Research in Developmental Disabilities 85 (2019) 92–103

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Structural validity of the Bruininks-Oseretsky test of motor proficiency – Second edition brief form (BOT-2-BF)

Ted Brown

Department of Occupational Therapy, School of Primary Health and Allied Care, Faculty of Medicine, Nursing and Health Sciences, Monash University, Peninsula Campus, Build G, 4thfloor, McMahons Road, PO.O. Box 527, Frankston, Victoria, 3199, Australia

ARTICLE INFO

No. of reviews completed 2

Keywords:
Motor skills
Assessment
Rasch measurement model
Dimensionality
Children

ABSTRACT

Background: Practitioners often assess the motor skills of children presented with suspected developmental delays. It is essential that the tests used to assess children have strong measurement properties including validity.
Aim: The structural validity of the Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF) was investigated.
Method: 123 healthy children (67 males & 56 females; M = 10 years, 2 months; SD = 1 year, 4 months) completed the BOT-2-BF. The Rasch Measurement Model (RMM) was used to assess the 14-item BOT-2-BF's dimensionality, hierarchical ordering, differential item functioning (DIF), and item and person separation reliability.
Results: Nine BOT-2-BF misfit RMM requirements. A second RMM analysis of a five-item BOT-2-BF version was completed. The five-item version did meet the RMM requirements of item fit, dimensionality, hierarchical ordering, lack of DIF, and adequate item and person separation reliability.
Implications: The 14-item version of the BOT-2-BF is not recommended for use as a screening scale to assess children's motor skill performance. However, a revised five-item version of the BOT-2-BF did appear to meet RMM expectations. Further psychometric assessment of the revised five-item version of the BOT-2-BF is recommended.

What this paper adds?

Children and adolescents presenting with suspected motor coordination difficulties are frequently assessed using standardized assessment tools such as the 14 item Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF). This study investigated the structural validity of the BOT-2-BF using the Rasch Measurement Model (RMM). The findings indicated that the 14 items version of the BOT-2-BF did not appear meet the requirements of structural validity. However a resultant five-item BOT-2-BF version did meet the RMM requirements of item fit, dimensionality, hierarchical ordering, lack of DIF, and adequate item and person separation reliability. This paper adds evidence to the body of knowledge about the psychometric properties of the BOT-2-BF and its suitability for assessing children and adolescents with known or suspected motor coordination difficulties, including developmental coordination disorder (DCD), dyspraxia, and other related psychomotor problems.

1. Introduction

The International Classification of Functioning, Disability and Health (ICF) (World Health Organization, 2001) has several

E-mail address: ted.brown@monash.edu.

https://doi.org/10.1016/j.ridd.2018.11.010
Received 7 June 2018; Received in revised form 19 October 2018; Accepted 15 November 2018
Available online 28 November 2018

components to its framework including body functions and structures. Children's motor skills are included under the ICF body functions and structures 23heading. Many practitioners assess children's fine, gross, and fundamental movement skills using standardized motor skill scales (Hands, Licari, & Piek, 2015; Kennedy, Brown, & Stagnitti, 2013; Wiart & Darrah, 2001; Yoon, Scott, Hill, Levitt, & Lambert, 2006). Frequently used motor skill assessment tools include the Peabody Developmental Motor Skills – 2nd edition (PDMS-2), Movement Assessment Battery of Children – 2nd edition (MABC-2), Berry Buktenica Developmental Test of Visual Motor Integration Skills 6th Edition (Berry VMI), Wide Range Assessment of Visual Motor Abilities (WRAVMA), Test of Gross Motor Skills – Second Edition (TGMD-2), Bayley Scales of Infant and Toddler Development – Third Edition (Bayley-III) Motor Scale Kit, and the Bruininks-Oseretsky Test of Motor Proficiency – 2nd edition (BOT-2).

The BOT-2 is a standardized, norm-referenced gross and fine motor skill control assessment designed to be used by health, research, and education professionals in the assessment of those between the ages of four and 21 years presenting with suspected mild to moderate motor control difficulties (Bruininks & Bruininks, 2005). The battery was originally developed in 1978 as the Bruininks-Oseretsky Test of Motor Proficiency (BOTMP) (Bruininks, 1978) and the battery was revised between 2002 and 2005 (Bruininks & Bruininks, 2005). The BOT-2 is designed for diagnosing motor impairment, screening, assisting research objectives, determining placement, and creating and appraising motor training (Cools, De Martelaer, Samaey, & Andries, 2009). Norms were established on a national representative sample of 1520 American children, adolescents, and young adults aged four to 21 years, providing gender specific and combined gender scores, and accounting for norms of the highest incidence of disability requiring specialized education services: emotional or behavioral disturbance, specific learning disability, mental retardation, developmental delay, and speech or language impairment (Bruininks & Bruininks, 2005).

The BOT-2 has 53 items arranged into four composites that are further divided into eight subscales: *fine motor manual control composite* (FMMCC) (fine motor precision [FMP], seven items; fine motor integration [FMI], eight items), *manual coordination composite* (MCC) (manual dexterity [MD], five items; upper-limb coordination [ULC], seven items), *body coordination composite* (BCC) (bilateral coordination [BC], seven items; balance [B], nine items), and *strength and agility composite* (SAC) (running speed & agility [RSA], five items; strength [S], five items) (Bruininks & Bruininks, 2005). This four-factor model reflected in the four composite scales was utilized because it was strongly supported by confirmatory factor analysis (Bruininks & Bruininks, 2005). Items are scored on an ordinal scale with response categories diverging between four categories to 13 categories dependent on individual items (Wuang, Lin, & Su, 2009). The full version of the BOT-2 is time intensive to administer and score, requiring approximately 10 min to set up, between 40–60 min to administer the complete form, and scoring as well as completing the record form taking approximately 20 min (Deitz, Kartin, & Kopp, 2007).

A short version of the BOT-2 titled the Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF) is also available (Bruininks & Bruininks, 2005) and is composed of 14 items selected from the eight BOT-2 subscales. It is essential for motor skill assessments to have documented evidence of their validity (Brown, 2010), and to date only a few studies have been published in the peer reviewed literature about the psychometric properties of the BOT-2-BF (Brahler, Donahoe-Fillmore, Mrowzinski, Aebker, & Kreill, 2012; Carmosino et al., 2014; Fransen et al., 2014; Long, Eldridge, Harris, & Cheung, 2016; Lucas et al., 2013; Venetsanou, Kambas, Aggeloussis, Fatouros, & Taxildaris, 2009).

One specific type of validity that is critical is structural validity (also referred to as internal validity) which considers how well the items of a respective scale fit or load together to measure the trait or attribute it claims to (Mokkink et al., 2006). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) proposal specified that evidence of six types of measurement properties should be reported about health-related instruments (including motor skill assessments): internal consistency, reliability, measurement error, content validity, construct validity, and criterion validity (Mokkink et al., 2006). More specifically, the components of construct validity evidence with regard to health and epidemiological scales that should be reported according to the COSMIN are structural validity, hypothesis testing, and cross-cultural validity (Mokkink et al., 2006). Structural validity is defined as the degree to which scores of a scale are an adequate indication of the dimensionality of the construct, attribute or factor being measured (Mokkink et al., 2010; Rios & Wells, 2014).

One statistical approach that can be used to evaluate features of the structural validity of a standardized scale is the Rasch Measurement Model (RMM) (Boone, Staver, & Yale, 2014; Bond & Fox, 2015). The RMM, a type of Item Response Theory, is a mathematical model which does not assume that each item of a scale has the same value or replicates the same level of difficulty (de Ayala, 2009). Evidence of the structural validity of a scale can take the form of differential item functioning (DIF) studies and dimensionality studies (Newton & Shaw, 2016) which the RMM can provide.

The RMM analysis output generates a hierarchical scale of items (easier to hardest) identifying both *person ability* and *item difficulty* (Bond & Fox, 2015). This is based on the principle that how a person responds to an item results from the interaction between the individual and the level of item difficulty. In other words, respondents with higher levels of person ability are expected, based on the premise of the RMM, to answer a larger number of scale items correctly, whereas for items with higher levels of difficulty, fewer respondents are expected to answer these items correctly (Lim, Rodger, & Brown, 2009). Advantages of the RMM over Classical Test Theory include: (i) allowing for generalizability across respondents and items where they are placed on the same dimension and therefore become directly comparable; (ii) generating fit statistics for both items and persons; (iii) allowing for participants who respond randomly or idiosyncratically to be identified; (iv) ability to convert ordinal level data to interval level logit data; and (v) ability for data-sets to handle missing data (Ganglmair & Lawson, 2003; Törmäkangas, 2011).

The RMM provides fit statistics about which scale items fit the Rasch model expectations in order to establish the relationships between the items and the weight of these within the overall construct, and to ascertain if participants responded in a consistent and logical manner. Additionally, where an item is affirmed by fewer respondents, this suggests that a construct may be more challenging to achieve. Principles of the RMM include *dimensionality, item functioning, hierarchical ordering, DIF, item and person separation*

reliability, and *rating scale functioning* (Bond & Fox, 2015).

2. Purpose

The purpose of this study was to examine components of the structural validity of the BOT-2-BF (including the *dimensionality, hierarchical ordering, DIF, rating response scale functioning/hierarchical ordering,* and *item and person separation reliability*) using RMM methodology with a sample of typically developing school-age children aged eight to 12 years. It is hypothesized that (1) dimensionality of the BOT-2-BF will also be confirmed; (2) the BOT-2-BF's 14 items will form a hierarchical index exhibiting adequate rating response scale functioning; (3) there will be a lack of DIF of the item calibrations of the BOT-2-BF across different groups of participants based on gender (e.g., males versus females); (4) the BOT-2-BF will exhibit adequate levels of reliability; and (5) the BOT-2-BF will represent an adequate continuum or spread of the construct being measured. This in turn will provide insights about the *structural validity* of the BOT-2-BF.

3. Method

3.1. Participants

A convenient sample of 123 children were recruited for this study from Melbourne, Victoria, Australia. Inclusion criteria were that the children were between eight and 12 years of age, had a working knowledge of the English language, and had consent provided by a parent/guardian for participation in the study. Children were excluded if they presented with a known history of physical, psychosocial or intellectual impairment (based on parental report) that would impact upon their ability to complete the BOT-2-BF.

3.2. Instrumentation

The BOT-2-BF is a short version of the BOT-2 composed of 14 items proportionally selected from the eight subtests of the BOT-2 Complete Form (Bruininks & Bruininks, 2005). The BOT-2-BF is designed to be used as a screening tool and thus takes less time to administer. It typically takes a child 15 to 20 min to complete. It is suitable for use in children and adolescents aged four to 21 years and yields a maximum total raw score of 88. Evidence of BOT-2-BF's content validity was reported with a correlation of 0.80 between the full version of the BOT-2 and itself (Cools et al., 2009; Deitz et al., 2007). Inter-rater reliability was assessed on 47 children between the ages of four and 21 and was found to be > 0.90 on the BOT-2-BF (Bruininks & Bruininks, 2005). Test-retest reliability was checked for three groups of children on two separate occasions based on age (4–7 years, n = 43; 8–12 years, n = 44; and 13–21 years, n = 47) with correlation coefficients ≥0.90 for each group for the BOT-2-BF (Bruininks & Bruininks, 2005).

3.3. Data analysis

The *Statistical Package for the Social Sciences Version 22.0* (SPSS) (IBM Corp., 2013) was used for the data entry, its storage, and retrieval. Descriptive statistics such as measures of central tendency and measures of variance were calculated as appropriate to the BOT-2-BF items using SPSS (see Table 1). The RMM computer program, Winsteps (version 3.70.0) (Linacre, 2011) was used for the data analysis (Wright & Linacre, 1998). RMM analysis is an iterative process with the objective of achieving 'best fit' of the data to the model by testing the model's assumptions. The intent of the RMM analysis was to determine (1) the dimensionality of the BOT-2-BF

Table 1
Descriptive Statistics for the Bruininks-Oseretsky Test of Motor Proficiency – second edition Short Form (BOT-2-BF) Items (N = 123).

BOT-2-BF item	Mean	Std. deviation	Variance	Minimum	Maximum	Percentiles		
						25	50	75
FMP3	6.90	.332	.110	5	7	7.00	7.00	7.00
FMP6	6.42	1.014	1.029	2	7	6.00	7.00	7.00
FMI2	4.95	.227	.051	4	5	5.00	5.00	5.00
FMI7	4.47	.601	.361	3	5	4.00	5.00	5.00
MD2	6.74	1.160	1.346	4	9	6.00	7.00	8.00
BC3	2.99	0.104	.011	2	3	3.00	3.00	3.00
BC6	3.94	.385	.148	1	4	4.00	4.00	4.00
B2	3.56	.225	.065	1	4	4.00	4.00	4.00
B7	3.71	.731	.534	1	4	4.00	4.00	4.00
RSA3	8.38	.820	.672	5	10	8.00	8.00	9.00
UC1	4.91	.282	.079	4	5	5.00	5.00	5.00
UC6	6.53	1.138	1.295	2	7	7.00	7.00	7.00
S2	4.53	1.698	2.882	0	8	3.50	4.00	5.00
S3	5.17	1.340	1.796	2	9	4.00	5.00	6.00

Note: BOT-2 = Bruininks-Oseretsky Test of Motor Proficiency – second edition; FMP = Fine Motor Precision; FMI = Fine Motor Integration; MD = Manual Dexterity; BC = Bilateral Coordination; B = Balance; RSA = Running Speed & Agility; UC = Upper-Limb Coordination; S = Strength.

based on goodness-of-fit analysis (also referred to as item-fit); (2) whether DIF of the item calibration estimates occurred across participant samples in terms of gender; (3) the hierarchical order and spacing of the BOT-2-BF based on item calibration (item difficulty parameter estimate); and (4) the item- and person-separation reliability of the BOT-2-BF. The data generated by the BOT-2-BF is polytomous thus the Rasch-Masters Partial Credit Model (Masters, 1982) within the Winsteps program was used.

3.4. Rasch Measurement Model Analysis Procedures

3.4.1. Item fit

The RMM evaluates the fit of the data to an unconditional probabilistic model. The logit values represent the difficulty of the items (item weights) in an instrument and items are ordered from easiest to most difficult to provide evidence of hierarchical ordering of scale items. The fit of the items to the RMM was determined by the infit and outfit MNSQ statistic, both of which are based on a chi-square distribution (Smith, 1992). Fit statistics should range between 0.60 and 1.40 to fit RMM expectations. It has also been reported that the infit and outfit ZSTD scores, according to the RMM requirements, should fall between +2 and -2. High or low fit statistics represent abnormalities in the response pattern to the item that may be related to a lack of unidimensionality, DIF, poorly placed items in terms of developmental sequencing, or poorly worded items (Wright & Linacre, 1998). This step indicates how items fit the RMM. Item fit is also one step in the confirmation of unidimensionality.

The infit and outfit statistics use slightly different methods for assessing an item's fit to the RMM. The infit statistic gives more weight to the performance scores of participants closer to the item value. The belief is that persons whose ability is close to the item's difficulty will give a more sensitive insight into that item's performance (Bond & Fox, 2001). The outfit statistic is not weighted, and therefore is more sensitive to the influence of outlying scores. "It is for this reason that users of the Rasch model routinely pay more attention to infit scores than outfit scores. Aberrant infit scores usually cause more concern than large outfit statistics" (Bond & Fox, 2001, p. 43).

3.4.2. Confirmation of unidimensionality

Unidimensionality, which will be partially established through item fit, will be further confirmed by principal component factor analysis with orthogonal Varimax rotation of the item residuals. Factor analysis is a mathematical process that determines linear combinations of the variables in order to explain the maximum amount of variance in the data (Nunnally & Bernstein, 1994). The first factor to explain the largest amount of variance is represented by the construct identified by the RMM analysis. Hence, factor analysis of the item residuals should not identify any additional factors (e.g., minimal component of variance explained) if the assumption of unidimensionality is upheld. The criterion specified for the percentage of variance a factor must account for in order for the items in that factor to be considered to compromise a unidimensional measure was 60% (Bond & Fox, 2015; Nunnally & Bernstein, 1994). The criterion specified for the minimum factor loading an item can have and still be considered part of the underlying latent trait was 0.40 (Nunnally & Bernstein, 1994).

3.4.3. Item hierarchical ordering

The average item calibrations from the RMM analysis (referred to as logits) defined the hierarchical order of the items along the continuum. The BOT-2-BF is mapped onto an item difficulty map based on their respective logit scores. Harder items were located at one end of the linear continuum and easier items were located at the opposite end. This provides evidence of the hierarchical ordering of the BOT-2-BF items.

3.4.4. DIF

DIF, as evaluated by examining the difference between logit scores for each of the BOT-2-BF items for the two participant group variables. In this instance, two sets of logit values of the scale items based first on gender (males versus females) were generated and examined for potential significant differences using *t*-test comparisons. If any significant differences were found between the two sets of RMM logit scores, then DIF would be present (Bond & Fox, 2015).

3.4.5. Reliability

The Winsteps program generates both person reliability statistics and item reliability statistics. The item and person separation indexes were converted into strata according to the formula [4(separation index) +1]/3 (Wright & Masters, 1982). The strata were utilized to pinpoint the number of discreet groups of items (based on difficulty) and people (based on ability). It is expected that scales should separate the items and people into at least two separate groups. The item reliability coefficient should ideally be > 0.80 and the item-separation index (ISI) should be > 3.0. The person reliability coefficient should ideally be > 0.80, the person separation index (PSI) should be > 2.0, and the person raw score reliability should be > 0.80 (Arnadottir & Fisher, 2008; Wright & Masters, 1982).

3.5. Procedure

Ethical approval for this study was granted by xxxxx University Human Research Ethics Committee and the Victorian Department of Education Human Research Ethics Committee. Four state primary schools within the Melbourne metropolitan region were approached to participate in the study. The primary schools were chosen for their differing metropolitan locations to increase the diversity of the sample. Consent was obtained from principals at three of the schools. After consent was obtained, information

Table 2
Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF) Rasch Measurement Model (RMM) Item Statistics (N = 123).

BOT-2-S scale items	RMM logit item measure	Logit item measure S. E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point measure correlation
BOT2 BC3: Jumping in place – same sides synchronized[a]	**4.47**	**.14**	**.26**	**-6.4**	**.21**	**-7.1**	**.39**
BOT2 B7: Standing on one leg on a balance beam – eyes open	2.85	.17	.98	-.1	.89	-.7	.57
BOT2 BC6: Tapping feet and fingers – same sides synchronized[a]	**2.18**	**.16**	**.18**	**-7.5**	**.35**	**-8.2**	**.05**
BOT2 B2: Walking forward on a line[a]	**2.08**	**.17**	**.16**	**-8.0**	**.15**	**-8.4**	**.05**
BOT2 FMI7: Copying a star[a]	**.92**	**.15**	**.83**	**-1.0**	**.83**	**-.9**	**.16**
BOT2 S2: Knee or full push-ups[a]	**.81**	**.15**	**4.30**	**9.9**	**4.29**	**9.9**	**.78**
BOT2 UC1: Dropping and catching a ball – both hands[a]	**.15**	**.13**	**.25**	**-7.0**	**.28**	**-6.2**	**.36**
BOT2 FMI2: copying a square[a]	**.10**	**.12**	**.29**	**-6.5**	**.34**	**-5.6**	**.11**
BOT2 S3: Sit-ups	-.20	.12	1.37	2.4	1.41	2.5	.67
BOT2 FMP6: Folding paper	−1.64	.12	1.00	.0	.94	-.3	.51
BOT2 UC6: Dribbling a ball – alternating hands	−1.78	.12	.87	-.8	.75	-1.5	.74
BOT2 MD2: Transferring pennies[a]	**−2.11**	**.13**	**1.49**	**2.5**	**1.70**	**3.1**	**.61**
BOT2 FMP3: Drawing lines through paths – crooked[a]	**−2.39**	**.14**	**.42**	**-4.0**	**.38**	**-4.1**	**.23**
BOT2 RSA3: One-legged stationary hop	−5.33	.14	1.07	.5	1.11	.8	.35

Note: RMM = Rasch Measurement Model; BOT-2 = Bruininks-Oseretsky Test of Motor Proficiency – second edition; MNSQ = Mean Square; ZSTD = z-standardized; FMP = Fine Motor Precision; FMI = Fine Motor Integration; MD = Manual Dexterity; BC = Bilateral Coordination; B = Balance; RSA = Running Speed & Agility; UC = Upper-Limb Coordination; S = Strength.
The bolded values are BOT-2-BF items that misfit the Rasch Measurement Model requirements.

[a] Misfitting item according to RMM requirements of MNSQ range between 0.60–1.4, ZSDT range between -2 and 2, and/or Point Measure Correlation < 0.20.

packages and consent forms were provided to school administration staff who then randomly distributed the packages to 250 students between eight and 12 years of age. Parents were asked to return signed consent forms in a reply-paid envelope to indicate willingness for both the parent and the child to participate in the study. A total of 156 signed consent forms were returned to the researcher; however, 33/156 children were excluded due to not meeting the inclusion criteria for the study. The final sample size for the study was 123 participants.

During one session the researcher met with each child individually for approximately 30 min to complete the BOT-2-BF. Prior to each session the researcher explained the purpose of the session and sought the child's verbal consent to participate. All 123 children provided verbal consent to participate in the study. The data collection sessions were completed within school grounds at a time negotiated with the child's classroom teacher to ensure minimal impact upon the child's learning. The majority of sessions were conducted indoors within a spare classroom; however, where space was inadequate, some items (e.g. shuttle run) were completed in an undercover outdoor area.

4. Results

4.1. Participants

The sample comprised 123 children (response rate of 49.2%); 67 males (54.5%) and 56 females (45.5%). Participants varied in age from eight years to 12 years, 2 months, with a mean age of 10 years, 2 months (standard deviation [SD] = 1 year, 4 months). The participants did not have any known history of physical, psychosocial or intellectual impairment based on parental report.

4.2. Item Fit: 14-item BOT-2-BF version

RMM fit of the 14 BOT-2-BF items was assessed using fit MNSQ statistics and standardized fit statistics (ZSDT). MNSQ values outside the range of 0.6 to 1.4 were identified as potential misfitting items. The item logit scores for the 14-item BOT-2-BF measure ranged from −5.33 to 4.47 (see Table 2). Nine out of the 14 BOT-2-BF items did not meet the MNSQ criteria indicating RMM misfit: Bilateral Coordination item 3 (BC3), Bilateral Coordination item 6 (BC6), Balance item 2 (B2), Fine Motor Integration item 7 (FMI7), Strength item 2 (S2), Upper-limb Coordination item 1 (UC1), Fine Motor Integration item 2 (FMI2), Manual Dexterity item 2 (MD2), and Fine Motor Precision item 3 (FMP3). BC3, BC6, B2, UC1, and FMI2 all had Infit MNSQs < 0.40 while S2 had an Infit MNSQ > 1.40 (see Table 2). BC6, B2, and FMP3 had Point Measure Correlations < 0.20. BC3, BC6, B2, S2, UC1, FMI2, MD2, and FMP3 had Infit ZSTD scores outside the +2 to −2 range (see Table 2).

4.3. Confirmation of Unidimensionality: 14-item BOT-2-BF version

Unidimensionality was assessed by a Rasch principal components analysis (PCA) of the residuals within the Winsteps program. By completing a Rasch PCA of the residuals (referred to as the first contrast), evidence of a component that explains a large percentage of variance (usually > 60%) of the residuals and a PCA eigenvalue for the first contrast of < 3.0 are expected. For the BOT-2-BF, the percentage of variance accounted for by the first factor was 82.9% with an eigenvalue of 1.5 (see Table 3). The percentage of unexplained variance in contrasts 1–5 of the PCA of the residuals was 1.2% (with a desired percentage of < 5%) (see Table 3). The 14-item version of the BOT-2-BF met the requirements for unidimensionality even though nine of its 14 items misfit the RMM MNSQ and ZSDT expectations.

4.4. Item hierarchical ordering: 14-item BOT-2-BF version

The item hierarchical ordering of the 14 BOT-2-BF items were examined. The item logit scores for the 14 BOT-2-BF items ranged from −5.33 to 4.47 (see Table 2) with an item mean of 0.0 logits (SD = 2.43) (see Table 3). The item-separation index was 14.98 and the number of separate item strata was 20.3 (see Table 3).

The Wright Person-Item map of the BOT-2-BF items is located in Fig. 1. It visually depicts the person ability distributions of the 123 participants mapped against the logit item difficulty scores for the BOT-2-BF items. It provides a visual representation of the hierarchical ordering of the BOT-2-BF items. In Fig. 1, it appears that several of the BOT-2-BF items have similar difficulty levels. For example, B2 and BC6 have logit scores of 2.08 and 2.18 respectively, FMI7 and S2 have logit scores of 0.92 and 0.81 respectively, and MD2 and Upper-limb Coordination item 6 (UC6) have logit scores of -2.11 and -1.78 respectively. From a hierarchical ordering perspective, there may be some item redundancy in relation to item difficulty for some of the 14 BOT-2-BF items.

4.5. DIF: 14-item BOT-2-BF version

The BOT-2-BF were examined for DIF based on gender (males versus females). Only one of the items exhibited DIF based on gender, that being Balance item 7.

4.6. Person- and item-separation reliability: 14-item BOT-2-BF version

Person and item reliability indices were calculated. "Person separation reliability measures how accurately persons can be

Table 3
Fit Parameters of the Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF) with Rasch Measurement Model (RMM) Requirements (N = 123).

Parameter	RMM requirements	BOT-2 Brief Form (14 items)	BOT-2 Brief Form Revised (5 items)
Model Requirements			
• Monotonicity	Visual inspection of item thresholds; all thresholds order	✓	✓
• Local independence	r > 0.3 between the standardized residuals of the Rasch analysis between any item pairs would violate local independence	✓	✓
• Unidimensionality	PCA analysis of item residuals	✓	✓
• Differential item functioning	Percentage of items demonstrating DIF is < 5%	1/14 items exhibited DIF; BOT2 B7 exhibits DIF based on gender	✓
Model Fit: summary of items			✓
• Item hierarchy: face validity	Item hierarchical ordering conforms to theoretical/clinical expectations	✓	✓
• Item mean (SD) logits	0.0	0.00 (2.43)	0.00 (2.96)
• Item reliability[a]	> 0.8	1.00 [excellent]	1.0
• Item-separation index	> 3.0	14.98	21.03
• Item spread		9.80	9.01
• Item Model Fit MNSQ Range Extremes[a]	0.60-1.4	0.16-4.30 [poor]	0.75-1.14 [good]
• Item Model Infit ZSTD Range Extremes	−2.0 to +2.0	9.9-(-8.0)	1.0-(-2.0)
• Item Model Outfit MNSQ Range Extremes[a]	0.60-1.4	0.15-4.29 [poor]	0.52-1.29 [good]
• Item Model Outfit ZSTD Range Extremes	−2.0 to +2.0	9.9-(-8.4)	1.9-(-2.6)
• # misfitting items	0	9 (64.3%)	0 (0%) See Table 4
• BOT-2 misfitting items		See Table 2	
• # Separate Item Strata[a]	> 3	20.3 [excellent]	28.3 [excellent]
Model Fit: summary of persons			
• Person spread		3.73	6.80
• Person mean (SD) logits		0.47 (0.66)	0.43 (1.05)
• # misfitting persons	< 5%	0 (0%)	0 (0%)
Measurement Quality: reliability and targeting			
• Person reliability[a]	> 0.80 for individual person measurement	0.63 [poor]	0.60 [poor]
• Person Separation Index (PSI)	> 2.0	1.30	1.23
• # of Separate Person Strata[a]	> 3 strata for individual person measurement	2.1 [fair]	1.97 [fair]
• Person Raw Score Reliability	> 0.80 for individual person measurement	0.69	.99
• Standard Error of Measurement (SEM)	SEM as low as possible	0.40	.11
• Precision of the person measure estimates	Expected to be higher in the middle of the measurement range	6.25	5.30
• Targeting – Targeting Index[a]	When average person measure is [-1, 1] targeting is good; when [-2, 2] targeting is fair	0.47 : 0.40 [good]	0.40 : 0.35 [good]
• Ceiling effect (% of persons with maximum score)[a]	< 2-5%	0% [excellent]	0% [excellent]
• Floor effect (% of persons with minimum score)[a]	< 2-5%	0% [excellent]	0% [excellent]
• Difference between person & item means	< 1.0 logit	0.47	0.43
Unidimensionality			
• % of variance accounted for by 1st factor[a]	> 60%	80.5% [excellent]	84.3% [excellent]
• PCA (eigenvalue for 1st contrast)	≤ 2.0	6.0	1.66
• Unexplained variance in contrasts 1-5 of PCA of residuals[a]	< 5%	15.3%	15.7%
Differential Item Functioning			
• DIF by gender (Item Number [DIF contrast])	> 0.5 logits; $p < .05$	1 (7.1%)	0.0 (0.0%)
• Items exhibiting DIF by gender		BOT2 B7	none

Note: RMM = Rasch Measurement Model; # = number; SD = Standard Deviation; PCA = Principle Components Analysis; DIF = Differential Item Functioning; MNSQ = Mean Square; ZSTD = z-standardized; *Person spread* is defined as the difference between maximum person logit score and minimum person logit score. *Item spread* is defined as the difference between maximum item logit score and minimum item logit score; ✓ denotes that the RMM quality assessment guideline requirement has been met, whereas × infers that the quality assessment guideline requirement has not been met.

[a] Refers to Rating Scale Instrument Quality Criteria categories of "poor, fair, good, very good and excellent" developed by W. P. Fisher (2007).

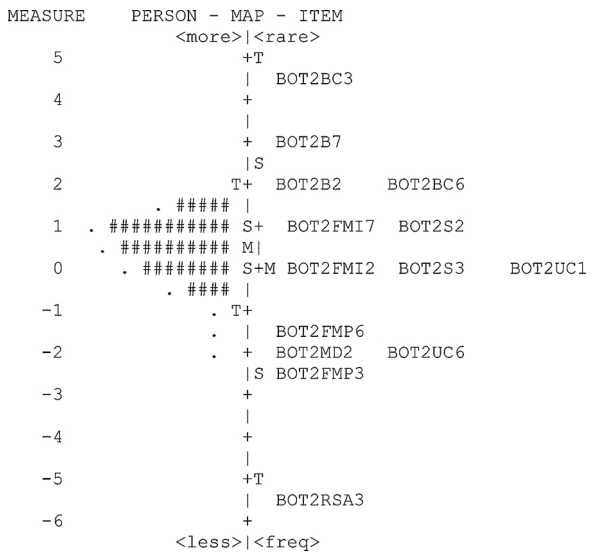

Fig. 1. BOT-2-BF 14-item Wright Person-Item Rasch Map (N = 123).
Note: Each "#" represents three participants and each "." represents one to two participants.

Table 4
Bruininks-Oseretsky Test of Motor Proficiency – second edition Brief Form (BOT-2-BF) Rasch Measurement Model (RMM) Item Statistics (N = 123).

BOT-2-S scale items[a]	RMM logit item measure	Logit item measure S. E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point measure correlation
BOT2 B7: Standing on one leg on a balance beam – eyes open	3.82	.14	.85	−.8	.84	−.7	.51
BOT2 S3: Sit-ups	1.5	.11	.76	−2.0	1.24	1.4	.75
BOT2 FMP6: Folding paper	.03	.13	1.06	.4	.97	−.1	.58
BOT2 UC6: Dribbling a ball – alternating hands	−.16	.14	.75	−1.4	.52	−2.6	.73
BOT2 RSA3: One-legged stationary hop	−5.19	.16	1.14	1.0	1.29	1.9	.50

Note: RMM = Rasch Measurement Model; BOT-2 = Bruininks-Oseretsky Test of Motor Proficiency – second edition; MNSQ = Mean Square; ZSTD = z-standardized; FMP = Fine Motor Precision; B = Balance; RSA = Running Speed & Agility; UC = Upper-Limb Coordination; S = Strength.
[a] No misfitting item according to RMM requirements of MNSQ range between 0.60–1.4, ZSDT range between -2 and 2, and/or Point Measure Correlation < 0.20.

differentiated on the measured variable, whereas item separation reliability refers to how well the test distinguishes between items along the measured variable" (Teman, 2013, p. 420). Person-separation reliability for the 14-item BOT-2-BF measure was 0.63, and item-separation reliability was 1.00 (see Table 3). The Person Raw Score reliability for the BOT-2-BF was 0.69.

4.7. Revised five-item BOT-2-BF version

The nine misfitting BOT-2-BF items were discarded. A second RMM analysis was completed with the five remaining BOT-2-BF items that exhibited model fit (see Table 4 and Fig. 2): B7, S3, FMP6, UC6, and RSA3. The item logit scores for the five-item BOT-2-BF measure ranged from −5.19 to 3.82 (see Table 4). All five BOT-2-BF items had infit and outfit MNSQ and ZSDT scores that fell within the RMM specified ranges (see Table 4). For the five-item BOT-2-BF version, the percentage of variance accounted for by the first factor was 84.3% with an eigenvalue of 1.66 (see Table 3). The percentage of unexplained variance in contrasts 1–5 of the PCA of the residuals was 15.7% (see Table 3).

The Wright Person-Item map of the five-item version of the BOT-2-BF is located in Fig. 2. It had a reasonable spread of logit item difficulty scores; however, there was the possibility for a ceiling and/or floor effect with B7 being markedly more difficult and RSA3 being notably easier than the other three items that make up the five-item BOT-2-BF version.

The five-item BOT-2-BF version was examined for DIF based on gender (males versus females) and it was found that none of the five items exhibited DIF based on gender. Person-separation reliability for the five-item BOT-2-BF version was 0.60, and item-separation reliability was 1.00 (see Table 3). The Person Raw Score reliability for the five-item BOT-2-BF version was 0.62. In summary, the five-item version of the BOT-2-BF was found to have good fit with the RMM, exhibited unidimensionality, did not demonstrate DIF based on gender, and had reasonable person- and item-separation statistics. Hierarchical ordering/rating response scale functioning was also adequate.

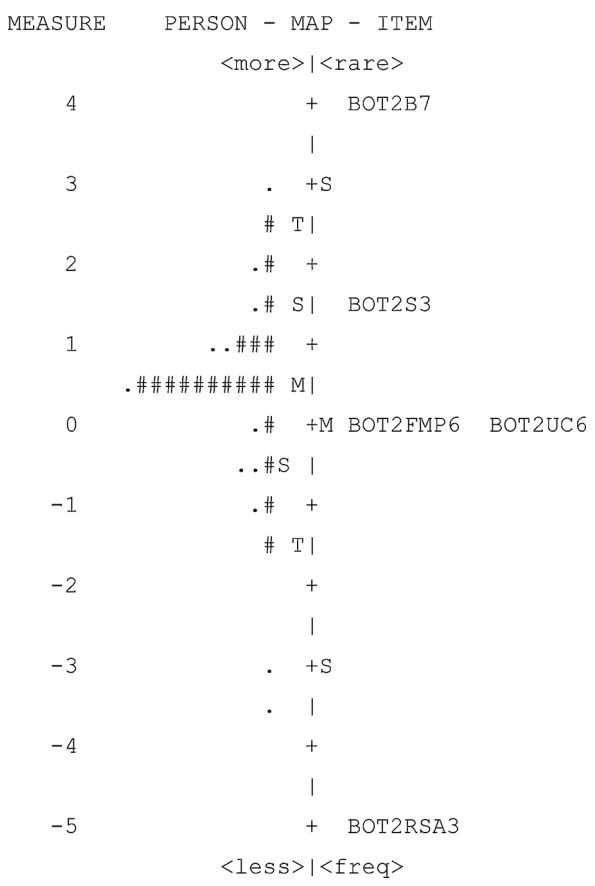

Fig. 2. BOT-2-BF five-item Wright Person-Item Rasch Map (N = 123).
Note: Each "#" represents four participants and each "." represents one to three participants.

5. Discussion

5.1. RMM item fit

Only five out of the 14 BOT-2-BF items met the RMM Infit and Outfit MNSQ requirements: Balance item 7 (B7), Strength item 3, (S3), Fine Motor Precision item 6 (FMP6), Upper-limb Coordination item 6 (UC6), and Running Speed and Agility item 3 (RSA3). Nine out of the 14 BOT-2-BF's exhibited RMM misfit: BC3, BC6, B2, FMI7, S2, UC1, FMI2, MD2, and FMP3. The results of the RMM analysis indicated that the 14 BOT-2-BF items cannot be summed together to calculate a composite score with confidence. Overall, the majority (64.3%) of the BOT-2-BF items failed to meet the RMM fit requirements. However, when the five BOT-2-BF items (e.g., B7, S3, FMP6, UC6, and RSA3) that did fit the RMM in the initial analysis, were put through a second Rasch analysis iteration, the five-item version did meet the RMM requirements.

Carmosino et al. (2014) completed a study involving 44 school-age typically developing children from the United States. They completed the items from the BOT-2 Manual Dexterity, Bilateral Coordination, Running Speed and Agility, and Upper Limb Coordination subtests. Using Pearson correlations, BOT-2 subtest item scores were correlated with their respective BOT-2 subtest total score. The strength of the correlation was used to determine if the items that made up the 14-item version of the BOT-2-BF were strongly associated with their subtest total score on the BOT-2 Complete. Results indicated that all the subtest items "in the Manual Dexterity, Running Speed and Agility, and Upper Limb Coordination subtests were significantly correlated ($p < 0.05$) with their overall subtest score" (Carmosino et al., 2014, p. 32). However, it was also discovered that two Bilateral Coordination subtest items did not significantly correlate with the overall BOT-2 Complete Bilateral Coordination subtest total score and that one of these items was included on the 14-item version of the BOT-2-BF; that item being *BC6: Tapping feet and fingers – same sides synchronized* (Carmosino et al., 2014). BC6 was an item that also was found to misfit in the first RMM analysis of the 14-item version of the BOT-2-BF in the current study. Hence the findings from the Carmosino et al. (2014) study and the current study concur with each other. It would appear that BC6 is a particularly problematic item.

Similar to Carmosino et al., Brahler et al. (2012) examined the correlations between the BOT-2 individual subtest items scores and the subtest total scores themselves involving a sample of 113 children aged 6 to 10 years old from the United States. The subtest item score and subtest total score correlations ranged from .071 to .865 (Brahler et al., 2012). The following subtest item scores on the

BOT-2-BF were found not to significantly correlate with their respective BOT-2 subtest total scores: FMI12: copying a square; FMI7: copying a star; FMP3: drawing lines through paths – crooked; and B2: walking forward on a line. FMI12, FMI7, FMP3, and B2 were also found to be misfitting items based on the Rasch analysis in the current study. It would appear that BOT-2-BF FMI12, FMI7, FMP3, and B2 are particularly problematic items.

The following items did have significant associations with their BOT-2 subtest total scores in the Brahler et al. (2012) study: FMP6: folding paper; B7: standing on one leg on a balance beam – eyes open; and S3: sit-ups. Again, similar to the current study, all these items fit the Rasch model. In summary, there was a large amount of agreement in the results between the current study and the findings published by Brahler et al. (2012) about items that make up the BOT-2-BF that are strong and those that should potentially be discarded. This provides insights into one component of the structural validity of the BOT-2-BF.

5.2. Dimensionality

Dimensionality of the 14-item BOT-2-BF version was examined via PCA of the residuals within the RMM Winsteps program. The percentage of variance accounted for by the first factor was 82.9%. It is possible that the dimensionality of the 14-item BOT-2-BF could be improved by removing the misfitting items. In this instance, the nine misfitting items were discarded and the RMM Winsteps analysis was rerun to determine if better model fit could be attained and if the unidimensionality of the five-item version of the BOT-2-BF was supported. Based on the outcomes of the RMM analysis of the five-item version of the BOT-2-BF, the first factor accounted for 84.3% of the variance with 17.1% of the raw variance explained by the persons and 67.3% of the raw variance explained by the items. This indicates that the dimensionality of the five-item version of the BOT-2-BF was supported.

There are limited studies reported in the external literature that have examined the validity of the 14-item version of the BOT-2-BF. However, McIntyre et al. (2017) investigated the convergent validity of the 14-item BOT-2-BF and the KorperKoordinationsTest fur Kinder (KTK) (a test that consists of four subscales that assess children's gross motor coordination) in a sample of 2485 Flemish typically developing children. "Moderately strong positive (r = 0.44–0.64) associations between BOT-2 total and gross motor composite scores and KTK Motor Quotient and weak positive correlations between BOT-2 Short Form fine motor composite and KTK Motor Quotient scores (r = 0.25–0.37) were found" (p. 1375). In the BOT-2 manual, evidence of the content validity of the BOT-2-BF is also recounted; a correlation of 0.80 between the full version of the BOT-2 and BOT-2-BF was reported (Deitz et al., 2007).

5.3. Hierarchical ordering

Referring to Fig. 1, the Wright Person-Item map for the 14-item BOT-2-BF version charts the person ability logit scores against the item difficulty logit scores. It provides a visual representation of how the difficulty levels of the 14-item BOT-2-BF version match the ability levels of the participant group in a hierarchical representation. For the most part, there is a reasonable spread in the difficulty levels of the 14-item BOT-2-BF version with item logit scores ranging from −5.30 to 4.47. However, there appears to be a mismatch between the person ability logit scores and the item difficulty logit scores. FMP6, FMP3, MD2, and UC6 are items that all of the participants found relatively easy, whereas BC3, B7, B2, and BC6 are all items that the participants found too hard to complete.

Therefore, eight of the BOT-2-BF items were not positioned in a useful range to indicate the ability level of the participants. In other words, there is a lack of spread of item difficulty levels in the 14-item BOT-2-BF version to cover the full range of participant ability levels. It is also interesting to note that of these eight BOT-2-BF items, five of them also generated MNSQ and ZSTD statistics outside the RMM acceptable specification ranges: BC3, BC6, B2, MD2, and FMP3. Likewise, there appears to be several 14-item BOT-2-BF version items with the same level of difficulty. For example, on Fig. 1, B2, and BC6 as well as MD2 and UC6 appear to have the same item difficulty level.

Referring to Fig. 2, the Wright Person-Item map for the five-item BOT-2-BF version also exhibits a reasonable range of item difficulties. However, there is a notable difficulty gap between the most difficult item at the top of the Wright Person-Item map and the easiest item at the bottom of the map. For example, B7 has a logit score of 3.82 and RSA3 has a logit value of −5.19. In other words, all the participants found RSA3 relatively easy and all the participants found B7 very challenging to complete correctly. It appears that the five-item BOT-2-BF version benefits from the inclusion of some additional items to fill in the difficulty gaps between the easiest and hardest items.

5.4. Differential item functioning

DIF provides information about whether or not the items of a scale are biased or act differently when completed by a specific subgroup of participants (Bond & Fox, 2015). This is particularly important when a scale or instrument is high-stakes or would have potential consequences for the test-takers. It also provides valuable information about specific items that may need to be revised or ultimately discarded. The 14-item BOT-2-BF and five-item BOT-2-BF versions were scrutinized for DIF based on gender (male versus female participants). The 14-item BOT-2-BF had one item that exhibited DIF while no items in the five-item BOT-2-BF demonstrated DIF based on gender. This is a definite strength of the five-item BOT-2-BF version.

5.5. Person and item reliability

Person reliability results ranged from 0.63 for the 14-item BOT-2-BF version to 0.60 for the five-item BOT-2-BF version which are both less than optimal. The desired level is to have person reliability scores > 0.80. The low person reliability scores can be partially

explained by the number of items per BOT-2-BF version (which ranged from a minimum of five items to a maximum of 14 items). Also, the misalignment between the person ability logit scores and the item difficulty logit scores on the two BOT-2-BF versions could also be an explanation for the low person reliability scores. It should be noted, however, that the person reliability score for the five-item BOT-3-BF version only decreased by 0.03 after the nine misfitting items were removed.

The PSI for the two BOT-2-BF versions were 1.30 and 1.23 respectively. The recommended level for the PSI is > 2.0 (Bond & Fox, 2015); therefore, the PSI for the 14- and five-item versions of the BOT-2-BF were less than optimal. Again, it should be noted that the PSI for the five-item BOT-3-BF version only decreased by 0.07 after the nine misfitting items were eliminated.

The Person Raw Score Reliability (deemed comparable to Cronbach's coefficient alpha) for the 14-item BOT-2-BF version was 0.69 and 0.62 for the five-item BOT-2-BF version. The recommended level for the Person Raw Score Reliability is > 0.80 (Boone et al., 2014). In short, the person reliability indices for the 14- and five-item versions of the BOT-2-BF were in the moderate range. The item reliability coefficients for the 14-item BOT-2-BF version was 1.00 and 1.00 for the five-item BOT-2-BF version. The desired range for the item reliability coefficient is > 0.80 (Bond & Fox, 2015).

6. Limitations

Limitations of this study include the convenience sampling approach used to recruit participants. Also, participants were recruited from one geographical region which may be a source of sampling bias. Only children who were typically developing were included in the sample and only one aspect of DIF based on gender was examined.

7. Further research

It is recommended this study be replicated with a larger sample size recruited from a larger geographical region using participants who have been randomly selected. It is also recommended that other aspects of the BOT-2-BF's validity be examined such as its relationship to other variables (convergent and divergent validity), consequences to participants of testing, and criterion validity. Further investigation of the BOT-2-BF items' DIF in relation to other participant traits (such as education, background, age, and ethnicity) is suggested. The concurrent validity of the BOT-2-BF could be examined in relation to other motor skill tests (such as the PDMS-2, MABC-2, or TGMD-2) that measure similar motor skill factors.

Given that a new five-item version of the BOT-2-BF was generated as a result of the study findings, it is recommended that other aspects of this version's validity be examined. Likewise, examination of the five-item BOT-2-BF's test-retest and inter-rater reliability is warranted.

8. Conclusion

This study investigated components of the BOT-2-BF's structural validity. The findings indicated that the dimensionality and RMM fit were not supported, nine of the BOT-2-BF items misfit the RMM expectations, and one item exhibited DIF based on gender: Balance Item 7. Hierarchical ordering and rating response scale functioning were adequate.

Given the lack of support for the structural validity of the 14-item BOT-2-BF version, a five-item version of the BOT-2-BF was trialled. Preliminary findings indicated that the five-item version of the BOT-2-BF met the RMM requirements for fit and dimensionality. None of the five items exhibited DIF based on gender, and hierarchical ordering was adequate. Item- and person-separation reliability indices were low but adequate. Further research of the five-item BOT-2-BF version is recommended.

Conflicts of interest

There are no conflicts of interest related to this article.

Funding

The study received no funding from any source.

Authors' contributions

Dr. Ted Brown designed the study, collected the data, performed the statistical analyses, completed the interpretation of the data findings and drafted the manuscript.

References

Arnadottir, G., & Fisher, A. (2008). Rasch analysis of the ADL scale of the A-ONE. *American Journal of Occupational Therapy, 62*, 51–60. https://doi.org/10.5014/ajot.62.1.51.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge/Taylor and Francis Group.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences.* Dordrecht, Netherlands: Springer.

Brahler, C., Donahoe-Fillmore, B., Mrowzinski, S., Aebker, S., & Kreill, M. (2012). Numerous test items in the complete and short forms of the BOT-2 do not contribute

substantially to motor performance assessments in typically developing children six to ten years old. *Journal of Occupational Therapy, Schools & Early Intervention, 5*(1), 73–84. https://doi.org/10.1080/19411243.2012.674746.
Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *Hong Kong Journal of Occupational Therapy, 20*(1), 30–42. https://doi.org/10.1016/S1569-1861(10)70056-5.
Bruininks, R. H. (1978). *Bruininks-Oseretsky Test of Motor Proficiency.* Circle Pines, MN: American Guidance Service.
Bruininks, R. H., & Bruininks, B. D. (2005). *Bruininks-Oseretsky Test of Motor Proficiency Second* (Edition manual). Minneapolis, MN: Pearson Assessments.
Carmosino, K., Grzeszczak, A., McMurray, K., Olivo, A., Slutz, B., Zoll, B., ... Brahler, C. J. (2014). Test items in the complete and short forms of the BOT-2 that contribute substantially to motor performance assessments in typically developing children 6-10 years of age. *Journal of Student Physical Therapy Research, 7*(2), ARTICLE 1. Retrieved from https://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1037&context=dpt_fac_pub.
Cools, W., De Martelaer, K., Samaey, C., & Andries, C. (2009). Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. *Journal of Sports Science and Medicine, 8*(2), 154–168. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761481/.
de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford.
Deitz, J. C., Kartin, D., & Kopp, K. (2007). Review of the Bruininks-Oseretsky Test of Motor Proficiency, Second Edition (BOT-2). *Physical & Occupational Therapy in Pediatrics, 27*(4), 87–102. https://doi.org/10.1080/J006v27n04_06.
Fisher, W. P., Jr (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*, 1095 Available from: http://www.rasch.org/rmt/rmt211.pdf.
Fransen, J., D'Hondt, E., Bourgois, J., Vaeyens, R., Philippaerts, R. M., & Lenoir, M. (2014). Motor competence assessment in children: Convergent and discriminant validity between the BOT-2 Short Form and KTK testing batteries. *Research in Developmental Disabilities, 35*(6), 1375–1383. https://doi.org/10.1016/j.ridd.2014.03.011.
Ganglmair, A., & Lawson, R. (2003). Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. *European Advances in Consumer Research, 6*, 162–167.
Hands, B., Licari, M., & Piek, J. (2015). A review of five tests to identify motor coordination difficulties in young adults. *Research in Developmental Disabilities, 41–42*, 40–51. https://doi.org/10.1016/j.ridd.2015.05.009.
IBM Corp (2013). *IBM SPSS statistics for windows, version 22.0.* Armonk, NY: IBM Corp.
Kennedy, J., Brown, T., & Stagnitti, K. (2013). Top-down and bottom-up approaches to motor skill assessment of children: Are child-report and parent-report perceptions predictive of children's performance-based assessment results? *Scandinavian Journal of Occupational Therapy, 20*(1), 45–53. https://doi.org/10.3109/11038128.2012.693944.
Lim, M., Rodger, S., & Brown, T. (2009). Using Rasch Analysis for establishing the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation, 16*(5), 251–260. https://doi.org/10.12968/ijtr.2009.16.5.42102.
Linacre, J. M. (2011). *Winsteps (version 3.70.0) [computer software].* Chicago, IL: Winsteps.com.
Long, S. H., Eldridge, B. J., Harris, S. R., & Cheung, M. M. H. (2016). Motor skills of 5-year-old children who underwent early cardiac surgery. *Cardiology in the Young, 26*(4), 650–657. https://doi.org/10.1017/S1047951115000797.
Lucas, B. R., Latimer, J., Doney, R., Ferreira, M. L., Adams, R., Hawkes, G., ... Elliot, J. (2013). The Bruininks-Oseretsky Test of Motor Proficiency-Short Form is reliable in children living in remote Australian Aboriginal communities. *BMC Pediatrics, 13*, 135. https://doi.org/10.1186/1471-2431-13-135.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
McIntyre, F., Parker, H., Thornton, A., Licari, M., Piek, J., Rigoli, D., ... Hands, B. (2017). Assessing motor proficiency in young adults: The Bruininks Oseretsky Test-2 Short Form and the McCarron Assessment of Neuromuscular Development. *Human Movement Science, 53*, 55–62. https://doi.org/10.1016/j.humov.2016.10.004.
Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., ... de Vet, H. C. W. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Medical Research Methodololgy, 6*(1), 2. https://doi.org/10.1186/1471-2288-6-2.
Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., ... de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.
Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*(2), 178–197. https://doi.org/10.1080/0969594X.2015.1037241.
Nunnally, J., & Bernstein, I. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.
Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260.
Smith, R. M. (1992). *Applications of Rasch measurement.* Chicago, IL: MESA Press.
Teman, E. D. (2013). A Rasch analysis of the Statistical Anxiety Rating Scale. *Journal of Applied Measurement, 14*(4), 414–434. Retrieved from https://pdfs.semanticscholar.org/a6a7/485b686ba0f57349d7885fe2958d4642f55f.pdf.
Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: an applicator's reflection. *Educational Research and Evaluation, 17*(5), 307–320. https://doi.org/10.1080/13803611.2011.630562.
Venetsanou, F., Kambas, A., Aggeloussis, N., Fatouros, I., & Taxildaris, K. (2009). Motor assessment of preschool aged children: A preliminary investigation of the validity of the Bruininks-Oseretsky Test of Motor Proficiency - Short Form. *Human Movement Science, 28*(4), 543–550. https://doi.org/10.1016/j.humov.2009.03.002.
Wiart, L., & Darrah, J. (2001). Review of four tests of gross motor development. *Developmental Medicine & Child Neurology, 43*, 279–285. https://doi.org/10.1111/j.1469-8749.2001.tb00204.x.
World Health Organization (2001). *International Classification of Functioning, Disability and Health.* Geneva, Switzerland: World Health Organization.
Wright, B. D., & Linacre, J. M. (1998). *A user's guide to WIN-STEPS Rasch-Model Computer Program.* Chicago, IL: MESA Press.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.
Wuang, Y.-P., Lin, Y.-H., & Su, C.-Y. (2009). Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency - second edition in children with intellectual disabilities. *Research in Developmental Disabilities, 30*(6), 1132–1144. https://doi.org/10.1016/j.ridd.2009.03.003.
Yoon, D. Y., Scott, K., Hill, M. N., Levitt, N. S., & Lambert, E. V. (2006). Review of three tests of motor proficiency in children. *Perceptual and Motor Skills, 102*(2), 543–551. https://doi.org/10.2466/PMS.102.2.543-551.