1 week after test 1*							*0.95*[F]			
	Verbecque et al. 2016	172	3-5	*Test 1*	*8.24*[A]	*1.97*	*7.92*[A]	*1.72*	*7.58*[A]	*1.60*				
	Williams et al. 2005	176	3-9	*Test 1 (baseline)*	*6.0*	*1.5*	*5.9*	*1.3*	*5.9*	*1.3*	*0.8*[A]	*0.75-0.84*	*0.6*	*1.86**
		173	3-9	*10-20 minutes after test 1*	*5.9*	*1.5*	*5.9*	*1.5*	*5.9*	*1.5*	*0.89*[A]	*0.86-0.92*	*0.4*	*1.38**
		151	3-9	*1 week after test 1*	*5.7*	*1.2*	*5.8*	*1.2*	*5.7*	*1.2*	*0.85*[A]	*0.81-0.89*	*0.46**	*1.29**

	Author	n	Age range (years)	TUG (seconds)									
				Test session			Rater 1		Rater 2		ICC		SEM	MDC_{95}
							Mean	*SD*	*Mean*	*SD*	*Mean*	*95% CI*		
Inter-rater reliability	Butz et al. 2015	10		*Test 1*							*0.999*[C]			
	Itzkowitz et al. 2016	22		*Test 1*							*0.988*[D]			

	Author	n	Age range (years)	TUG (seconds)									
				Measurement characteristics			Session 1		Session 2		ICC		SEM	MDC_{95}
				Live/video	*Duration between tests*	*Trial used for analysis*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*95% CI*		
Test-retest reliability	Nicolini-Panisson & Donadio 2014	178	3-18	*Live*	*1 week*	*Best of 3*					*0.95*[F]			
		459	3-18	*Live*	*1-2 hours*	*Best of 3*					*0.95*[F]			
	Williams et al. 2005	173	3-9	*Live*	*1 week*	*Average of 3*	*5.90*	*1.3*	*5.70*	*1.1*	*0.83*[B]	*0.77-0.88*	*0.54**	*1.49**
		83	3-5	*Live*	*1 week*	*Average of 3*	*6.7*	*1.2*	*6.50*	*1.0*	*0.61*[B]	*0.39-0.75*	*0.75**	*2.08**
		90	5-9	*Live*	*1 week*	*Average of 3*	*5.2*	*0.8*	*5.0*	*0.8*	*0.83*[B]	*0.73-0.89*	*0.33**	*0.91**
		173	3-9	*Live*	*10-20 minutes*	*Average of 3*	*5.9*	*1.3*	*5.9*	*1.5*	*0.89*[B]	*0.86-0.92*	*0.43**	*1.20**
		83	3-5	*Live*	*10-20 minutes*	*Average of 3*	*6.7*	*1.2*	*7.0*	*1.3*	*0.82*[B]	*0.72-0.88*	*0.51**	*1.41**
		90	5-9	*Live*	*10-20 minutes*	*Average of 3*	*5.2*	*0.8*	*4.9*	*0.8*	*0.76*[B]	*0.61-0.85*	*0.39**	*1.09**

Legend: **n**: number of children; **SD**: standard deviation; **ICC**: Intra-class correlation coefficient; **SEM** (Standard Error of measurement) = $SD_{test\ 1}$* (1-ICC)$^{1/2}$; **MDC_{95}** (Minimal Detectable Change) = SEM*1.96*$2^{1/2}$; **[A]** values differ significantly; ***** values have been calculated.

Applied ICC models: **A**= one way random effects, absolute agreement, single rater, ICC (1,1); **B**= one way random effects, absolute agreement, multiple raters/measurements, ICC (1,3); **C**= two-way random effects, absolute agreement, single rater/measurement, ICC (2,1); **D**= two-way mixed effects, consistency, single rater/measurement, ICC(3,1); **E**=two-way mixed effects, consistency, multiple raters/measurements, ICC (3,2); **F**= not reported.