The Trunk Impairment Scale - modified to ordinal scales in the Norwegian version

Bente Gjelsvik[1,5], Kyrre Breivik[2], Geert Verheyden[3], Tori Smedal[1], Håkon Hofstad[4,5], Liv Inger Strand[5,1]

[1]Department of Physiotherapy, Haukeland University Hospital, Bergen, Norway

[2]Centre for Child and Adolescent Mental Health, Uni Health, Bergen, Norway

[3]University of Applied Sciences, Bochum, Germany

[4]Department of Physical Medicine and Rehabilitation, Haukeland University Hospital, Bergen, Norway

[5]Department of Public Health and Primary Health Care, Physiotherapy Research Group, University of Bergen, Norway

Abstract

Purpose: To translate the Trunk Impairment Scale (TIS), a measure of trunk control in patients with stroke, into Norwegian (TIS-NV), and to explore its construct validity, internal consistency, intertester and test-retest reliability.

Method: The TIS was translated according to international guidelines. 201 patients with acute stroke were recruited for the validity study, and 50 inpatients with acquired brain lesions were recruited for the study of intertester and test-retest reliability. Construct validity was analysed using explorative factor analysis, confirmatory factor analysis and item response theory, internal consistency with Cronbach's alpha test, and intertester and test-retest reliability with kappa and intraclass correlation coefficient tests.

Results: The back-translated version of TIS-NV was validated by the original developer. The subscale Static sitting balance was removed from the test. Six testlets were hierarchically constructed by combining items from the subscales Dynamic sitting balance and Coordination, and renamed modified TIS-NV (TIS-modNV). After these modifications the TIS-modNV fitted well to a locally dependent unidimensional item response theory model. The test demonstrated good construct validity, excellent internal consistency, as well as high intertester and test-retest reliability for the total score.

Conclusions: The TIS-modNV is a valid and reliable scale for use in clinical practice and research.

Main text

Patients with disability due to neurological lesions constitute one of the greatest challenges for society and health services in developed countries [1]. The most common cause of brain damage in adults is stroke, and in Norway approximately 15.000 persons suffer a stroke each year [2]. Rehabilitation should be beneficial for the individual patient as well as for society [3], and adequate assessment tools are needed to examine relevant functional aspects.

Impaired balance is a common physical deficit post stroke [4;5], and improved balance has been found to be associated with improved rehabilitation outcomes [6], ability to perform daily activities [7], and walking [8]. Impaired balance increases the risk of falls [9], and may thus imply social problems and high economic costs [10]. The trunk seems particularly important for balance as it stabilizes the pelvis and spinal column [11], being a prerequisite for coordinated use of the extremities in functional activities such as reaching and gait [12]. Impaired trunk control seems common post stroke [13], and trunk control assessed in patients early after stroke has been found predictive of long-term functional improvement [14;15] and length of institutional stay [16;17].

To adequately assess function and disability, therapists need assessment tools for the different domains of function according to the International Classification of Functioning, Disability and Health (ICF) [18]. The Trunk Impairment Scale (TIS) addresses the body domain of the ICF [19;20], and was developed to evaluate postural control of the trunk in patients suffering from stroke [13]. The TIS originally consists of three subscales; Static sitting balance, Dynamic sitting balance, and

Coordination, containing 3, 10 and 4 items, respectively. Patients must be able to sit independently for 10 seconds to be tested. The test has not demonstrated a ceiling effect, and is therefore appropriate to use in a wide range of functional deficits in patients suffering from stroke [17].

Studies using classical test theory (CTT) have found measurement properties of the TIS to be satisfactory for different patient populations: stroke [13], Parkinson's disease [21], multiple sclerosis [22] and traumatic brain injury [23]. Good ability to predict function over time was furthermore demonstrated in patients with sub-acute stroke [24].

Even if previous studies using CTT have given important psychometric information, there are several problems with the assumptions underlying CTT such as sample dependency, item equivalence and standard error of measurement [25]. If the data can meet certain rather strict assumptions, Item Response Theory (IRT) overcomes many of these limitations [26;27]. IRT also provides rather sophisticated psychometric information that is difficult to obtain by the use of CTT. Two important assumptions of traditional IRT models are that the scale must be essentially unidimensional and the individual items of the scale locally independent [27]. The local independency assumption can be relaxed in certain situations, e.g. if it has a negligible impact on the IRT parameters [28]. Alternatively, local dependency might be taken into account directly in the model by using measurement models such as a bifactor model [28;29] or a locally dependent unidimensional IRT model [30].

In IRT, degree of trunk control is considered as a latent variable, which has a relationship with each item that is described by an item characteristic curve. This curve illustrates how the probability of affirming an item is conditioned on the respondent's trait level [31]. Different IRT models are various equations for modelling the item characteristic curve. In Rasch models, the item characteristic curves are allowed to vary in the difficulty/severity parameter which reflects the location on the trait where an individual has a 50% chance of endorsing or passing the specific item. The Rasch model allows conversion of raw data into interval scores, however, it is particularly restrictive as it assumes that the items should be equally related (equal discrimination parameters) to the latent construct in question. In less restricted IRT models, the item characteristic curves are also allowed to vary in their discrimination parameter (guessing parameters are not considered here) which depicts how well the item differentiates between individuals with different levels on the latent construct (for an introduction to IRT, see [32]).

Verheyden and Kersten [33] used Rasch analysis to investigate the internal validity of the TIS subscales, resulting in removal of the subscale Static sitting balance due to a high ceiling effect and not fitting the Rasch model. The Dynamic sitting balance and Coordination subscales were initially not found to fit the Rasch model due to local dependency between two or more items, but fit was achieved by combining the problematic items into testlets. A testlet consists of a group of items related to a single content area that is developed as a unit [34]. We explored whether our data fitted better to the Rasch model and other less restricted IRT models. In contrast to Verheyden and Kersten [33] we focused on the total scale and hypothesized that a strong general factor would underlie the subscales. Moreover, from a clinical point of

view we regard the total scale as important, as its score is meant to reflect the degree of trunk control in sitting, and such information might for instance be important for prognostic estimation.

The aim of the present study was first to translate the TIS into Norwegian (TIS-NV), and then to explore its construct validity, internal consistency, and intertester and test-retest reliability.

METHOD

The methods are described in three steps; translation and cross-cultural adaptation of TIS, examination of construct validity and internal consistency of the measure, and finally examination of intertester and test-retest reliability.

Translation and cross-cultural adaptation

We translated the TIS into Norwegian following international guidelines [35] after consent from the test developer. Three bi-lingual physiotherapists translated the TIS separately into Norwegian. The three versions were compared, and consensus was reached for a first draft. This draft along with the individually translated versions were further discussed by an expert panel consisting of three neurorehabilitation physiotherapists, all knowledgeable in English and research methodology, and compared with the original English version. Consensus was reached for a second Norwegian draft of TIS. This version was examined clinically, and adjustments were made in cooperation between the translators and the clinicians, resulting in a final Norwegian version, named TIS-NV. A bi-lingual colleague with no previous experience with the TIS translated the TIS-NV back into English.

Construct validity and internal consistency

A cross-sectional design was used. All patients admitted to the Stroke unit at the Department of Neurology (Haukeland University Hospital) between December 2008 and September 2010, were considered for inclusion. Eligible patients had to live in Bergen and at home prior to the stroke, be included 2-7 days after stroke onset and within 120 hrs after admission to the stroke unit, be awake and give informed consent either by themselves or their carers, and achieve a score between 2 and 26 on the National Institutes of Health Stroke Scale (NIHSS) [36-38]. Exclusion criteria were serious psychological illness, drug addiction, co-morbidity that might affect the progress from stroke, or poor knowledge of Norwegian.

Information about age, gender, type of brain lesion, lesion side, most affected body side and time since stroke were collected for all participants. Three physiotherapists were responsible for testing the patients as soon as possible after inclusion with several clinical tests, including TIS-NV, Postural Assessment Scale for Stroke [39], 5m timed walk [40] and timed Up-and-Go [41]. In order to standardize the test procedure, the physiotherapists underwent training for all measures. All patients were tested in a separate room at the physiotherapy department.

Intertester and test-retest reliability

A cross-sectional design was used for the intertester study, and a longitudinal design for the test-retest reliability study. Patients with stroke or other brain damage were recruited by their treating physiotherapists from the Department of Physical Medicine and Rehabilitation (Haukeland University Hospital) between May and September 2009 and between May and September 2010. The included patients were in a subacute or chronic stage post brain injury and involved in multidisciplinary inpatient

rehabilitation, understood verbal instructions, were able and willing to give informed consent, and had no other physical or mental disorders that could affect performance of the TIS-NV.

Information about age, gender, type of brain damage, lesion side, most affected body half and time since brain damage were collected. Two neurorehabilitation physiotherapists; SD and BG, performed the testing. SD worked mainly with patients suffering from stroke for the last 8 years and attended basic and advanced Bobath courses. BG is an advanced Bobath Instructor (IBITA[1]).

The test procedure was standardised for all patients: the location was the same, all patients received the same instructions for the TIS-NV from tester 1 (BG), and performed each test item three times. Patients were tested simultaneously but scored independently by both testers, and again two hours later by BG alone. Test scores were not summarized to avoid BG remembering the results of the first test.

Statistical analysis

For examining construct validity, explorative factor analysis (EFA), confirmatory factor analysis (CFA) and IRT analyses were carried out by the use of the Mplus 6.0 program [42] using the WLSMV estimator (Weighted Least Squares with Mean and Variance adjustments). This particular estimator takes the ordinal nature of the data into account [43]. The IRT parameters (graded response parameters; [44]) were derived by translation of the CFA parameters by the use of formulas described by Brown [45]. Six testlets making ordinal scales were constructed from the items of the subscales Dynamic sitting balance and Coordination, and further analysed using

[1] International Bobath Instructors' Training Association, IBITA

CFA. The graded response model is a popular IRT model, when estimating ordered polytomous (>2 categories) data. In this particular model each item has one discriminate parameter (alpha) but as many difficulty parameters (thresholds, beta's) as there are response categories minus one. In the present study, all of the testlets, except two, had three thresholds. The remaining two testlets (3 and 4) had only two thresholds as they were constructed by the use of two original items instead of three. In line with most research, the latent construct was scaled to have a mean of 0 and standard deviation of 1.

The unidimensional assumption of the IRT model was tested by the use of explorative and confirmatory factor analysis. In CFA the unidimensional assumption of traditional IRT models was tested by the use of testing the fit of a 1 factor model in CFA, assessed by the use of chi square, Bentler's Comparative Fit Index (CFI; [46]) and the root mean square error of approximation (RMSEA, [47]). CFI ≥ 0.96 and RMSEA ≤ 0.05 have been proposed as cut off values for indicating good fit when using categorical indicators [48]. In EFA the unidimensional assumption was tested by assessment of the eigenvalues, where a high ratio (e.g. >3) of the first over the second eigenvalues was considered as supporting essential unidimensionality [31]. To assess local independence, modification indexes of the one factor model was explored to see whether there were any non ignorable correlations (r>=0.20) between the items error variances after the latent variable was taken into account.

All collected data on the TIS-NV were transformed to the six testlets before analysing internal consistency and reliability, using the software programme PASW 18 (SPSS Inc.). Internal consistency was examined by Cronbach's α. Acceptable value was set at Cronbach's alpha 0.70-0.95. Intraclass correlation coefficients (ICC) were calculated to examine relative and absolute intertester and test-retest reliability of the

total score. Both ICC 1.1 and ICC 3.1 were used to examine whether there was a systematic error in scores between the two testers and between repeated measurements. If no systematic error was part of the variability, the value of ICC 3.1=ICC 1.1. Reference values for ICC: < 0.50=low; 0.50-0.69=moderate; 0.70-0.89=high, and 0.90-1.00 very high [49].

The within subject standard deviation (S_w) is a value of absolute reliability, expressed in the unit of the measurement tool. For intertester reliability, the difference between a score and the true value of an individual is expected to be less than 1.96 S_w for 95% of the observations. The difference between two repeated measurements of the same individual is expected to be less than $\sqrt{2}$ X 1.96 S_w = 2.77 S_w for 95% of the observations [50], called the smallest detectable change (SDC) [51].

Reliability of the separate testlets was examined by *kappa* statistics. Reference values for kappa (*κ)*: < 0.20=poor, 0.21-0.40=weak; 0.41-0.60=moderate, 0.61-0.80=high, and 0.81-1.0=very high [52]. A prerequisite for the use of kappa is a symmetrical cross-table based on the same scoring alternatives being used by the two testers or by repeated testing [52]. Percentage agreement (%) was used when kappa could not be calculated, 80% agreement considered acceptable.

The study was approved by the Regional Committee for Medical and Health Research Ethics in Western Norway.

RESULTS

Translation and cross-cultural adaptation

Some of the terms used in the TIS were not straight forward to translate, for instance, the word “should” in the Coordination subscale, items 1 and 3, could be interpreted as “ought to” or “must”. The understanding of the items was discussed with the test developer, and consensus regarding interpretation and phrasing was reached for both the English and Norwegian versions. The back-translated version of the TIS-NV was validated by the original developer. TIS-NV formed the basis for the next part of the study; the examination of measurement properties.

Construct validity and internal consistency

A total of 201 patients with stroke were assessed for the present study (table 1). More male than female patients participated, and most had ischemic strokes with an even distribution between hemispheres for the localisation of the strokes.

Insert table 1 about here

Initially, we examined whether the items of the TIS-NV fitted a unidimensional CFA model. A poor fit was demonstrated, both according to the chi-square = 563.70, df = 119, $p<0.001$ and the RMSEA fit index (RMSEA=0.136, CFI=0.93). Post-hoc modification indexes revealed that this poor fit was mainly due to local dependence between Dynamic sitting balance items 1-3, 4-6, 7-8 and 9-10, Coordination items 1-2 and 3-4. Most of the patients (96%) obtained the maximum score on item 1 of the Static sitting balance subscale and the correlation between items 2 and 3 on this scale was very high (r=0.98). This subscale was therefore removed. Based on clinical judgement, testlets were constructed making hierarchically organized ordinal scales, by combining items within the subscales Dynamic sitting balance and Coordination.

Items 1-3 of Dynamic sitting balance were recoded to testlet 1; items 4-6 to testlet 2; items 7-8 to testlet 3; items 9-10 to testlet 4; items 1-2 of Coordination were recoded to testlet 5; and items 3-4 to testlet 6 (table 2), making the scoring levels mutually exclusive.

Insert table 2 about here

EFA analyses revealed a large ratio (5.7) of the first (4.045) to second eigenvalue (0.710) which was well above the proposed 3.0 cut-off to support essential unidimensionality as there seemed to be one dominant factor. Rerunning the unidimensional CFA model using the six testlets still resulted in a poor fit according to RMSEA index (RMSEA=0.145, CFI=0.96). Modification indexes revealed that there were rather large correlations between the error terms (local dependency) of testlet 1 and 2, and testlets 3 and 4. Allowing these error terms to co-vary in a locally dependent unidimensional IRT model (table 3, model 1), resulted in a very good fit to the data (Chi-square=6.002, df=7, p=0.54; RMSEA=0.00, CFI=1.00).The local dependencies for the latter model were moderate to strong; 0.37 between testlet 1 and 2, and 0.52 between testlets 3 and 4. One plausible way to interpret this model is that it consists of a strong general factor and two smaller content specific factors (testlets 1 - 2 and testlets 3 - 4), which is reflected by the two local dependencies [30]. We interpret these two factors as reflecting problems with lower and upper trunk, respectively.

Insert table 3 about here

The testlets had a strong relationship with the general factor (standardized beta 0.70-0.86) (table 3). Constraining the factor loadings to be equal with each other led to a significantly poorer fit (Delta Chi-square=20.29, df=4, p<0.001), and thus did not support the use of Rasch models. Allowing for local dependencies in the model (MII vs. MI) had a moderate impact on the loadings (especially the loadings associated with testlets 3 and 4). This fact led us to translate the Mplus factor parameters into IRT parameters based on MII which included the correlated error terms.

IRT discriminating parameters (alpha) for testlets 5 and 6 can be classified as rather high (>1.6) (table 3). The difficulty parameters (beta's) ranged from -1.27 to 0.89 dependent on the specific item and the threshold in question. The last threshold (beta 3 on all testlets except 3 and 4 of which beta 2 was the last threshold) was rather similar across testlets. They revealed that an individual had to be 0.43 -0.89 standard deviation above the mean to be likely to pass the particular threshold. There was more diversity with regard to the testlets' first threshold (beta1, ranged from -1.27 to -0.22), where the threshold related to testlets assessing lower trunk control (1 and 2) and coordination (5) were lower than the testlets assessing upper trunk control (3 and 4) and coordination (6). The patients need lesser trunk control to score at least 1 on testlets 1, 2 and 5 than on testlets 3, 4 and 6.

The analyses support the notion of a general underlying factor, which we call "trunk control". After modification of the scale by constructing testlets, the modified TIS-NV was renamed to TIS-modNV (appendix).

The TIS-modNV demonstrated high internal consistency (table 4). Cronbach's alpha did not increase if any of the testlets were deleted, which demonstrated that each testlet contributed to alpha.

Insert table 4 about here

Intertester and test-retest reliability

This part of the analysis was performed with the TIS-modNV on fifty patients with brain lesions of different causes, primarily stroke (table 1).

Intertester reliability. *Kappa* was high for testlet 1, moderate for testlets 2, 4 and 5, and low for testlet 3 (0.40). Kappa could not be calculated for testlet 6, as the two testers had used different response alternatives. This testlet received 80% agreement (table 5). The total sum score demonstrated normal distribution, and ICC 1.1 was 0.77 (95%CI 0.63-0.86), which is high. The SDC was 2.63.

Insert table 5 about here

Test-retest reliability. Forty-nine patients participated in the retest. One patient dropped out of the second test due to poor condition. *Kappa* was high for testlets 1, 3, 4 and 5, low for testlet 2 and moderate for testlet 6 (table 5). ICC 1.1 was high, 0.85, for the total sum score (0.85, 95%CI 0.75-0.91). The SDC was 2.90. Thus, to demonstrate a real improvement in trunk control as measured using the TIS-modNV,

an individual patient must improve 3 points or more on the 0-16 point scale on repeated testing.

The scatter plots (figures 1 and 2) demonstrate that the testlet scale had no ceiling effect.

Insert figures 1 and 2 about here

DISCUSSION

The aim of this study was to translate the TIS into Norwegian and examine psychometric properties of this version in patients with stroke. The original developers used Rasch analysis to examine the possibility for transforming the TIS item scores to interval levels using data from a mixed sample of patients in acute and chronic stages post stroke (n=162). The study resulted in omitting the subscale Static sitting balance [33], and this was in line with our conclusion after examining it in a sample of 201 patients with acute stroke. However, our data did not fit the Rasch model as the items did not seem equally related to the general latent construct. From a clinical point of view, it became evident that several items measured the same ability but to different degrees, and different aspects of trunk control, e.g. lower trunk, pelvis and hip stability (lower trunk) for selective movement of shoulder girdles, and upper trunk and contralateral pelvic stability (upper trunk) for selective movement of the unilateral pelvis, were identified in the construction of testlets. The underlying construct of all the testlets was examined using CFA which demonstrated good

construct validity, and resulted in a modified version (TIS-modNV. Appendix), containing six testlets with hierarchically organized ordinal scales. The TIS-modNV demonstrated good construct validity, excellent internal consistency, as well as high intertester and test-retest reliability for the total score and, can be applied with confidence in clinical practice as well as research.

Translation

Translation should ensure cross-cultural adaptation [35].TIS was developed in Belgium which is a North-European country and culturally similar to Norway, and published in English in 2004 [13]. We believe that we achieved a good translation that reflected the developers' intention.

Construct validity and internal consistency

We wanted to examine the construct validity of the TIS-NV specifically in relation to the Static sitting balance subscale, as this subscale could be more relevant for use in the acute stroke population. Our sample contains data from 201 patients with acute stroke, which is well above the minimum number (N=100) of subjects recommended by Terwee et al. [51] to be included in a factor analysis. Modeling the underlying general construct by the use of IRT turned out to be complex. First, a total of 96% of our participants obtained the maximum score on item 1 of Static sitting balance. This was surprising since our patients had suffered acute strokes and were mostly tested within 7 days of stroke onset. Based on our results, we support Verheyden and Kersten's [33] decision in maintaining a prerequisite of sitting for 10 seconds in the starting position, and to remove the Static sitting balance subscale from the test. Second, the results of the analyses strongly suggest that the original items should not

be treated as separate when modelling the latent trait. In line with Verheyden and Kersten [33] we found a large degree of local dependency when using the original items. In the present study we combined items that empirically seemed to analyse similar aspect of trunk control, although hierarchically more difficult, into 4 testlets (table 2); Dynamic sitting balance items 1-3 and 4-6 for lower trunk control; 7-8 and 9-10 for upper trunk control. Similarly, the four original Coordination items where recoded into two testlets; 1-2 and 3-4 for lower and upper trunk control respectively, as the original items also seemed to be hierarchically dependent. Finally, the present analyses suggested that a locally dependent unidimensional IRT model [30] was the most appropriate way to model the general trunk control construct when using the TIS-modNV. The testlets did not have a similar relationship with the underlying construct, and did therefore not fit the Rasch model. The data did not fit a traditional IRT either, due to the fact that rather strong local dependencies between two pairs of testlets (relating to lower and upper trunk) existed after the general latent construct was taken into account. We believe that these two local dependencies reflect two content specific factors, relating to lower and upper trunk control, which exists in addition to the general latent construct. When these local dependencies were built into the model, the model had a very good fit to the data.

In the final model, the testlets related to coordination (5, 6) had a noticeably stronger relationship with the underlying latent construct than the testlets assessing lower/ upper trunk control. Lower and upper trunk can be seen as aspects of the construct trunk control as the patient moves in one plane only. The coordination items require an overall trunk control where the stabilizing requirements change between the two sides to allow alternate movement of the opposite sides. This movement requires

dynamic trunk control in three movement planes, and may therefore capture the underlying construct to a greater degree.

The most noticeable finding with regard to the items difficulty parameters was that obtaining the lowest score on the lower trunk (1 and 2) and coordination (5) testlets seemed to be the best indicator of severe trunk impairment. In fact individuals as low as -1.20 standard deviation below the mean of this patient population had at least a 50% chance of obtaining a score on these testlets. Patients may find it easier to stabilise against a base of support and to move the upper trunk than vice versa.

Several studies indicate that trunk control is an important aspect of balance and function [11;53-57]. Impairment in trunk control is a common problem in patients after brain damage [12;14;17;23;24;58-62]. Instability and deficits in movement control constitute some of these impairments. The testlets of the TIS-modNV seem to capture such problems and are therefore relevant indicators of the construct. Additionally, analysis of internal consistency was found to be excellent for the TIS-modNV.

Reliability

Intertester reliability of the total TIS-modNV scores was high in our study (ICC=0.77). *Kappa* was moderate to high for all testlets apart from testlet 3 (0.40), where testers agreed on the scores in 32 out of 50 patients (64%). In testlet 3, the two testers evaluated the patients' ability to lift the pelvis unilaterally while maintaining an upright posture. This movement requires finely tuned coordination between the two sides of the body. When impairments affect coordination and make the movement difficult to

initiate and perform, patients may compensate which makes it difficult for testers to judge whether the movement was “appropriate”, as described in the test guidelines. Furthermore, the two testers were positioned facing the patient, and tester 1 sat straight across the patient to instruct each item, while tester 2 had to sit to one side. This might have affected the viewing angle, causing different evaluation in some cases.

For the total sum of the TIS-modNV, the test-retest analysis demonstrated that there was no systematic shift in the data as ICC 1.1 was identical to ICC 3.1. The test-retest results demonstrated moderate to high *kappa*-values for all testlets, except for testlet 2. Analysis of the cross tables revealed that there was agreement for 30 out of 49 patients (61%), which demonstrated weak test-retest reliability for this testlet. This may have been due to a learning effect, as the patients were performing the original items 1-3 (testlet 1) and 4-6 (testlet 2) nine times in total during both test rounds. No other testlets had the same amount of repetition. The reliability of the sum score seemed to be higher than the reliability of the individual testlets.

Limitations of the present study

Two hours between test and retest was chosen. Time of day, as well as the patients’ stability (or variability) in motor performance could have affected test results. Our intention was to provide no treatment between the test sessions, but this could not be avoided for all patients; a few had occupational therapy, but none had physiotherapy during the two hours. All patients attended active rehabilitation, and a longer time span might deprive patients of treatment, which was considered unethical. Furthermore, participants in the reliability study had a wide range of lesions and

ages, and as such we did not examine a homogeneous group. Using a mixed sample for the reliability study could be seen as a limitation; however, in the time span available, it was not possible to recruit stroke patients only. Nevertheless, our sample should be representative for patients whom therapists meet and treat in a neurorehabilitation unit.

Conclusion and implications for practice and research

Adequate measurement properties were demonstrated for the TIS-modNV, allowing Norwegian physiotherapists to evaluate trunk control with a reliable and valid scale in Norwegian language. The results from the present study suggest that the testlet scale should be used instead of the original scale by both researchers and clinicians. Moreover, when interested in obtaining specific patients' standing on the general latent construct, the most reliable score is probably gained by calculation of their estimated IRT factor score derived directly from the statistical model. Such a score would take the correlated error terms and the differential weighting of the items into account. Being aware of the fact that the use of factor scores is often not practical in clinical settings, we believe the simple sum score of the testlets should be a viable option. As all the testlets have reasonable high loadings on the general factor, we believe a simple sum score should reflect this general factor to a high degree [63].

More research is clearly needed on the practical use of this scale. For instance, it would be of great interest to explore the relative merit of using the total scale versus the specific testlets in predicting clinical outcomes. Even if we believe that the total scale will often be the best choice due to the higher reliability, it is far from certain that this will always be the case. Whether specific lesion localisations lead to specific

impairments in trunk control, as explored by analysis of the individual testlets, remains to be assessed.

The developments of TIS-NV into TIS-modNV have not changed the original items of the scale, but highlighted the underlying construct and how the items should be constructed and scored. The individual testlets may give guidelines for treatment, while the total sum of the testlets is recommended for use as an outcome measure in clinical practice. It is recommended that therapists using the TIS-modNV as well as the previous versions should train themselves in the observation and scoring, in order to score as reliable as possible.

Acknowledgements

The authors wish to thank the Department of Physiotherapy at Haukeland University Hospital, and specifically physiotherapists Helene Christiansen and the members of the expert panel Kari Øen Jones, Olav Gjelsvik† and Torunn Grenstad for active participation in the translation process, Mona Kristin Aaslund for back-translating the TIS-NV, Torunn Grenstad, Veronica Bøe, Odd Arne Bergset and Silje Daltveit for their dedicated work in testing the patients. Silje Daltveit also collected and plotted the test results for the validity study.

Grants for the study have been received by Bente Gjelsvik from Haukeland University Hospital, Western Norway Health Region and the Norwegian Fund for Post-Graduate Training in Physiotherapy.

Declarations of interest

The authors report no declarations of interest.

APPENDIX

TRUNK IMPAIRMENT SCALE – Modified Norwegian version (TIS-modNV)

Forutsetning: pasienten kan opprettholde utgangsstillingen i 10 sek.
Utgangsstillingen for hver deltest er den samme: Pasienten sitter på kanten av en seng eller behandlingsbenk uten rygg- og armstøtte. Lårene har full kontakt med sengen eller benken, føttene har hoftebreddes avstand og er plassert flatt på gulvet. Pasient er barfot. Knevinkelen er 90^0. Armene hviler på beina. Dersom det er hypertonus til stede, regnes posisjonen i affisert arm som en del av utgangsstillingen. Hodet og trunkus er i midtlinjeposisjon.

1.	Utgangsstilling. *Pasienten instrueres i å berøre sengen eller benken med **den mest affiserte albue** (ved å forkorte den mest affiserte siden og forlenge den minst affiserte siden) og returnere til utgangsstillingen.*	
	INSTRUKSJON: **Kan du berøre sengen/benken med …albue?**	
	Pasienten faller, trenger støtte fra en arm eller albuen berører ikke sengen eller benken	0
	Pasienten beveger aktivt uten hjelp, albuen berører seng eller benk, men uten passende trunkal forkorting/forlengning	1
	Pasienten viser passende forkorting/forlengning, men med kompensasjon	2
	Pasienten beveger uten kompensasjon	3
	(Mulige kompensasjoner er: (1) bruk av arm, (2) kontralateral hofteabduksjon, (3) hoftefleksjon (dersom albuen berører seng eller benk lenger distalt enn proksimale halvdel av femur), (4) knefleksjon, (5) føttene glir	
2.	Utgangsstilling. *Pasienten instrueres i å berøre sengen eller benken med **den minst affiserte albue** (ved å forkorte den mest affiserte siden og forlenge den minst affiserte siden) og returnere til utgangsstillingen.*	
	INSTRUKSJON: Kan du gjøre det samme igjen, men til motsatt side?	
	Pasienten faller, trenger støtte fra en arm eller albuen berører ikke sengen eller benken	0
	Pasienten beveger aktivt uten hjelp, albuen berører seng eller benk, men uten passende trunkal forkorting/forlengning	1
	Pasienten viser passende forkorting/forlengning, men med kompensasjon	2
	Pasienten beveger uten kompensasjon	3
	(Mulige kompensasjoner er: (1) bruk av arm, (2) kontralateral hofteabduksjon, (3) hoftefleksjon (dersom albuen berører seng eller benk lenger distalt enn proksimale halvdel av femur), (4) knefleksjon, (5) føttene glir)	
3.	Utgangsstilling. *Pasienten instrueres i å løfte **mest affisert bekkenhalvdel** fra sengen eller benken (ved å forkorte mest affisert side og forlenge minst affisert side) og returnere til utgangsstilling*	
	INSTRUKSJON: **Kan du løfte… hofte/bekkenhalvdel?**	
	Pasienten viser ingen eller omvendt trunkal forkorting/forlengning	0
	Pasienten viser passende trunkal forkorting/forlengning, men med kompensasjon	1
	Pasienten viser passende forkorting/forlengning og beveger seg uten kompensasjon	2
	(Mulige kompensasjoner er: (1) bruk av armer, (2) skyver fra med ipsilateral fot (hælen mister kontakt med gulvet)	
4.	Utgangsstilling. *Pasienten instrueres i å løfte **minst affisert bekkenhalvdel** fra sengen eller benken (ved å forkorte mest affisert side og forlenge minst affisert side) og returnere til utgangsstilling*	
	INSTRUKSJON: Kan du gjøre det samme på andre siden?	
	Pasienten viser ingen eller omvendt trunkal forkorting/forlengning	0
	Pasienten viser passende forkorting/forlengning, men med kompensasjon	1
	Pasienten viser passende forkorting/forlengning og beveger seg uten kompensasjon	2
	(Mulige kompensasjoner er: (1) bruk av armer, (2) skyver fra med ipsilateral fot (hælen mister kontakt med gulvet)	
5.	Utgangsstilling. *Pasienten instrueres i å **rotere øvre del av trunkus 6 ganger** (hver skulder skal beveges fremover 3 ganger), **mest affisert** side beveges først, hodet bør holdes i ro i utgangsstillingen.*	
	INSTRUKSJON: **Roter vekselvis øvre del av kroppen 3 ganger. Hold hodet i ro. Start med å bevege…side frem.**	
	Mest affisert side beveges ikke 3 ganger	0
	Rotasjon er asymmetrisk	1
	Rotasjon er symmetrisk	2
	Rotasjon er symmetrisk, og oppgaven tar mindre enn 6 sekunder	3
6.	Utgangsstilling. *Pasienten instrueres i å **rotere nedre del av trunkus 6 ganger** (hvert kne skal beveges fremover 3 ganger), **mest affisert** side beveges først, øvre del av trunkus bør holdes i ro i utgangsstillingen. Dersom pasienten spontant setter seg lenger ut på kanten av sengen eller benken, tillates dette.*	
	INSTRUKSJON: Skyv vekselvis høyre og venstre kne frem 3 ganger. Hold overkroppen i ro. Start med …side.	
	Mest affisert side beveges ikke 3 ganger	0
	Rotasjon er asymmetrisk	1
	Rotasjon er symmetrisk	2
	Rotasjon er symmetrisk, og oppgaven tar mindre enn 6 sekunder	3
	TIS-modNV total	/16

TIS-modNV - Back-translated version

Prerequisite: the patient can maintain the starting position for 10 secs.
The starting position for each item is the same: The patient is sitting on the edge of a bed or plinth without back and arm support. The thighs make full contact with the bed or plinth, the feet are hip width apart and are positioned flat on the floor. The patient is barefooted. The angle of the knees is 90^0. The arms are resting on the thighs. If there is hypertonia present, the position of the affected arm is counted as part of the starting position. The head and trunk are in a midline position.

1.	*From the starting position, the patient is instructed to touch the bed or plinth with the **most affected elbow** (by shortening the most affected trunk side and elongating the least affected trunk side) and return to the starting position.*	
	The patient falls, needs support from an arm, or the elbow does not touch the bed or plinth	0
	The patient moves actively without help, the elbow touches the bed or plinth, but without appropriate trunk shortening/elongation	1
	The patient demonstrates appropriate trunk shortening/elongation, but with compensations	2
	The patient moves without compensations.	3
	(Possible compensations are: (1) use of arm, (2) contralateral hip abduction, (3) hip flexion (if the elbow touches the bed or plinth more distally than the proximal half of femur), (4) knee flexion, (5) sliding of the feet)	
2.	*From the starting position, the patient is instructed to touch the bed or plinth with the **least affected elbow** (by shortening the least affected trunk side and elongating the most affected trunk side) and return to the starting position.*	
	The patient falls, needs support from an arm, or the elbow does not touch the bed or plinth	0
	The patient moves actively without help, the elbow touches the bed or plinth, but without appropriate trunk shortening/elongation	1
	The patient demonstrates appropriate trunk shortening/elongation, but with compensations	2
	The patient moves without compensations	3
	(Possible compensations are: (1) use of arm, (2) contralateral hip abduction, (3) hip flexion (if the elbow touches the bed or plinth more distally than the proximal half of femur), (4) knee flexion, (5) sliding of the feet)	
3.	*From the starting position, the patient is instructed to lift the **most affected side of the pelvis** from the bed or plinth (by shortening the most affected trunk side and elongating the least affected trunk side) and return to the starting position.*	
	The patient demonstrates no or the opposite trunk shortening/elongation	0
	The patient demonstrates appropriate trunk shortening/elongation, but with compensations	1
	The patient demonstrates appropriate trunk shortening/elongation and moves without compensations	2
	(Possible compensations are: (1) use of upper extremities, (2) pushing off with the ipsilateral foot (the heel loses contact with the floor))	
4.	*From the starting position, the patient is instructed to lift the **least affected side of the pelvis** from the bed or plinth (by shortening the most affected trunk side and elongating the least affected trunk side) and return to the starting position.*	
	The patient demonstrates no or the opposite trunk shortening/elongation	0
	The patient demonstrates appropriate trunk shortening/elongation, but with compensations	1
	The patient demonstrates appropriate trunk shortening/elongation and moves without compensations	2
	(Possible compensations are: (1) use of upper extremities, (2) pushing off with the ipsilateral foot (the heel loses contact with the floor))	
5.	*From the starting position, the patient is instructed to **rotate the upper part of the trunk 6 times** (each shoulder must be moved forwards 3 times); the **most affected** side moves first, the head should be maintained in the starting position.*	
	The most affected side is not moved 3 times	0
	The rotation is asymmetrical	1
	The rotation is symmetrical	2
	The rotation is symmetrical and the task takes less than 6 seconds	3
6.	*From the starting position, the patient is instructed to **rotate the lower part of the trunk 6 times** (each shoulder must be moved forwards 3 times); the **most affected** side moves first, the head should be maintained in the starting position.*	
	The most affected side is not moved 3 times	0
	The rotation is asymmetrical	1
	The rotation is symmetrical	2
	The rotation is symmetrical and the task takes less than 6 seconds	3
	TIS-modNV total	/16

Reference List

1. Academy of Medical Sciences. Restoring neurological function. Putting the neurosciences to work in neurorehabilitation. London: Academy of Medical Sciences; 2004.

2. Helsedirektoratet. Hjerneslag - Nasjonale retningslinjer for behandling og rehabilitering ved hjerneslag. Oslo: Helsedirektoratet; 2010.

3. Ellekjaer H, Solberg R. Hjerneslag - like mange rammes, men prognosen er bedre. Tidsskr Nor Legeforen 2007;6(127):740-3.

4. Feld JA, Rabadi MH, Blau AD, Jordan BD. Berg balance scale and outcome measures in acquired brain injury. Neurorehabil Neural Repair 2001;15(3):239-44.

5. Karthikbabu S, Nayak A, Vijayakumar K, Misri ZK, Suresh BV, Ganesan S, Joshua AM. Comparison of physio ball and plinth trunk exercises regimens on trunk control and functional balance in patients with acute stroke: a pilot randomized controlled trial. Clin Rehabil 2011.

6. Wee JY, Wong H, Palepu A. Validation of the Berg Balance Scale as a predictor of length of stay and discharge destination in stroke rehabilitation. Arch Phys Med Rehabil 2003;84(5):731-5.

7. Wee JY, Hopman WM. Stroke impairment predictors of discharge function, length of stay, and discharge destination in stroke rehabilitation. Am J Phys Med Rehabil 2005;84(8):604-12.

8. Kollen B, van de Port I, Lindeman E, Twisk J, Kwakkel G. Predicting improvement in gait after stroke: a longitudinal prospective study. Stroke 2005;36(12):2676-80.

9. Campbell GB, Matthews JT. An integrative review of factors associated with falls during post-stroke rehabilitation. J Nurs Scholarsh 2010;42(4):395-404.

10. de Oliveira R, Cacho EWA, Borges G. Post-stroke motor and functional evaluations - A clinical correlation using Fugl-Meyer assessment scale, Berg balance scale and Barthel index. Arquivos de Neuro-Psiquiatria 2006;64(3B):731-5.

11. Kibler WB, Press J, Sciascia A. The role of core stability in athletic function. Sports Med 2006;36(3):189-98.

12. Ryerson S, Byl NN, Brown DA, Wong RA, Hidler JM. Altered trunk position sense and its relation to balance functions in people post-stroke. J Neurol Phys Ther 2008;32(1):14-20.

13. Verheyden G, Nieuwboer A, Mertin J, Preger R, Kiekens C, De Weerdt W. The Trunk Impairment Scale: a new tool to measure motor impairment of the trunk after stroke. Clin Rehabil 2004;18(3):326-34.

14. Hsieh CL, Sheu CF, Hsueh IP, Wang CH. Trunk control as an early predictor of comprehensive activities of daily living function in stroke patients. Stroke 2002;33(11):2626-30.

15. Wang CH, Hsueh IP, Sheu CF, Hsieh CL. Discriminative, predictive, and evaluative properties of a trunk control measure in patients with stroke. Physical Therapy 2005;85(9):887-94.

16. Duarte E, Marco E, Muniesa JM, Belmonte R, Diaz P, Tejero M, Escalada F. Trunk control test as a functional predictor in stroke patients. J Rehabil Med 2002;34(6):267-72.

17. Verheyden G, Vereeck L, Truijen S, Troch M, Herregodts I, Lafosse C, Nieuwboer A, De Weerdt W. Trunk performance after stroke and the relationship with balance, gait and functional ability. Clin Rehabil 2006;20(5):451-8.

18. World Health Organization. International classification of functioning, disability and health, ICF. Geneva: World Health Organisation; 2001.

19. Blum L, Korner-Bitensky N. Usefulness of the Berg Balance Scale in stroke rehabilitation: A systematic review. Physical Therapy 2008;88(5):559-66.

20. Tyson SF, Connell LA. How to measure balance in clinical practice. A systematic review of the psychometrics and clinical utility of measures of balance activity for neurological conditions. Clin Rehabil 2009;23(9):824-40.

21. Verheyden G, Willems AM, Ooms L, Nieuwboer A. Validity of the trunk impairment scale as a measure of trunk performance in people with Parkinson's disease. Arch Phys Med Rehabil 2007;88(10):1304-8.

22. Verheyden G, Nuyens G, Nieuwboer A, Van Asch P, De Weerdt W. Reliability and Validity of trunk assessment for people with multiple sclerosis. Phys Ther 2006;86:66-76.

23. Verheyden G, Hughes J, Jelsma J, Nieuwboer A, De Weerdt W. Assessing Motor Impairment of the Trunk in Patients with Traumatic Brain Injury: Reliability and Validity of the Trunk Impairment Scale. SA Journal of Physiotherapy 2006;62(2):23-8.

24. Verheyden G, Nieuwboer A, De Wit L, Feys H, Schuback B, Baert I, Jenni W, Schupp W, Thijs V, De Weerdt W. Trunk performance after stroke: an eye catching predictor of functional outcome. J Neurol Neurosurg Psychiatry 2007;78(7):694-8.

25. Streiner DL, Norman ED. Item response theory. In: Streiner DL, Norman ED, editors . Health measurement scales. 4 ed. Oxford: Oxford University press; 2008. p. 299-330.

26. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of psychometric methods. NIHR Health Technology Assessment programme 2009;13(12).

27. Steinberg L, Thissen D. Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychol Methods 1996;1(1):81-97.

28. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual Life Res 2007;16 Suppl 1:19-31.

29. Cai L, Yang JS, Hansen M. Generalized full-information item bifactor analysis. Psychol Methods 2011; Advance online publication. doi: 10.1037/a0023350.

30. Ip EH. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. Br J Math Stat Psychol 2010;63(Pt 2):395-416.

31. Morizot J, Ainsworth AT, Reise SP. Toward modern psychometrics. Application of item response theory models in personality research. In: Robins RW, Fraley C, Kreuger RF, editors. Handbook of research methods in personality psychology. New York: Guilford Press; 2007. p. 407-23.

32. Embretson SE, Reise SP. Item response theory for psychologists. New Jersey: Lawrence Erlbaum; 2000.

33. Verheyden G, Kersten P. Investigating the internal validity of the Trunk Impairment Scale (TIS) using Rasch analysis: the TIS 2.0. Disabil Rehabil 2010;32(25):2127-37.

34. Wainer H, Kiely GL. Items Clusters and Computerized Adaptive Testing - A Case for Testlets. Journal of Educational Measurement 1987;24(3):185-201.

35. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25(24):3186-91.

36. Goldstein LB, Samsa GP. Reliability of the National Institutes of Health Stroke Scale. Extension to non-neurologists in the context of a clinical trial. Stroke 1997;28(2):307-10.

37. Kasner SE. Clinical interpretation and use of stroke scales. Lancet Neurol 2006;5(7):603-12.

38. Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. Stroke 2002;33(5):1261-6.

39. Benaim C, Perennou DA, Villy J, Rousseaux M, Pelissier JY. Validation of a standardized assessment of postural control in stroke patients: the Postural Assessment Scale for Stroke Patients. Stroke 1999;30(9):1862-8.

40. Salbach NM, Mayo NE, Higgins J, Ahmed S, Finch LE, Richards CL. Responsiveness and predictability of gait speed and other disability measures in acute stroke. Arch Phys Med Rehabil 2001;82(9):1204-12.

41. 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):142-8.

42. Muthen LK, Muthen BO. Mplus user guide. 6 ed. Los Angeles: Muthen & Muthen; 1998-2010.

43. Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods 2004;9(4):466-91.

44. Samejima F. Estimation of latent ability using a response pattern of graded scores. 17[Suppl], 1-98. 1969. Richmond, VA, Psychometric Society. Psychometric Monograph.

45. Brown TA. Data issues in CFA. In: Brown TA, editor. Confirmatory Factor Analysis for Applied Research. New York, London: The Guilford Press; 2006. p. 363-411.

46. Bentler PM. Comparative fit indexes in structural models. Psychol Bull 1990;107(2):238-46.

47. Steiger J, Lind J. Statistically-based tests for the number of common factors. The Annual Meeting of the Psychonomic Society, Iowa City, IA 1980.

48. Yu C-Y. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes University of California, Los Angeles; 2002.

49. Domholdt E. Statistical analysis of relationships: the basics. In: Domholdt E, editor . Rehabilitation Research. Principles and applications. 3 ed. Elsevier Saunders; 2005. p. 351-63.

50. Bland JM, Altman DG. Measurement error. BMJ 1996;313(7059):744.

51. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. Quality criteria were proposed for measurement properties of health status questionnaires. Clin Epidemiol 2007;60:34-42.

52. Altman DG. Some common problems in medical research. In: Altman DG, editor. Practical statistics for medical research. 1 ed. Chapman & Hall/CRC; 1991. p. 396-439.

53. Borghuis J, Hof AL, Lemmink KA. The importance of sensory-motor control in providing core stability: implications for measurement and training. Sports Med 2008;38(11):893-916.

54. Cholewicki J, Panjabi MM, Khachatryan A. Stabilizing function of trunk flexor-extensor muscles around a neutral spine posture. Spine (Phila Pa 1976) 1997;22(19):2207-12.

55. Hodges PW, Cresswell AG, Daggfeldt K, Thorstensson A. Three dimensional preparatory trunk motion precedes asymmetrical upper limb movement. Gait Posture 2000;11(2):92-101.

56. Horak FB. Postural orientation and equilibrium: what do we need to know about neural control of balance to prevent falls? Age Ageing 2006;35 Suppl 2:ii7-ii11.

57. Lee LJ, Coppieters MW, Hodges PW. Anticipatory postural adjustments to arm movement reveal complex control of paraspinal muscles in the thorax. J Electromyogr Kinesiol 2009;19(1):46-54.

58. Di Monaco M, Trucco M, Di Monaco R, Tappero R, Cavanna A. The relationship between initial trunk control or postural balance and inpatient rehabilitation outcome after stroke: a prospective comparative study. Clin Rehabil 2010;24(6):543-54.

59. Michaelsen SM, Dannenbaum R, Levin MF. Task-specific training with trunk restraint on arm recovery in stroke: randomized control trial. Stroke 2006;37(1):186-92.

60. Perlmutter S, Lin F, Makhsous M. Quantitative analysis of static sitting posture in chronic stroke. Gait Posture 2010;32(1):53-6.

61. Verheyden G, Vereeck L, Truijen S, Troch M, Lafosse C, Saeys W, Leenaerts E, Palinckx A, De Weerdt W. Additional exercises improve trunk performance after stroke: a pilot randomized controlled trial. Neurorehabil Neural Repair 2009;23(3):281-6.

62. Verheyden G, Nieuwboer A, De Wit L, Thijs V, Dobbelaere J, Devos H, Severijns D, Vanbeveren S, De Weerdt W. Time course of trunk, arm, leg, and functional recovery after ischemic stroke. Neurorehabilitation and Neural Repair 2008;22(2):173-9.

63. Gustafsson JE, Åberg-Bengtsson L. Unidimensionality and interpretability of psychological instruments. In: Embretson SE, editor. Measuring psychological constructs: Advances in model-based approaches.Washington DC: American Psychological Association; 2010. p. 97-121.

Tables

Table 1. Characteristics of the study samples.

Variables	Reliability study, N=50	Validity study, N=201
Gender, male/female; n(%)	31(62)/19(38)	117(58)/84(42)
Age; mean SD, min-max	51.5 SD 13.7,22-77	72 SD 14,27-98
Diagnosis; n(%)		
Stroke		
▪ Ischemic	33(66)	174(86.5)
▪ Haemorrhagic	8(16)	19(9.5)
▪ Undiagnosed		8(4)
Traumatic brain injury	6(12)	
Intracerebral tumor	3(6)	
Localisation of lesion; n(%)		
▪ Right hemisphere	26(52)	78(38.8)
▪ Left hemisphere	17(34)	76(37.8)
▪ Bilateral	7(14)	10(5)
▪ Brainstem		18(9)
▪ Cerebellum		11(5.5)
▪ MRI not performed/inconclusive		8(4)
Most affected body half, right/left/bilateral; n(%)	17(34)/27(54)/6(12)	104(52)/93(46)/4(2)
Weeks since brain lesion; mean SD,min-max	39 SD 66.2,2-359	4,7 SD 2.2,1-18*

*Only 1 patient was tested this late; 89% of the patients were tested within 7 days of the stroke.

Table 2. Overview of transformations.

TIS-NV items	TIS-modNV items	Trunk control
DSB 1,2,3*	Testlet 1	Lower trunk control
DSB 4,5,6*	Testlet 2	
DSB 7,8*	Testlet 3	Upper trunk control
DSB 9,10*	Testlet 4	
Coo 1,2**	Testlet 5	Coordination/lower trunk stability
Coo 3,4**	Testlet 6	Coordination/upper trunk stability

*DSB = Dynamic sitting balance subscale items.

**Coo = Coordination subscale items.

Table 3. Factor IRT parameter.

	Factor loadings		IRT Parameter MII*			
	MI*	MII**	Alpha	Beta1	Beta2	Beta3
Testlet 1	0.73	0.70	0.97	-1.27	-0.36	0.76
Testlet 2	0.76	0.72	1.03	-1.51	-0.86	0.43
Testlet 3	0.81	0.73	1.06	-0.22	0.81	-----
Testlet 4	0.80	0.72	1.03	-0.74	0.58	-----
Testlet 5	0.84	0.87	1.72	-1.20	0.20	0.71
Testlet 6	0.83	0.86	1.66	-0.83	0.55	0.89
Correlated error terms						
Testlet 1 with Testlet 2	------	0.36				
Testlet 3 with Testlet 4	-------	0.53				

*MI = Unidimensional IRT model.

**MII = Locally dependent unidimensional IRT model.

Table 4. Internal consistency.

	Cronbach's alpha (95%CI)	Cronbach's alpha if Item Deleted
Total sum testlets	.85 (.82 .88)	
Testlet 1		.83
Testlet 2		.83
Testlet 3		.83
Testlet 4		.83
Testlet 5		.82
Testlet 6		.82

Table 5. Intertester and test-retest reliability of each testlet by Kappa (κ) statistics.

	Intertester N=50	Test-retest N=49
Testlets	κ (% of agreement)	κ (% of agreement)
Testlet 1	.80 (86)	.66 (76)
Testlet 2	.58 (74)	.34 (61)
Testlet 3	.40 (64)	.69 (82)
Testlet 4	.51 (72)	.77 (88)
Testlet 5	.44 (76)	.66 (88)
Testlet 6	* (80)	.53 (76)

*Kappa could not be calculated.

Image files

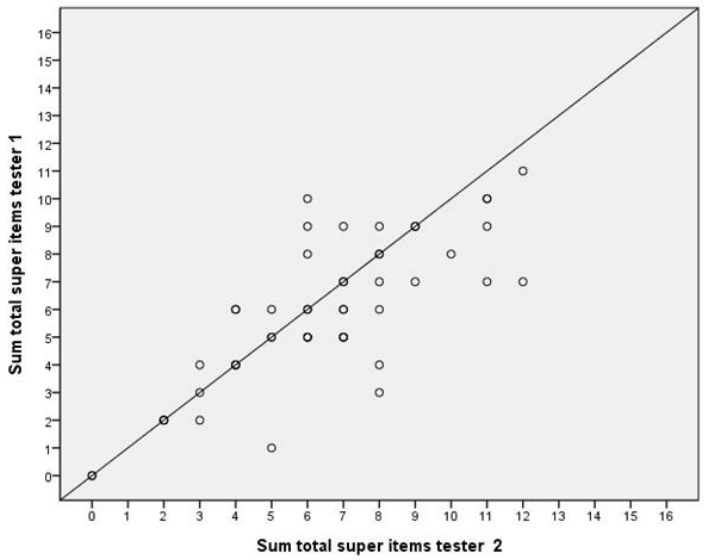

Figure 1. Graphical representation of intertester reliability data of the sum score (scale 0-16) (n=50). Maximum score is 16. 13 plots represent overlapping data for 30 patients

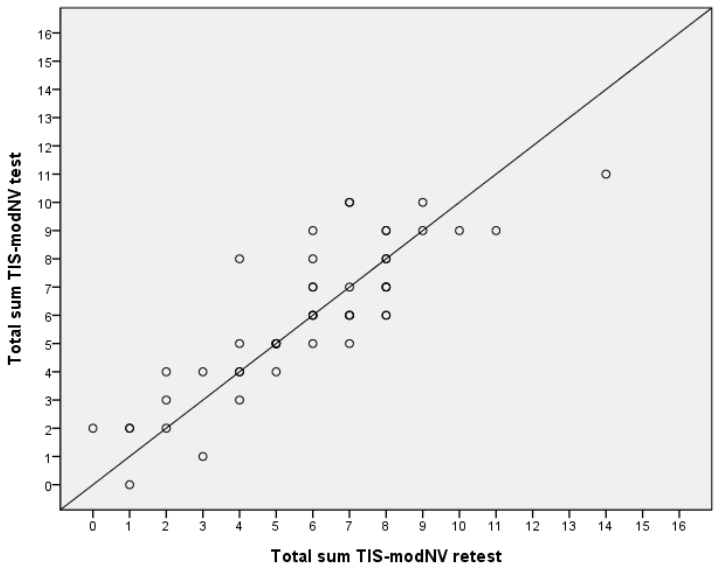

Figure 2. Graphical representation of test-retest reliability data (n=49) of the sum score (scale 0-16). 11 plots represent overlapping data for 28 patients