Title: Reliability and concurrent validity of a modified Timed Up and Go test for healthy preschoolers

Authors

- Ann Hallemans[1,2], ann.hallemans@uantwerpen.be
- Katrijn Klingels[3,4], katrijn.klingels@uhasselt.be
- Tamaya Van Criekinge[1,2], tamaya.vancriekinge@uantwerpen.be
- Luc Vereeck[1,2], luc.vereeck@uantwerpen.be
- Evi Verbecque[1,2,3], evi.verbecque@uhasselt.be

Author contribution:

- Conception of the study: E. Verbecque
- Data collection: T. Van Criekinge, E. Verbecque
- Analyses and interpretation of data: A. Hallemans, L. Vereeck, T. Van Criekinge, K. Klingels, E. Verbecque
- Drafting the manuscript: A. Hallemans, L. Vereeck, T. Van Criekinge, K. Klingels, E. Verbecque
- Each of the authors has read the final version of the manuscript and concurs with the content in the manuscript.

Author affiliations

1. Department of Rehabilitation Sciences and Physiotherapy, Faculty of Medicine and Health Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium.
2. Multidisciplinary Motor Centre Antwerp (M²OCEAN), University of Antwerp, Universitair Ziekenhuis Antwerpen route 41, Wilrijkstraat 10 2650 Edegem.
3. Rehabilitation Research Center (REVAL) Pediatric rehabilitation, Faculty of rehabilitation sciences and physiotherapy, Hasselt University, Agoralaan building A, 3590 Diepenbeek, Belgium.
4. Research Group for Neurorehabilitation (eNRGy), Department of Rehabilitation Sciences, KU Leuven, Tervuursevest 101 - box 1501, 3001 Leuven.

Corresponding author: evi.verbecque@uhasselt.be.

Keywords: “postural balance”[Mesh], “child, preschool”[Mesh], “reproducibility of results”[Mesh], reliability, “concurrent validity”

What is known:

- A large variety of Timed Up and Go test protocols is available for children
- The protocols differ in the instructions on walking speed (self-selected/fastest), the use of an extra motivation (e.g. touch a star on the wall) and the applied outcome (average/best performance)

What is new:

- The best Timed Up ang Go test performance induces more consistent test results between raters and sessions and also less standard- and measurement error in 3- to 5-year-old children
- The best Timed Up and Go test performance should be preferred over the average performance to achieve both reliable and valid test results in 3- to 5-year-old children

Abstract

The purpose of this study was to assess the interrater and test-retest reliability and the concurrent validity of the modified Timed Up and Go test for preschoolers. As such, we aim to determine the most suited outcome of the modified Timed Up and Go test: the best or the average performance. Thirty-two children (age 3-5) performed three Timed Up and Go test trials as fast as possible on two separate occasions. During the first session, two researchers recorded the time to perform the task simultaneously. For reliability analyses, intra-class correlation coefficients (ICC's) and the minimal detectable change were determined. A Pearson correlation coefficient was calculated to determine concurrent validity between the Timed Up and Go test and the balance subscale of the Movement Assessment Battery for Children, 2nd edition. Interrater (ICC>0.97) and test-retest (ICC>0.75) reliability were good both for the average and the best Timed Up and Go test performance. A minimal detectable change of 1.86 seconds was found for the best performance, and 2.30 seconds for the average performance. Only the best Timed Up and Go test performance correlated significantly with the balance subscale of the Movement Assessment Battery for Children, 2nd edition, though fair (r=-0.347, p=0.007).

Conclusion: The modified Timed Up and Go test for preschoolers using the best performance is reliable and recommended to reduce standard- and measurement error.

List of Abbreviations in alphabetical order

- ICC: Intra-class correlation coefficient
- MABC-2: Movement Assessment Battery for Children, 2nd edition
- MDC: Minimal Detectable Change
- SEM: Standard Error of Measurement
- TUG: Timed Up and Go test

Introduction

The Timed up and go test (TUG) is widely used in clinical practice to assess a child's functional balance control during walking [1,2]. Therefore, the TUG is an often-applied outcome measure to assess the effectiveness of physiotherapeutic interventions in various pediatric populations, e.g. sensorineural hearing loss, acquired brain injury and cerebral palsy [3-5].

The original protocol by Podsiadlo & Richardson (1991) [6] was developed to assess functional mobility in elderly people and consists of getting up from a chair, walk three meters, turn around, walk back and sit down again. The subject is expected to walk at a comfortable and safe pace and is allowed three trials, of which the best is used as a final result. Although the TUG's original test protocol is very straightforward [6], several modifications have been made in the past 15 years to enhance its administration in children [7]. Protocol alterations mainly comprise: 1) the use of self-selected walking speed [8-10] instead of the fastest performance [11-13], 2) allowing two trials [8,9,11] instead of three [10,12,13], 3) averaging trials [8-11] versus the best performance [12,13] and 4) providing an extra motivation such as star on the wall that needs to be touched [8,10,12] or a Duplo brick [13] the children need to grasp and transport.

All these protocol changes can affect the child's final test result. The diverging distribution of the reported normative data that are currently available in literature reflects the influence of the applied protocol on the TUG performance [7].

Protocol changes can also influence the reliability of the TUG results. In children with severe motor impairments due to central nervous system disorders, the TUG is highly reliable within a session (ICC>0.85) [14-16], between raters (ICC>0.84) [17] and between sessions (ICC>0.91) [15,17-19] regardless of the applied protocol. However, in typically developing children, when the protocol does not add an extra motivation [11], allows self-selected walking speed [8,10] or averages TUG trials [8,10,14,20], intraclass correlation coefficients (ICC) have been shown to be lower or more variable (ICC=[0.61;0.99]). These specific protocol features tend to influence the reproducibility of a child's performance over a one-week interval, especially in young children (age 3-5, ICC=0.61 [10]). This may

have repercussions for the TUG's usefulness as a screening tool for balance deficits in young children with mild motor impairments.

In 2016, we suggested a modified protocol for children who are preschoolers (age 3-5), emphasizing fast but safe walking speed and in which we added a motivational aspect to the original protocol by Podsiadlo & Richardson (1991) [6], a Duplo brick the children need to grasp and transport [13]. This TUG protocol has been shown to discriminate between performances of 3-, 4- and 5-year-olds [13]. However, the reliability of this protocol has not yet been investigated, thereby limiting its clinical applicability.

In clinical practice, knowledge of which TUG outcome is most useful is imperative for therapeutic purposes (identification of dysfunction and assessment of a treatment effect). In case of an outcome consisting of repeated measures, as is the case for the TUG, it always remains the question whether trials should be averaged or the best performance should be used.

The main aim of this study is therefore to establish whether within the modified TUG protocol, the average or the best of three trials is most reliable and valid in 3- to 5-year-old healthy children. For this purpose, the within-session, interrater- and test-retest reliability of our modified TUG protocol for preschoolers and the concurrent validity with the Movement Assessment Battery for Children, 2nd edition (MABC-2) will be investigated.

Materials and methods

Participants

One hundred brochures with study information were distributed to one regular regional preschool in the vicinity of Antwerp, Belgium. Children were included for testing after their parents had given written informed consent. The study protocol was approved by the ethical committee of the Antwerp University Hospital (B300201629848). Data were collected in March 2018.

Parents were asked to fill in a questionnaire to inquire into features of the pregnancy and the child's birth, presence of any diagnosed developmental disorders, severe visual or hearing impairments and the use of aids such as glasses, orthoses and cochlear implants. These characteristics were used to describe the sample. Children were excluded from the study if they had a diagnosis of a neurodevelopmental or

a neuromotor disorder, severe visual impairments or when they were uncooperative, i.e. unwilling to perform the requested tasks.

Measurement

All children performed the modified TUG [13] and the balance subscale of the Movement Assessment Battery for Children, 2nd edition (MABC-2) [21]. The tests were administered in the school's hallway during class activities in the morning so they would not be disturbed or distracted by other children. This way, the test administrations were done at approximately the same time of the day during test and retest sessions. The children wore every-day clothes and shoes during the assessments. Time measurements were recorded with a digital stopwatch. Children were permitted to rest between trials and/or tests.

The modified Timed Up and Go test

The modified TUG for preschoolers requires the child to get up from a sitting position on the cue 'go', walk as fast as possible but safely towards a Duplo brick (s)he needs to grasp and transport back and sit down again. Details of the protocol have been described elsewhere [13]. Test instructions were explained beforehand and if necessary the test was also demonstrated. A red Duplo brick®, the children needed to grasp and transport, was used to improve motivation. Each child performed one practice trial at self-selected walking speed to verify they had understood the test instruction, and three trials as fast as possible [7]. Each child was allowed to rest between trials, for approximately ten seconds, i.e. the time it takes for the raters to write down the test result, get in position again and ask the child whether (s)he is ready to continue.

Movement Assessment Battery for children – second edition (MABC-2)

The balance subtest of the MABC-2, a reliable and valid test for assessing motor performance in children this age [21], was administered in all children to screen for balance deficits. This subtest comprises three items: standing on one leg (preferred and non-preferred) for maximum 30 seconds, walking on tip-toes on a 4.5 m long line and bipedal hopping (maximum of 5 hops). Raw scores were converted to standard scores and percentiles. The standard scores were used to determine concurrent validity with the modified TUG. Performances at or below the 16th percentile were considered indicative for a balance deficit and were used to describe the sample. The MABC-2 subscale was administered by rater 2 (EV).

Data collection and variables of interest

The data were collected on two days. During the first day, data of the TUG (TUG-1) and the balance subtest of the MABC-2 were collected (n=32). Eighteen children were randomly selected for a retest session (TUG-2) in a blinded manner, using opaque envelopes. The time between test and retest was either 1 (n=8) or 2 weeks (n=10). This time difference for retest was a result of logistical reasons, which did not induce significant differences between test and retest sessions after one week nor after two weeks ($p>0.05$). Environmental factors were similar in test and retest except for the presence of 2 assessors at the first test session and only 1 assessor at the retest session.

For *interrater reliability*, TUG-1 performance was simultaneously recorded by two raters (TVC, EV): one rater gave the instructions to the children and the cue to begin timing, whereas the other rater only timed the child's performance on the cue (TUG-1). Both raters were physical therapists, each having at least 5 years of experience. Interrater reliability analyses were performed on the best of 3 trials (TUG-1-B3) and the average of the three trials (TUG-1-A3) measured during the first session.

Within-session reliability of the TUG was established using the three trials performed by each child during the first test session (TUG-1) measured by both raters (TVC and EV).

The TUG-2 was administered by either rater 1 or rater 2. To establish *test-retest reliability*, TUG-1 and TUG-2 results from the same rater were used for analyses. The best of three trials on both sessions (TUG-1&2-B3) and the average of the three trials on both sessions (TUG-1&2-A3) were analyzed.

To address concurrent validity, the TUG-1 test results (rater 1) and those of the MABC-2 balance subscale (rater 2) were compared. Both raters were blind to the other results.

Statistical Analysis

Normal distributions of the TUG times were confirmed with the Kolmogorov-Smirnov test, allowing the use of parametric statistics. Differences between trials, raters and sessions were established with paired samples t-tests. Significance was set at $p<0.05$.

Reliability was investigated within a session (within-session reliability), between raters (interrater reliability) and between sessions (test-retest reliability). Within-session and test-retest reliability was

determined by calculating ICC's, for which a two-way mixed effects model with absolute agreement [22] for single measures was applied. The ICC's for interrater reliability were calculated with a two-way random effects model with consistency [22]. The ICC values were interpreted as poor (ICC < 0.5), moderate ($0.5 \leq ICC < 0.75$), or good ($ICC \geq 0.75$) [23]. For interrater and test-retest reliability, the ICC values were calculated for the best of three trials as well as the average of three trials. Subsequently, the standard error of measurement (SEM, formula: $SD_{test\ 1}*(1-ICC)^{1/2}$), the minimal detectable change (MDC_{95}, formula: $SEM*1.96*2^{1/2}$) and measurement error, expressed as the 95% confidence interval of the mean difference in TUG time recorded by the two raters (interrater reliability) or during the two test sessions (test-retest reliability), were computed [23].

Concurrent validity was established by correlating the standard scores of the MABC-2 balance subscale with z-scored TUG-1 results (rater 1) using previously reported normative data [13]. The z-scored values were chosen to reduce the influence of age on the TUG data, similar to the standard scores applied in the MABC-2. The relationship between TUG and the balance subscale of the MABC-2 was investigated with a Pearson correlation coefficient ($r \leq 0.25$: little or no relationship, $0.25 < r \leq 0.50$: fair relationship, $0.50 < r \leq 0.75$: moderate to good relationship, $r > 0.75$: good to excellent relationship [23]).

The data were statistically analyzed with IBM SPSS for Windows (25.0).

Results

Participants

Thirty-three children between age 3 and 5 enrolled the study, 32 of which the data were complete (mean age (SD) in months: 55.0 (9.9)). One three-year-old refused further testing after performing one of the assessments. The sample consisted of 9 three-year-olds, 11 four-year-olds and 12 five-year-olds. None of the children wore glasses or orthoses. A description of the sample is presented in *Table 1*.

Reliability

Within-session reliability. The TUG time was similar for the first and second trial, but significantly decreased between the first and the third trial ($p<0.01$) and between the second trial and the third trial ($p<0.01$) for both raters. The distribution of the TUG-1 trials is presented in *Figure 1A*. The three trials were moderately consistent (ICC=0.69) for both raters (*Table 2*).

Interrater reliability. The TUG results differed significantly between raters for both the TUG-1-B3 and the TUG-1-A3 ($p<0.001$). The distribution of TUG-1-B3 and TUG-1-A3 for each rater is shown in *Figure 1B*. The consistency between raters for both TUG-1-B3 and TUG-1-A3 was strong (ICC=0.99) (*Table 3*). The measurement error for the TUG-1-B3 was 0.20 seconds and for TUG-1-A3 was 0.26 seconds. The Bland-Altman plots depicting the measurement error are graphically presented in *Figure 2A*.

Test-retest reliability. The TUG performance was similar between test and retest sessions for both best performance (TUG-1&2-B3, p=0.074) and the average performance (TUG-1&2-A3, p=0.610), as shown in *Figure 1C*. Both best of three and averaged TUG results were strongly reproducible between sessions (TUG-B3: ICC=0.78, TUG-A3: ICC=0.75; *Table 3*). The mean difference between test and retest was 0.39 seconds for TUG-1&2-B3 and 0.13 seconds for TUG-1&2-A3 (*Table 3*). The SEM between sessions was 0.67 for the TUG-1&2-B3 and 0.83 for the TUG-1&2-A3 and the MDC was respectively 1.86 seconds and 2.30 seconds. The measurement error and its 95% confidence interval for test-retest reliability is shown in *Figure 2B.*

Concurrent validity

The relationship between the TUG-1-B3 and the balance subscale of the MABC-2 was fair and significant (r=-0.347, p=0.007), which was not the case for the TUG-1-A3 (r=-0.251, p=0.053).

Discussion

With this study, reliability and concurrent validity of a modified TUG protocol in typically developing preschoolers (age 3-5), have been established. The overall aim was to determine the most reliable outcome for the modified TUG: the best of three trials or the average of three trials. Although both are consistent between raters and sessions, as shown by the similar ICC values, the results of the present study endorse the use of the best of three trials for several reasons. Compared to the average of three trials, the best of three trials gives 1) less variable results between raters and sessions (*Figure 1B and 1C*), 2) smaller standard and measurement errors (*Table 3 and Figure 2*) and 3) a fair, but significant relationship with the balance subscale of the MABC-2.

Determination of the most reliable TUG outcome (best versus average) is much needed in very young children because of the large intra-individual variability in their performances, which is a typical feature of this developmental stage [24]. By using the best performance, this variability is not considered, whereas this is the case when performances are averaged. Although the best of three trials is recommended in the original TUG protocol [6], no previous studies investigated the effect of the chosen outcome on the protocol's reliability in children [1,2,7]. Especially when it comes to applying the TUG in clinical practice, it is necessary to know the degree to which measures can be used interchangeably between raters, but also which outcome is most reliable over time to determine the treatment effect.

Based on current results we can conclude that the modified TUG can be used reliably between raters. Both the best and average performance result in strong ICC values and small standard and measurement errors. The measurement error of the TUG time between the two raters is approximately 0.2-0.3 seconds (*Table 3*) for best and average performance. This can be explained by the applied methodology. One rater gave the instructions, whereupon the other rater commenced timing in response to the verbal cue 'go'. This induced a small systematic difference between raters, which is expected and acceptable. Nevertheless, because of this methodology, interrater reliability in the present study refers to reproducibility of TUG time in one session measured by two raters simultaneously. Interrater reliability emphasizing the reproducibility of the instructions given by different raters has not been addressed yet. As instructions are key in motivating children, this should be established in future research.

The reproducibility over time (test-retest reliability) of the modified TUG is good in preschool children, with results mainly in favor of the best TUG performance. The difference in ICC values is negligible, but more importantly, the variability across subjects as acknowledged with the SEM and MDC is smaller for the best performance compared to average performance. The MDC values were 1.86 seconds for the best performance (27.7%) and 2.30 seconds for the average performance (31.7%). An improvement of TUG performance larger than these MDC values, would mean that a true improvement occurred in the child's performance. These are higher values compared to children with Down syndrome (reported MDC =1.26 seconds [19]), but rather similar to another sample consisting of typically developing preschoolers (SEM=0.68 seconds and MDC=1.88 seconds (28.5%), *calculated based on the*

reported standard deviation and ICC [10]). The protocol by Williams et al. (2005) applied in typically developing children, comprises self-selected walking speed, touching a star on the wall and averaging three trials. When we averaged TUG performance in our protocol, the ICC values, SEM and MDC values are higher compared to theirs (ICC=0.61 [10] versus ICC=0.75 in our study), indicating our TUG results are more consistent but also contain more standard error. However, when we used best performance, similar MDC values and higher ICC values (0.78) compared to Williams et al. (2005) were noted, meaning the reproducibility of this outcome over time was higher overall.

The similarities in standard error (high SEM and MDC values) between Williams et al. (2005) and our test results, indicate that repeated performances in this specific age group (age 3-5 years old) are quite variable, which is a typical feature of this developmental stage [24]. We therefore investigated the reproducibility of each trial recorded within one session. Indeed, the ICCs for within-session reliability were only moderate in our sample of preschoolers. This highlights the children's variable performances within one session, which induces a larger standard error thereby influencing the ICC values [23]. So far, no ICCs for within-session reliability have been published specifically for preschool children, but they have been reported for groups partially consisting of preschoolers [10,12]. Clearly, when preschoolers are part of a group with a larger age range, within-session reliability is more consistent (ICC=[0.80;0.93] [10,12]) compared to our results (ICC=0.69). Thus, the moderate agreement between trials of one session is probably a result of the investigated age group. Furthermore, the children's performances tend to improve with each trial, indicating a learning effect. The main question that remains is how many trials are needed to provide a stable performance, which should be investigated in future research.

To map concurrent validity, the balance subscale of the MABC-2 was correlated to the best and average TUG performance. The balance subscale of the MABC-2 was chosen as a reference test for balance control, because this test is often used for screening purposes in children presenting with mild motor impairments. A fair significant correlation with the MABC-2 balance subscale was found for the best TUG performance (r=-0.347, p=0.007), whereas this was insignificant for the average performance (r=-0.251, p=0.053). This suggests the best performance is a valid outcome to assess balance control in

preschoolers. A fair correlation, however, is not very surprising because of the differing task constraints that are being addressed with both tests. The MABC-2 balance subscale combines balance during walking backward, hopping forward and standing on one leg, whereas the TUG assesses balance during walking, turning and transferring from and to a seated position. A review on concurrent validity of the TUG in typically developing children, revealed similar correlation coefficients with other balance tests, consisting of different balance tasks, endorsing the influence of the requested balance tasks on the magnitude of their relationship [2]. This task-dependency of balance control highlights the need for a weighted selection of the most suitable balance assessment considering the child's difficulties.

Study limitations

Only 33% of parents agreed that their child could participate in the present study, which might have induced selection bias. Based on the results of the MABC-2 balance subscale, 22% (n=7/32) of our children did show balance deficits, which exceeds the expected 16%. This means that both children with normal balance and balance deficits, without any known disorders that might impede balance control, participated in this study. This makes the current sample representative for children in regular preschools as they also comprise a mix of typically developing children and children at risk for mild motor problems. Hence, the modified TUG protocol can be used reliably in a heterogeneous sample of children in regular preschools. However, the modified TUG protocol's reliability and validity have not yet been investigated in children with established diagnoses known to be associated with balance problems, such as children with sensorineural hearing loss, developmental coordination disorder, or cerebral palsy. Furthermore, the modified TUG protocol's sensitivity, specificity and responsiveness need to be established in future research, preferably in multiple populations.

The protocol set-up requires the investigator to commence timing on the cue 'go' and not when the child's bottom leaves the seat. Starting at the cue introduces a cognitive element, as this considers response time, to what is essentially a test of motor or physical ability. The effect of this cognitive element should be investigated in future research. Also, in case the practice trial revealed that the child did not correctly understand the instructions, a test demonstration was given. This means that in rare cases the child also had a demonstration, while the majority only had the instructions.

As suggested by the original protocol [6], three trials should be administered. Here we show that the best performance of these trials is reliable and valid in preschoolers. However, there seems to be a practice effect after the second trial in this sample as the third trial differs significantly from the other two. In our previous study with a larger sample (n=172) [13], there was a significant difference between the three trials, suggesting a practice effect occurs after the first trial. Thus, even though the best performance after three trials is reliable and valid, further investigations into the exact number of trials is needed. The ideal number of trials allows for a true representation of the child's ability, which is characterized by natural variability in performance, but also for the occurrence of a plateau, indicating practice is no longer an issue.

Finally, the children's weight was not measured during the assessments. As overweight children experience difficulties in their dynamic balance during walking [25], the modified TUG in overweight preschoolers could be of interest to identify such problems in future research.

In conclusion, based on the reliability and validity results of this modified TUG protocol, we recommend to use the best performance of three trials in preschoolers, thereby reducing the influence of intra-individual variability, that is prominently present in this specific age group. To implement the protocol in clinical practice, its psychometric properties should be further investigated in children with specific underlying pathologies.

Compliance with ethical statements

Conflict of interest

The authors declare that there is no conflict of interest.

Funding

There is no source of funding.

Ethical approval

The study was approved by the ethical committee of the Antwerp University Hospital (B300201629848).

Informed consent

All parents gave written informed consent before their child participated in the study.

References

1. Nicolini-Panisson RD, Donadio MV (2013) Timed "Up & Go" test in children and adolescents. Revista paulista de pediatria : orgao oficial da Sociedade de Pediatria de Sao Paulo 31 (3):377-383. doi:10.1590/s0103-05822013000300016

2. Verbecque E, Lobo Da Costa PH, Vereeck L, Hallemans A (2015) Psychometric properties of functional balance tests in children: a literature review. Developmental medicine and child neurology 57 (6):521-529. doi:10.1111/dmcn.12657

3. Baque E, Barber L, Sakzewski L, Boyd RN (2017) Randomized controlled trial of web-based multimodal therapy for children with acquired brain injury to improve gross motor capacity and performance. Clinical rehabilitation 31 (6):722-732. doi:10.1177/0269215516651980

4. Cetin SY, Erel S, Bas Aslan U (2019) The effect of Tai Chi on balance and functional mobility in children with congenital sensorineural hearing loss. Disability and rehabilitation:1-8. doi:10.1080/09638288.2018.1535629

5. Katz-Leurer M, Rotem H, Keren O, Meyer S (2009) The effects of a 'home-based' task-oriented exercise programme on motor and balance performance in children with spastic cerebral palsy and severe traumatic brain injury. Clinical rehabilitation 23 (8):714-724. doi:10.1177/0269215509335293

6. Podsiadlo D, Richardson S (1991) The timed "Up & Go": a test of basic functional mobility for frail elderly persons. Journal of the American Geriatrics Society 39 (2):142-148

7. Verbecque E, Schepens K, There J, Schepens B, Klingels K, Hallemans A (2019) The Timed Up and Go Test in Children: Does Protocol Choice Matter? A Systematic Review. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 31 (1):22-31. doi:10.1097/pep.0000000000000558

8. Butz SM, Sweeney JK, Roberts PL, Rauh MJ (2015) Relationships among age, gender, anthropometric characteristics, and dynamic balance in children 5 to 12 years old. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 27 (2):126-133. doi:10.1097/pep.0000000000000128

9. Itzkowitz A, Kaplan S, Doyle M, Weingarten G, Lieberstein M, Covino F, Vialu C (2016) Timed Up and Go: Reference Data for Children Who Are School Age. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 28 (2):239-246. doi:10.1097/pep.0000000000000239

10. Williams EN, Carroll SG, Reddihough DS, Phillips BA, Galea MP (2005) Investigation of the timed 'up & go' test in children. Developmental medicine and child neurology 47 (8):518-524

11. Habib Z, Westcott, S., Valvano, J. (1999) Assessment of balance abilities in Pakistani children: a cultural perspective. Pediatr Phys Ther 11:73-82

12. Nicolini-Panisson RD, Donadio MV (2014) Normative values for the Timed 'Up and Go' test in children and adolescents and validation for individuals with Down syndrome. Developmental medicine and child neurology 56 (5):490-497. doi:10.1111/dmcn.12290

13. Verbecque E, Vereeck L, Boudewyns A, Van de Heyning P, Hallemans A (2016) A Modified Version of the Timed Up and Go Test for Children Who Are Preschoolers. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 28 (4):409-415. doi:10.1097/pep.0000000000000293

14. Baque E, Barber L, Sakzewski L, Boyd RN (2016) Test-re-test reproducibility of activity capacity measures for children with an acquired brain injury. Brain injury 30 (9):1143-1149. doi:10.3109/02699052.2016.1165869

15. Carey H, Martin K, Combs-Miller S, Heathcock JC (2016) Reliability and Responsiveness of the Timed Up and Go Test in Children With Cerebral Palsy. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 28 (4):401-408. doi:10.1097/pep.0000000000000301

16. Newman MA, Hirsch MA, Peindl RD, Habet NA, Tsai TJ, Runyon MS, Huynh T, Zheng N (2018) Reliability of the sub-components of the instrumented timed up and go test in ambulatory children

with traumatic brain injury and typically developed controls. Gait & posture 63:248-253. doi:10.1016/j.gaitpost.2018.05.014

17. Dhote SN, Khatri PA, Ganvir SS (2012) Reliability of "Modified timed up and go" test in children with cerebral palsy. Journal of pediatric neurosciences 7 (2):96-100. doi:10.4103/1817-1745.102564

18. Katz-Leurer M, Rotem H, Lewitus H, Keren O, Meyer S (2008) Functional balance tests for children with traumatic brain injury: within-session reliability. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 20 (3):254-258. doi:10.1097/PEP.0b013e3181820dd8

19. Martin K, Natarus M, Martin J, Henderson S (2017) Minimal Detectable Change for TUG and TUDS Tests for Children With Down Syndrome. Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association 29 (1):77-82. doi:10.1097/pep.0000000000000333

20. Iatridou G, Dionyssiotis Y (2013) Reliability of balance evaluation in children with cerebral palsy. Hippokratia 17 (4):303-306

21. Smits-Engelsman BCM (2010) Movement ABC-2-NL Dutch manual. Pearson, Amsterdam

22. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine 15 (2):155-163. doi:10.1016/j.jcm.2016.02.012

23. Portney LG, Watkins, MP. (2009) Foundations of Clinical Research Applications to Practice, vol 3rd ed. Pearson Prentice Hall, Upper Saddle River, New Jersey

24. Gallahue D, Ozmun J (2006) Understanding Motor Development: Infants, Children, Adolescents, Adults. 6th edn. McGraw-Hill, New York

25. Steinberg N, Nemet D, Pantanowitz M, Eliakim A (2018) Gait Pattern, Impact to the Skeleton and Postural Balance in Overweight and Obese Children: A Review. Sports (Basel) 6 (3). doi:10.3390/sports6030075

Tables and Figures

Table 1

Table 1: Sample characteristics

	All children	**3-year-olds**	**4-year-olds**	**5-year-olds**
Boys/Girls (n)	15/17	1/8	6/5	8/4
Age (months)				
▪ *Mean (SD)*	55.0 (9.9)	41.9 (2.8)	54.5 (2.5)	65.3 (2.9)
MABC-2-balance (percentile)				
▪ *Mean (SD)*	50.0 (30.7)	44.9 (30.2)	54.8 (37.8)	41.3 (24.5)
▪ *Min-Max*	5.0-99.5	9.0-99.0	5.0-99.5	5.0-95.0
▪ *<P16 (n)*	7	2	3	2

Legend: **n**: number; **SD**: standard deviation; **Min**: minimum; **Max**: maximum; **P16**: percentile 16; MABC-2: Movement Assessment Battery for Children second edition.

Table 2

Table 2: Within-session reliability of the modified Timed Up and Go test

	TUG-1						**Reliability**	
	Trial 1 (s)		**Trial 2 (s)**		**Trial 3 (s)**		**ICC**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*95% CI*
Rater 1	7.66	2.08	7.45	1.72	6.99	1.36	0.69[A]	0.50-0.82
Rater 2	7.37	1.88	7.18	1.58	6.73	1.37	0.69[A]	0.50-0.83

Legend: **SD**: standard deviation; **ICC**: Intra-class correlation coefficient; **s**: seconds; **TUG-1**: baseline TUG administration. [A]= two-way mixed effects model, consistency, single rater, ICC (2,1).

Table 3

Table 3: Interrater and test-retest reliability of the modified Timed Up and Go test

		TUG Performance (s)				**Reliability**					
		Rater 1		**Rater 2**		**ICC**		**SEM**	**MDC_{95}**	**Measurement error**	
Interrater reliability (TUG-1)		*Mean*	*SD*	*Mean*	*SD*	*Mean*	*95% CI*			*Mean*	*95% CI*
	B3	6.59	1.22	6.37	1.24	0.99^{A}	$0.97\text{-}0.99^{A}$	0.14 s	0.39 s	0.20	[-0.21;0.61]
	A3	7.19	1.43	6.91	1.39	0.99^{A}	$0.98\text{-}0.99^{A}$	0.15 s	0.42 s	0.26	[-0.15;0.67]
Test-retest reliability (TUG-1 & TUG-2)		**TUG-1**		**TUG-2**		**ICC**		**SEM**	**MDC_{95}**	**Measurement error**	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*95% CI*			*Mean*	*95% CI*
	B3	6.70	1.43	6.32	1.31	0.78^{B}	$0.50\text{-}0.91^{B}$	0.67 s	1.86 s	0.39	[-1.30;2.08]
	A3	7.25	1.66	7.12	1.34	0.75^{B}	$0.50\text{-}0.90^{B}$	0.83 s	2.30 s	0.13	[-1.99;2.24]

Legend: **SD**: standard deviation; **ICC**: Intra-class correlation coefficient; **MDC_{95}**: Minimal Detectable Change (formula: $SEM*1.96*2^{1/2}$); **s**: seconds; **SEM**: Standard Error of Measurement (formula: $SD_{test\ 1}* (1\text{-}ICC)^{1/2}$); **B3**: best of three trials; **A3**: average of three trials; **TUG-1**: baseline TUG administration; **TUG-2**: retest TUG administration. ^{A}= two-way random effects model, consistency, single measurement, ICC (3,1); ^{B}= two-way mixed effects model, absolute agreement, single rater, ICC (2,1).

Figure legends

Figure 1: Distribution of TUG test results compared between trials, raters and test sessions

Figure 2: Bland-Altman plots for the measurement error of the TUG results

Figure 1: Distribution of TUG test results compared between trials, raters and test sessions

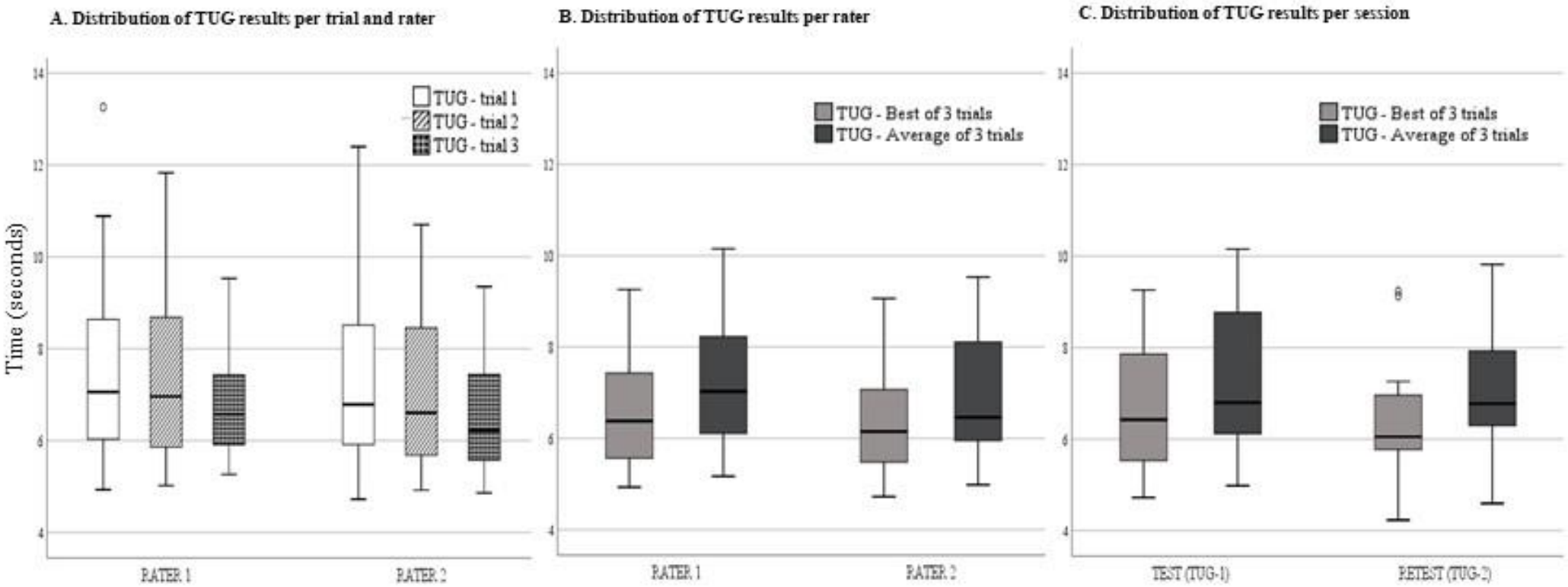
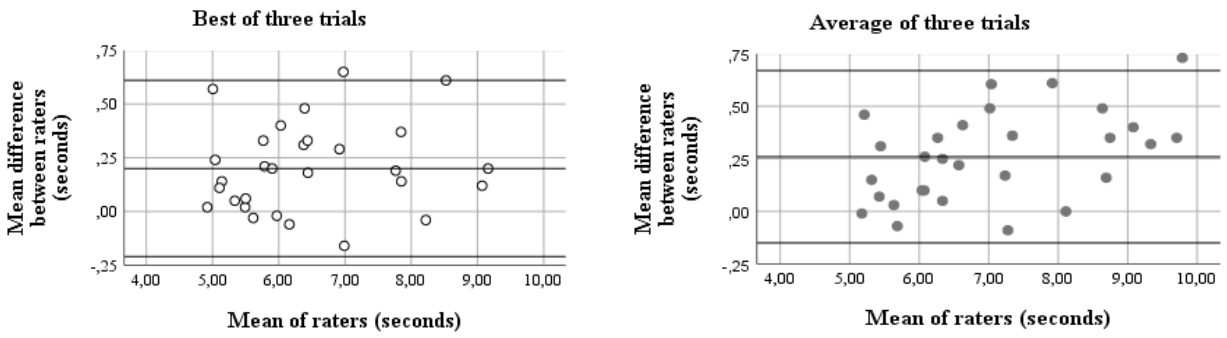
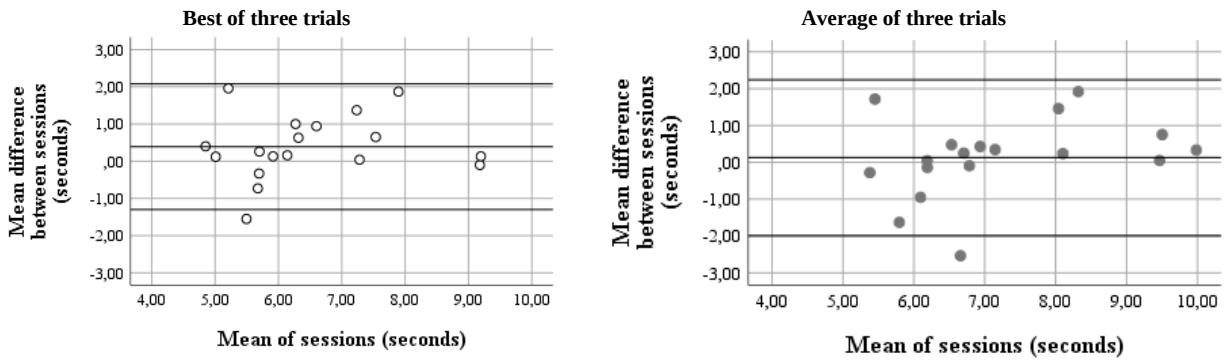


Figure 2: Bland-Altman plots for the measurement error of the TUG results

A. Interrater reliability (best of three trials versus average of three trials)



B. Test-retest reliability (best of three trials versus average of three trials)



Legend: horizontal lines represent the mean difference and the upper and lower bound of the 95% confidence interval.