Original Article

doi: 10.16926/par.2026.14.05

Diagnostic Qualities of the Bruininks-Oseretsky Test of Motor Proficiency, Short Form: Comparative Study

Iva Šeflová [1ABCDE], **Josef Chudoba** [2CD], **Michael Duncan** [3CD]

[1] Department of Physical Education and Sports, Faculty of Science, Humanities and Education, Technical University of Liberec, Liberec, Czech
[2] Institute of New Technologies and Applied Informatics, Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Liberec, Czech
[3] Centre for Physical Activity, Sport and Exercise Sciences, Coventry University, Coventry, United Kingdom

Authors' Contribution: A – Study Design, B – Data Collection, C – Statistical Analysis, D – Manuscript Preparation, E – Funds Collection

Abstract: Background: The Bruininks-Oseretzky Test of Motor Proficiency, Second Edition (BOT-2), in its complete form, is considered a clinical, objective diagnostic tool that focuses on the motor proficiency of individuals with typical development and moderate motor skill difficulties. The BOT-2 Short Form (SF), an alternative tool with 19 selected test items, is designed for quick screening of motor proficiency. This study aimed to evaluate the predictive validity of the German BOT-2 SF for below-average and well-below-average motor proficiency compared to the BOT-2 CF. Methods: The motor proficiency of 637 Czech children (46.6% girls) aged 6.0-11.0 years was assessed using the BOT-2 CF. We compared differences in sensitivity, predictive validity for motor difficulties, and feasibility between the BOT-CF and SF. Results: The BOT-2 SF yielded lower total motor composite scores than the BOT-2 CF ($t(636) = 8.84$, $p < .001$). This is reflected in the low precision and predictive value of a positive test result. The BOT-2 SF sensitivity was moderate (77%), the specificity was high (94%), and the accuracy was high (93%). Conclusions: The BOT-2 SF is a suitable assessment tool for low motor proficiency when used as an alternative to the BOT-2 CF. Improving the diagnostic quality of the BOT-2 SF can be achieved by adjusting the cutoff point. Further improving the BOT-2 SF's properties as a screening tool would require revising the selection of test items to eliminate the ceiling effect.

Keywords: Motor Competence; Predictive Validity; Feasibility; Precision

Corresponding author: Iva Šeflová, iva.seflova@tul.cz

Recevied: 28.07.2025; Accepted: 18.09.2025; Published online: 7.01.2026

Citation: Šeflová I, Chudoba J, Duncan M. Diagnostic Qualities of the Bruininks-Oseretsky Test of Motor Proficiency, Short Form: Comparative Study. Phys Act Rev 2026; 14(1): 53-65. doi: 10.16926/par.2026.14.05

INTRODUCTION

Motor competence (MC) is defined as an individual's degree of proficient performance in a broad range of motor skills, as well as the underlying mechanisms, including quality of movement, motor coordination, and motor control [1]. Research has identified the role of MC in several key physical health [2] and cognitive outcomes and social-emotional health [3]. Our findings align with the prevailing global trend of low MC in children and adolescents [4]. Determining proficiency in children's actual MC is essential for the development of meaningful interventions. There are several objective assessment tools for determining MC in the scientific literature. The selection of a particular method and its precision is paramount in effectively answering research questions, with the the method of measurement and the nature of the feature under assessment, constituting other pivotal considerations [5]. The variety of purposes specific to educational, sport, and clinical settings, as well as the diversity of sociocultural conditions, complicates the establishment of a universal gold standard measure of MC. The following text is intended to provide a comprehensive overview of the subject matter.

One of the most widely used and accepted tools for assessing MC, due to its high reliability and validity, is the Bruininks-Oseretsky Test of Motor Proficiency, 2nd Ed. (BOT-2). It is designed to assess the motor proficiency of all children, from those typically developing to those with mild to moderate MC problems [6]. The BOT-2 exists in a US original complete form (BOT-2 CF) with 52 test items and normative criteria for the US population ages 4-21. Based on the BOT-2 CF, a selection of 14 test tasks was created in a short form (BOT-2 SF). In terms of MC assessment, the BOT-2 theoretically offers a unique opportunity for both screening with the BOT-2 SF and the potential for subsequent diagnosis and specification of groups at risk of developmental delay. From a methodological perspective, the BOT-2 CF can be considered a diagnostic tool. Diagnostic tests are generally considered to provide definitive information about the presence of a disease or target condition [7]. In our study, the BOT-2 CF identifies subjects with below average MC as positive and others as negative.

A standard diagnostic test may be too resource-consuming, costly, or invasive to be practical for widespread use [8]. The BOT-2 CF demonstrates low feasibility compared to other MC tools [9,10]. Such scenarios require an alternative test that can better balance diagnostic yield with pragmatism [8]. Our study considers the BOT-2 SF as an alternative tool. Such alternative screening tests are known to be diagnostically imperfect and sometimes ambiguous [7]. Consequently, it is crucial to determine how these tests can identify the likely presence or absence of the condition of interest, i.e., their predictive validity. In the case of BOT-2 SF, the original American version, there is discussion about revising 14 selected items to improve psychometric quality [11,12]. Once the predictive validity of screening tools is determined, evaluating such tests in relation to practical utility and feasibility is important, which assesses aspects such as the test's availability, the measurement's feasibility in practice, and the test subject's acceptability [13].

Our study aimed to analyse the predictive validity of the German BOT-2 SF, which has 19 test items for motor proficiency below and well below average. In relation to predictive validity, we evaluated and compared the practical usability and feasibility of both test instruments. We defined three null hypotheses:

- HA_0: The BOT-2 SF does not provide comparable predictive validity as the BOT-2 CF.
- HB_0: Shifting the cut-off points will not address any differences in the psychometric quality of the BOT-2 SF.
- HC_0: The BOT-2 SF does not show significantly better feasibility than the BOT-2 CF.

We contribute to the issue with a comparative study based on measuring Czech school children aged 6-11 years with the BOT-2 CF adapted for German-speaking countries. This version may be more appropriate for the European context [14]. To our

knowledge, the BOT-2 SF and CF have not yet been compared to the German version of the test battery.

MATERIAL AND METHODS

Participants and Study Settings

This study was carried out in the Czech Republic between 2020 and 2022. Testing that was temporally affected by the SARS-CoV-2 pandemic was always carried out in compliance with current hygiene regulations. It took place in school premises during periods when the wearing of protective masks was not mandated in the classroom. Data were collected from 637 children aged 6.0-11.0 years (297 girls, 46.6 %). Age groups were evenly distributed, with age M = 8.42±1.30 years.

Participants were selected by quota sampling to ensure the representation of different socio-cultural groups in the population. In our selection process, we accounted for quotas based on the characteristics of age, gender, school size as expressed by the number of students, urban and rural location, associated estimated ethnicity and minority representation, and related socioeconomic status (lower in small schools outside larger cities). The study involved pupils from six primary schools in three regions. Sports schools and schools for students with special educational needs were not included in the study.

The initial requirement for the total sample size was a minimum of 40 children in each normative age group of a given gender. This number takes into account the condition of the variability of the data (sample > 30 when assessing relative frequency), the required width of the confidence interval (precision of the estimate in the Total motor composite and subcategories), and the estimate of the effect size under investigation (differences due to age and gender) [15]. Thirty-six trained researchers collected data during physical education classes. Their reliability was verified in pre-tests [16].

Methods – Motor Proficiency

Motor proficiency was determined using the complete form of the BOT-2, German version. Validity and reliability for measuring motor proficiency in the fine and gross motor categories were tested in a German standardisation study with n = 1177 German, Austrian and Swiss children [17]. The BOT-2 CF contains 53 test items and assesses Total Motor Composite (TMC) and level of fine and gross motor skills in 4 motor area composites 1 - 4, with 8 subtests I – VIII:

1. Fine Manual Control (I. Fine Motor Precision, II. Fine Motor Integration).
2. Manual Coordination (III. Manual Dexterity, VII. Upper-limb Coordination).
3. Body Coordination (IV. Bilateral Coordination, V. Balance).
4. Strength and Agility (VI. Running Speed and Agility, VIII. Strength).

The German version of the BOT-2 SF contains 19 selected test items, of which 11 focus on fine motor skills (7 pencil and paper items in categories I. and II., two items each in categories III. A VII.) and 8 tasks on gross motor skills (2 selected tasks each for categories IV., V., VI. and VIII.). The BOT-2 SF only allows the evaluation of the TMC with respect to age and gender, not the individual fine and gross motor skill subcategories. After calculating TMC values (evaluated as standard score with M = 50 and SD = 10), we categorized the results into performance categories: 'well-above average' (standard score of 70 and above), 'above average' (60 to 69), 'average' (41 to 59), 'below average' (31 to 40), and 'well-below average' (30 and below). In the present study, children in the 'below average' category were classified as 'at risk' for developmental delay in motor skills [18], while those in the 'well-below average' category were classified as 'at risk' for developmental coordination disorder (DCD) [19].

Table 1 Assessed feasibility categories

Assessed feasibility categories	Good (1)	Average (2)	Poor (3)
Administration Time	Less than 10 minutes.	10–20 minutes.	More than 20 minutes.
Equipment	Equipment available in schools and homes.	Equipment that could be exchanged for more easily accessible equipment.	Equipment that schools were unlikely to possess, or a test kit, incurs purchase costs.
Space	Less than 6 m space required.	6-10 m.	More than 10 meters, requiring an outdoor space, gym or large open classroom.
Assessment type	Product only.	Process and product.	Process only.
Items	Less than 6 items.	6-12 items.	More than 12 items.
Evaluator training	Training time less than half a day.	Half a day to one and a half days.	Training more than 1.5 days.
Qualifications required	Able to be delivered by any qualified staff.	Requiring school teacher level qualifications.	Requires higher than school staff qualifications.
Commonness of tasks*	Usual tasks included in the curriculum.	One unusual item.	More than one unusual item.
Cost-effectiveness of equipment	No cost.	One-time purchase of equipment	Need for ongoing costs, e.g. purchase of individual evaluation protocols or time-limited licenses of evaluation software.
Time required to complete all test items	Less than 45 minutes class period	90-45 minutes.	More than 90 minutes.

*Commonness of tasks in terms of their inclusion in curriculum documents in the context of the cultural environment.

Methods - feasibility

To assess feasibility, we used and adapted the categories proposed by Klingberg et al. [20]. In Table 1, we added three additional attributes to the Klingberg et al. [20] proposed attributes with ratings in three categories (1) good, (2) average, and (3) poor.

Methods - Comparative Study

To compare the BOT-2 SF and the BOT-2 CF, we methodologically followed the steps of comparative analysis [21]:

1. Specification of the object of comparison: the selected psychometric and descriptive characteristics of the BOT-2 SF and the BOT-2 CF.
2. Definition of the compared characteristics, traits, contextual variables and comparability assessment. We assessed changes in TMC variables using the BOT-2 CF and SF as two tools. The other subcategories assessed in the BOT-2 CF, such as Fine Manual Control, Manual Coordination, Body Coordination, Strength, and Agility, could not be determined using the standard BOT-2 SF manual. We compared the predictive validity of the BOT- 2 CF and SF for capturing 1) at risk for developmental delay in the TMC below average (TMC≤40) category and 2) at risk for developmental coordination

disorder (DCD) in the well- below average (TMC≤30) category. To assess and compare the feasibility of the BOT-2 CF and SF, we used adapted categories proposed by Klingberg et al. [20] (Table 1).

3. The determination of specific comparison techniques was described in the Statistical Analysis section.
4. The method of evaluating the obtained information and the systematics of the outputs was described and interpreted by changes in selected psychometric characteristics. For descriptive characteristics, we assessed and interpreted changes in descriptive feasibility categories.

Statistical Analysis

Lilliefors test was used to test the normality of the data. To evaluate the agreement of the means of the two samples (gender differences), we used a two-sample t test for the normality of the data.When data normality was not met, we used the Mann–Whitney test. We used a one-factor analysis of variance (ANOVA) when the data were normal to assess the agreement of the means of more than two samples (age differences). When the data normality condition was unmet, we used the Kruskal–Wallis test. Two-factor ANOVA was used to compare the agreement between two samples (gender differences), where the effect of age was expected.

Hypothesis tests were conducted at the 5% significance level. For ANOVA, Kruskal–Wallis test, and two-factor ANOVA, post hoc analyses of intercomparisons were performed if the H0 hypothesis of agreement of means/medians was rejected.

The effect size was ascertained for the difference in means using Cohen's d for two samples or Hedges g for more than two samples. Fisher eta was used to assess the variance η. Cohen's d classified effect sizes as small (d = 0.2), medium (d = 0.5), and large (d ≥ 0.8)

Hypothesis tests were conducted at the 5% significance level. For ANOVA, Kruskal–Wallis test, and two-factor ANOVA, post hoc analyses of intercomparisons were performed if the H0 hypothesis of agreement of means/medians was rejected.

The effect size was ascertained for the difference in means using Cohen's d for two samples or Hedges g for more than two samples. Fisher eta was used to assess the variance η. Cohen's d classified effect sizes as small (d = 0.2), medium (d = 0.5), and large (d ≥ 0.8) [22].

To determine predictive validity, we used the following variables: sensitivity, specificity, accuracy, positive and negative predictive values, positive and negative likelihood ratios, the Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC).

We assessed sensitivity as the probability of a positive outcome conditional on the individual being truly positive. The sensitivity of a test was defined as the proportion of people with disease who will have a positive result [23]. Specificity expresses the probability that the test will be negative in a healthy person. Test accuracy assessed the overall agreement of the test with reality. The predictive value of a positive (negative) test expresses the probability that a person is actually positive (negative) on a positive (negative) test [24].

A positive likelihood ratio (LR+) was assessed as the probability that a positive test would be expected in a person divided by the probability that a positive test would be expected in a person without difficulties. A negative likelihood ratio (LR-) was assessed as the probability of a person testing negative who has a disease divided by the probability of a person testing negative who does not have difficulties [25]. Findings with LRs greater than 1 argue for the diagnosis of interest; the bigger the number, the more convincingly the finding suggests that difficulties. Findings whose LRs lie between 0 and 1 argue against the diagnosis of interest; the closer the LR is to 0, the less likely the difficulties. Findings whose LRs equal 1 lack diagnostic value [26].

We set the thresholds for sensitivity and specificity measures according to the following criteria. The first is the sufficiency of the sample size, which increases the quality and precision of the estimates made. The next one is the representativeness of the sample related to the frequency of occurrence of the phenomenon under study in the population [27]. Another criterion is the severity of the pathology [28]. Considering the criteria, we evaluate a threshold of 80% as sufficient for our purposes.

Using ROC analysis, we validated the cut-off point for a diagnostically acceptable solution to the probability of false negative and false positive DCD results using the BOT-2 SF. The performance of a diagnostic variable was quantified by calculating the Area Under the Curve (AUC), which is a standard expression of a test's diagnostic performance [29]. Statistical data processing was performed using Matlab software (The MathWorks, UK).

Ethics Statement

This research study was reviewed and approved by the Ethics Commission of the Technical University of Liberec, Czech Republic. on 23. 10. 2019. The procedures involved in the study were undertaken in accordance with the ethical standards of the responsible Czech National Committee on Human Experimentation and the Helsinki Declaration of 2000. The participants' legal guardians provided written informed consent to participate in this study and for anonymised data collection.

RESULTS

Our cross-sectional study [30] published complete results of BOT-2 CF measurements, including the effect of age and gender on Total motor composite, 4 categories of gross and fine motor skills, and each of the 7 subcategories. Here, we present relevant data on the differences between BOT-2 CF and SF.

Precision

We evaluated the accuracy of BOT-2 SF based on differences in TMC, as BOT-2 SF does not allow for assessing fine and gross motor skills subcategories. TMC determined using the BOT-2 CF is average for our group. The TMC of the BOT-2 SF is also in the average category but shows lower mean values. The difference in TMC BOT-2 CF and BOT-2 SF is statistically significant ($t(636) = 8.84$, $p<0.001$) (Table 2). Figure 1 shows a histogram of the relative frequency of differences in TMC BOT 2 CF and BOT 2 SF results.

Table 2. Total motor composite results BOT-2 CF and BOT-2 SF total and by age

Indicator	Age categories [years]					Total
	6.0 to 6.9 M (SD)	7.0 to 7.9 M (SD)	8.0 to 8.9 M (SD)	9.0 to 9.9 M (SD)	10.0 to 10.9 M (SD)	
	n = 97	n = 185	n = 129	n = 125	n = 101	n = 637
TMC BOT-2 CF	50.27 (10.44)	48.01 (9.70)	46.97 (9.76)	43.16 (10.77)	42.82 (10.30)	46.37 (10.49)
TMC BOT-2 SF	48.05 (11.56)	45.27 (10.98)	44.07 (10.09)	41.82 (10.06)	41.01 (10.57)	44.44 (10.28)
Mean difference	1.72 (4.89)*	2.49 (5.37)*	2.56 (5.40)*	1.01 (6.09)	1.41 (5.52)*	1.93 (5.50)*
t-value	t(96) = 3.46	t(184) = 6.31	t(128) = 5.38	t(124) = 1.85	t(100) = 2.57	t(636) = 8.84
p-value	<0.001	<0.001	<0.001	0.066	0.012	<0.001

Note: M = mean, SD = standard deviation, TMC = Total Motor Composite.

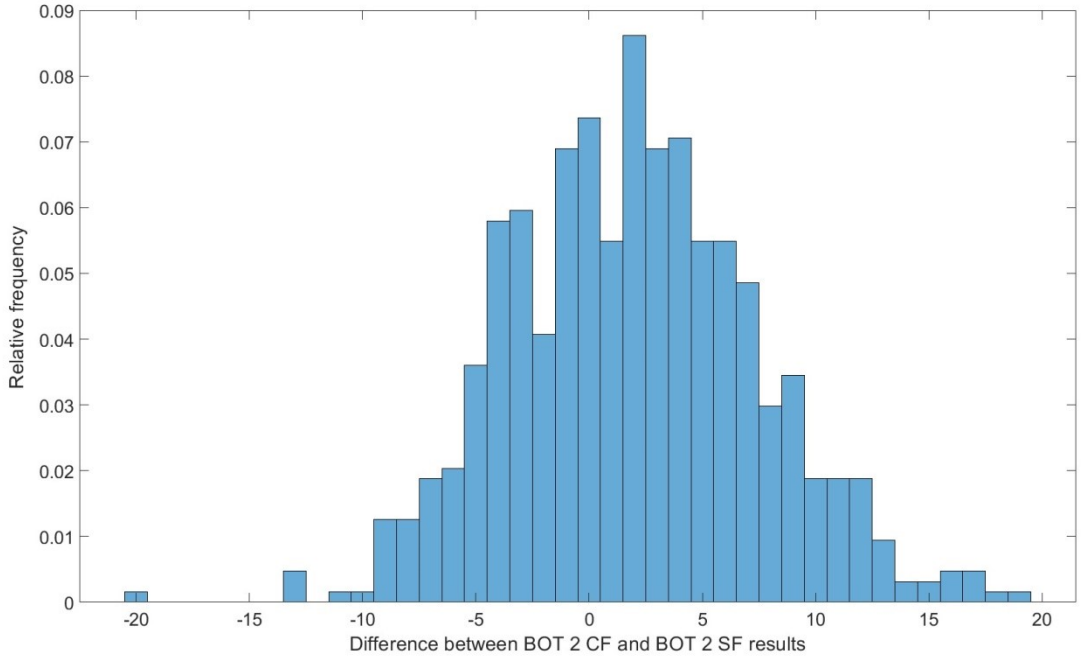

Figure 1. Histogram of the relative frequency of differences between the Total motor composite BOT-2 CF and BOT-2 SF

The distribution of TMC in performance categories BOT-2 CF and BOT-2 SF was: well-above average (CF n=2, 0.31% and SF n=1, 0.1%), above average (CF n=68, 10.68% and SF n= 40, 6.28%), average (CF n=380, 59.65% and SF n=389, 61.10%), below average (CF n=144, 22.61% and SF n=140, 21.98%), well-below average (CF n=43, 6.75% and SF n=67, 10.51%).

Age and gender differences

We analysed the effect of age on BOT-2 CF results. At the 5% significance level, the null hypothesis H0 was tested using ANOVA (mean TMC is the same for all age categories). We reject H0 ($F(4.632)=11.14$, $p < 0.01$, Hedges' $g = 0.73$). Post hoc analysis revealed that the TMC value decreased significantly when comparing the older age groups of 10 and 9 years with the younger age groups of 8, 7 and 6 years ($p<0.001$). The older age groups of 10 and 9 years scored significantly lower on the TMC than the younger age groups of 8, 7 and 6 years ($p<.001$). We observed a consistent decrease for BOT-2 SF (ANOVA, $F(4,632) = 8.48$, $p <.001$). The effect of age and gender on BOT-2 CF performance was examined using a two-factor ANOVA ($p=0.05$). We found identical mean TMC values for girls and boys of all age categories ($t(635)=1.77$, $p=0.08$, Cohen's $d=0.15$). We do not observe the same trend for TMC BOT-2 SF, where TMC values for girls and boys are significantly different ($t(635) = 2.81$, $p = 0.005$). Girls achieve better values.

Predictive Validity for Motor Skills Difficulties

To determine predictive validity, we used the basic indicators of predictive test validity: sensitivity, specificity, likelihood ratio and predictive values for below average TMC scores (TMC≤40) and for below average TMC scores (TMC≤30) (Table 3). In Figure 2, the ROC curve graphically depicts the sensitivity and specificity of TMC results. The area under the curve, AUC, is 0.92. According to the breakdown used in the literature, a test with an AUC above 0.75 can be considered satisfactorily discriminating, and above 0.90 can be considered excellently discriminating [13].

New Cut-off Point Proposal

Table 3 contains an expression of the predictive validity values when the cut-off point for the diagnosis of DCD was set at 32 points. For setting the optimal cut-off point value, we used the method that defines the cut-off point as the value whose sensitivity and specificity are the closest to the value of the area under the ROC curve, and the absolute value of the difference between the sensitivity and specificity values is minimum [31].

Table 3. Predictive validity of the BOT-2 SF for below average (well-below average) *(new cut-off point)* Total Motor Composite

Variable		Estimated value	Confidence interval 95 %	
			Lower Limit	Upper Limit
Sensitivity		0.77 (0.77) *(0.79)*	0.70 (0.62) *(0.67)*	0.82 (0.87) *(0.87)*
Specificity		0.86 (0.94) *(0.95)*	0.82 (0.92) *(0.93)*	0.82 (0.92) *(0.93)*
For any particular test result, the probability that it will be:	Positive	0.32 (0.11) *(0.12)*	0.29 (0.08) *(0.10)*	0.36 (0.13) *(0.15)*
	Negative	0.68 (0.89) *(0.88)*	0.64 (0.87) *(0.85)*	0.71 (0.92) *(0.90)*
For any particular positive test result, the probability that it is:	True positive*	0.70 (0.49) *(0.63)*	0.63 (0.38) *(0.52)*	0.75 (0.61) *(0.73)*
	False Positive	0.30(0.51) *(0.37)*	0.25 (0.39) *(0.27)*	0.37 (0.62) *(0.48)*
For any particular negative test result, the probability that it is:	True Negative#	0.90 (0.98) *(0.98)*	0.87 (0.97) *(0.96)*	0.92 (0.99) *(0.99)*
	False Negative	0.10 (0.02) *(0.02)*	0.08 (0.01) *(0.01)*	0.13 (0.03) *(0.04)*
Likelihood Ratios:	Positive conventional LR+	5.50 (13.41) *(15.67)*	4.87 (11.16) *(13.46)*	6.21 (16.10) *(18.24)*
	Negative conventional LR-	0.27 (0.25) *(0.22)*	0.23 (0.18) *(0.17)*	0.31(0.33) *(0.28)*

* Positive predictive value, # Negative Predictive Value

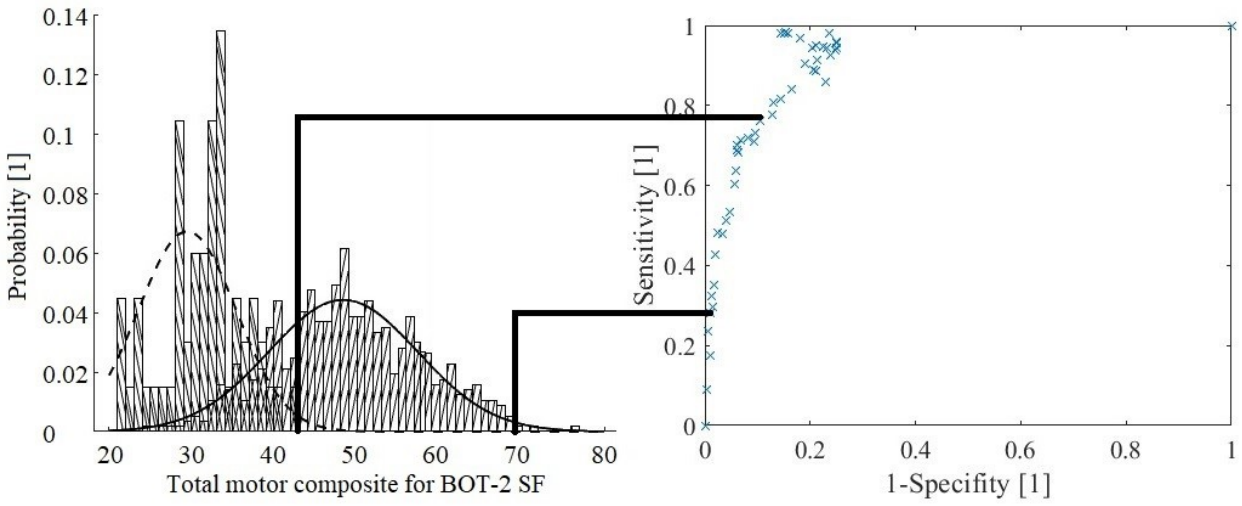

Figure 2. Receiver Operating Characteristic curve for the BOT-2 SF

Table 4. BOT-2 CF and BOT-2 SF Feasibility

Category	BOT-2 CF	BOT-2 SF
Administration Time*	(3) More than 20 minutes.	(2) 10 - 20 minutes.
Equipment	(3) Test set with purchase cost.	(3) Test set with purchase cost.
Space*	(3) More than 10 meters, gym.	(2) Ordinary room 6 - 10 m.
Assessment type	(1) Product.	(1) Product.
Items	(3) 53 test items > 12 items	(3) 19 test items > 12 items
Evaluator training	(2) Half a day to one and a half days.	(2) Half a day to one and a half days.
Qualifications required	(2) Able to be delivered by any qualified staff.	(2) Able to be delivered by any qualified staff.
Commonness of tasks	(1) Tasks usual for Europe, in the curriculum.	(1) Tasks usual for Europe, in the curriculum.
Cost-effectiveness of equipment	(3) Cost of evaluation kit and license.	(3) Cost of evaluation kit and license.
Time required to complete test items**	(2) 90 - 45 minutes.	(1) Less than 45 minutes.
Point score total	**(23) points**	**(20) points**

*Change from poor to average, **Change from poor to good

Feasibility

In the feasibility point evaluation, BOT-2 CF received a total of 23 points, and BOT-2 SF received 20 points. Changes in the feasibility of both tools occurred in 4 of the 10 attributes assessed (Table 4).

DISCUSSION

The present study compared and analysed the differences in selected psychometric and descriptive characteristics of the German version of the BOT-2 CF and BOT-2 SF. To the best of our knowledge, this was the first study evaluating the German versions of the BOT-2 in this way. The results of this study are important as providing evidence of feasibility and diagnostic qualities of the BOT-2 SF and CF enables researchers and practitioners to make evidence-based decisions regarding which tool might be most appropriate to use and in what circumstances. Consequently, the current study extends the existing literature on the topic of MC in children and adds new insights into assessment approaches for MC monitoring during childhood. These findings are in line with broader research underlining the role of physical activity for both motor and psychosocial outcomes, including its links with anxiety and sleep quality [32] and its importance for injury prevention in youth sport [33].

Precision

When TMC is assessed by the BOT-2 SF, the overall proportion of children categorised in the below average category of TMC drops from 70.6 % to 67.5%. Similarly, Jírovec et al. [34] found lower TMC in Czech school children using the original US version of the BOT-2 SF than using BOT- 2 CF.

The distribution of the TMC variable shows that significantly fewer individuals were rated as well-above average and above average when using BOT-2 SF instead of BOT-2 CF. Likewise significantly more children were in the well-below average category.

In general, a screening test such as the BOT-2 SF would be expected to have a normal distribution of TMCs across all performance categories and sensitivity at both ends of the distribution. We do not find a normal distribution in the BOT-2 CF, where there is a ceiling effect in 26 out of 53 test items (the criterion was a coefficient of variation less than 0.25). The same phenomenon is repeated in selected test items of the BOT-2 SF in the German version of the test. Our findings indicate that the BOT- 2 CF and SF category at the high end of the TMC spectrum may not be sufficiently discriminatory for the Czech sample [30]. Brahler et al. [11] and Jírovec et al. [32] observed a similar phenomenon in the original US test battery. The BOT-2 SF might be a useful tool to reveal mainly delayed but not above-average (advanced) psychomotorically developed children.

The ceiling effect of test tasks should be considered as one of the important criteria when selecting test tasks for the BOT-2 SF. In agreement with Carmosino's et al. [12] proposal for the revision of the US selection, a revision of the BOT-2 SF items for the German selection would be worth considering as well.

Predictive Validity for Motor Skills Difficulties

In the TMC well-below average category, the BOT-2 SF shows intermediate sensitivity, high specificity, and high accuracy. However, there is a low predictive value of a positive test reflecting the likelihood that a child has a truly very below average motor score on a positive test. In the category of TMC below average (TMC≤40), where children at risk of developmental delay are found, the BOT-2 SF shows a concordant mean sensitivity value, and still high specificity and accuracy values. The positive predictive value is intermediate.

When compared with the results of Korean school children [35], where the sensitivity for below average TMC is 83% and specificity is 92%, the values for our Czech sample are lower. Using the original US BOT-2 SF with a different selection of test items, Jírovec et al. [34] found a similar sensitivity of 83%, but an even significantly lower specificity of 42.9%. These discrepancies may be the result of using a different version of BOT-2 with different normative criteria. They may also be influenced by our use of a larger dataset.

We recorded the highest LR+ values for the BOT-2 SF test for well-below average TMC, indicating high confidence in determining below average results. The diagnostic power of the BOT-2 SF, assessed by the AUC variable expressing the area under the ROC curve, is considered excellentin terms of discriminative ability. Summarizing the above results, we reject the nul hypothesis HA_0 and accept alternative hypothesis HA_1: The BOT-2 SF provides comparable diagnostic qualities for low MC and for developmental delay and risk of DCD as the BOT-2 CF. We evaluated a new cut-off point of 32 points as the best diagnostic point, which offers the maximum sum of sensitivity and specificity values and a balance between the probability of false positive and false negative conclusions. The cut-off represents a more accurate criterion for assessing the risk of DCD using the BOT-2 SF test. However, it does not solve the problem of selecting items with a ceiling effect in the BOT-2 SF.

Although the BOT-2 SF showed very good diagnostic values, shifting the cut-off did not address the low predictive value of a positive test. Furthermore, we expected the BOT-2 SF as a screening tool to have good discriminatory properties at both ends of the distribution. However, the BOT-2 SF lacked this due to the ceiling effect of the test items. We accept the hypothesis HB_0: Shifting the cut-off points will not address any differences in the psychometric quality of the BOT-2 SF.

Feasibility Analysis

Reducing the number of items in the SF from the original 53 to 19 alternatives significantly reduces the total time for performing test items, time for administration and evaluation, and space requirements for implementation. This improves the ease of measurement. Test availability, including cost-effectiveness pricing, remained unchanged.

In terms of the evaluation method, the BOT-2 SF allows the evaluation of TMC by age and gender, which takes up to 20 minutes. In contrast, the BOT-2 CF evaluation, which allows for the assessment and comparison of gross and fine motor skills in addition to TMC, 4 motor area composites with 8 subtests I. - VIII., is very time-consuming, with the basic evaluation often taking over an hour and the more detailed one over 90 minutes. The German version does not offer an automatic evaluation program, so the individual items need to be looked up in the manual by using the BOT-2 CF. This requires a minimum of 79 manual spreadsheet conversions for the basic evaluation of an individual (more detailed evaluations e.g., confidence interval, percentile rank, and age equivalent would be even more time-consuming). The evaluation is equally challenging in terms of errors of inattention.

Overall, these arguments reduced the feasibility of the BOT-2 CF compared to the BOT-2 SF. As the BOT-2 ranked among the more demanding and less practical tools for screening investigation [20], we accept the hypothesis HC_0 that the BOT-2 SF does not show significantly better feasibility for screening investigation than the BOT-2 CF.

Based on the psychometric and descriptive characteristics of the BOT-2 SF, the current research study contributes to the discussion regarding the use of the BOT-2 SF for specific purposes. The BOT-2 SF appears to be a suitable tool for use in the school environment to assess an individual's level of motor skills. The BOT-2 SF can be used as part of the process of determining eligibility for special educational services and occupational therapy or physical therapy needs, and for screening for motor delays. This is in line with the recommendations for the BOT-2 SF in the original American version [36]. It can also be used for designing and evaluating education programs with emphasis to low motor competence.

Clinically the BOT-2 is useful for assessing people with suspected motor sill difficulties, and in injury prevention and rehabilitation with the SF serving as an efficient screening tool and the CF providing a more comprehensive, detailed assessment.

The strengths of this study included (i) the large sample of children and (ii) the use of the full German version of BOT-2. Limitations included (i) the limited number of studies using the German version of BOT-2 for results comparsion, (ii) possible bias of testing results by implementation during the SARS-CoV-2 pandemic.

CONCLUSION

The BOT-2, in short form, is an alternative test to the BOT-2, in complete form, which is a suitable assessment tool for motor difficulties, including the risk for developmental delay and risk of DCD. Overall, the BOT-2 SF provides lower Total Motor Composite values than the BOT-2 CF and classifies significantly more children in the well-below average category. This is reflected in the low predictive value of a positive test. Moving the cut-off point to 32 points increases the selected variables of diagnostic qualities for low motor competence. However, even with the new cut-off point, the BOT-2 SF test's weaknesses as a screening tool will remain the imbalance of the Total Motor Composite at both ends of the distribution and the low detection of above-average results. The selection of test items for the BOT-2 SF as a screening tool requires revision to exclude items that reach the ceiling effect.

Based on the evaluation of changes in the descriptive attributes, the BOT-2 SF shows only non-significant changes in feasibility compared to the BOT-2 CF.

Funding Statement: This work was supported by the Technology Agency of the Czech Republic under Grant TA ČR Éta 3 TL03000221 and the Technical University of Liberec under Grant SGS N. 21584.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

REFERENCES

1. Burton AW, Miller DE. Movement skill assessment. Champaign, IL: Human Kinetics; 1998.
2. Barnett LM, Webster EK, Hulteen RM, De Meester A, Valentini NC, Lenoir M, Pesce C, Getchell N, Lopes VP, Robinson LE, et al. Through the Looking Glass: A Systematic Review of Longitudinal Evidence, Providing New Insight for Motor Competence and Health. Sports Med. 2022; 52: 875–920. doi: 10.1007/s40279-021-01516-8
3. Hill PJ, Mcnarry MA, Mackintosh KA, Murray MA, Pesce C, Valentini NC, Getchell N, Tomporowski PD, Robinson LE, Barnett LM. The Influence of Motor Competence on Broader Aspects of Health: A Systematic Review of the Longitudinal Associations Between Motor Competence and Cognitive and Social-Emotional Outcomes. Sports Med. 2024; 54: 375–427. doi: 10.1007/s40279-023-01939-5
4. Duncan MJ, Foweather L, Bardid F, Barnett AL, Rudd J, O'Brien W, Foulkes JD, Roscoe C, Issartel J, Stratton G, et al. Motor Competence Among Children in the United Kingdom and Ireland: An Expert Statement on Behalf of the International Motor Development Research Consortium. J Mot Learn Dev. 2022; 10: 7–26. doi: 10.1123/jmld.2021-0047
5. Urbánek T, Denglerová D, Širůček J. Psychometrika: měření v psychologii. 1. vyd. Praha: Portál; 2011.
6. Bruininks R, Bruininks B. Bruininks-Oseretsky test of motor proficiency. 2nd ed. Minneapolis: Pearson; 2005.
7. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 2017; 5: 307. doi: 10.3389/fpubh.2017.00307
8. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, Dmochowski RR. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Mex). 2021; 57: 503. doi: 10.3390/medicina57050503
9. Hands B, Licari M, Piek J. A review of five tests to identify motor coordination difficulties in young adults. Res Dev Disabil. 2015; 41–42: 40–51. doi: 10.1016/j.ridd.2015.05.009
10. Šeflová I, Vašíčková J, Kalfiřt L, Suchomel A. Current Approaches to Motor Competence Assessment in School-Age Children. Phys Act Rev. 2022; 10: 39–50. doi: 10.16926/par.2022.10.20
11. Brahler CJ, Donahoe-Fillmore B, Mrowzinski S, Aebker S, Kreill M. 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. J Occup Ther Sch Early Interv. 2012; 5: 73–84. doi: 10.1080/19411243.2012.674746
12. Carmosino K, Grzeszczak A, McMurray K, Olivo A, Slutz B, Zoll B, Donahoe-Fillmore B, Brahler CJ. 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. J Stud Phys Ther Res. 2014; 7.
13. Coaley K. An introduction to psychological assessment and psychometrics. Los Angeles: SAGE; 2010.
14. Vinçon S, Green D, Blank R, Jenetzky E. Ecological validity of the German Bruininks-Oseretsky Test of Motor Proficiency – 2nd Edition. Hum Mov Sci. 2017; 53: 45–54. doi: 10.1016/j.humov.2016.10.005
15. Maxwell SE, Kelley K, Rausch JR. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. Annu Rev Psychol. 2008; 59: 537–563. doi: 10.1146/annurev.psych.59.103006.093735
16. Šeflová I, Kalfiřt L, Indráčková K. Use of the Bruininks-Oseretsky Test of Motor Proficiency, second edition in school practice. Trends Sport Sci. 2018; 195–199. doi: 10.23829/TSS.2018.25.4-4
17. Blank R, Jenetzky E, Vinçon S. Bruininks-Oseretsky Test of Motor Proficiency | Second Edition. 2. Ausg. Deutschsprachige Version. Frankfurt am Main: Pearson; 2014.
18. Brian A, Pennell A, Taunton S, Starrett A, Howard-Shaughnessy C, Goodway JD, Wadsworth D, Rudisill M, Stodden D. Motor Competence Levels and Developmental Delay in Early Childhood: A Multicenter Cross-Sectional Study Conducted in the USA. Sports Med. 2019; 49: 1609–1618. doi: 10.1007/s40279-019-01150-5
19. Tamplain P, Cairney J. Low Motor Competence or Developmental Coordination Disorder? An Overview and Framework to Understand Motor Difficulties in Children. Curr Dev Disord Rep. 2024; 11: 1–7. doi: 10.1007/s40474-024-00294-y

20. Klingberg B, Schranz N, Barnett LM, Booth V, Ferrar K. The feasibility of fundamental movement skill assessments for pre-school aged children. J Sports Sci. 2019; 37: 378–386. doi: 10.1080/02640414.2018.1504603
21. Ragin CC, Shulman D, Weinberg A, Gran B. Complexity, Generality, and Qualitative Comparative Analysis. Field Methods. 2003; 15: 323–340. doi: 10.1177/1525822X03257689
22. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. J Grad Med Educ. 2012; 4: 279–282. doi: 10.4300/JGME-D-12-00156.1
23. Glaros AG, Kline RB. Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. J Clin Psychol. 1988; 44: 1013–1023. doi: 10.1002/1097-4679(198811)44: 6<1013: : AID-JCLP2270440627>3.0.CO; 2-Z
24. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 2007; 96: 338–341. doi: 10.1111/j.1651-2227.2006.00180.x
25. Bolin E, Lam W. A Review of Sensitivity, Specificity, and Likelihood Ratios: Evaluating the Utility of the Electrocardiogram as a Screening Tool in Hypertrophic Cardiomyopathy: Statistical Concepts in Screening Tests. Congenit Heart Dis. 2013; n/a–n/a. doi: 10.1111/chd.12083
26. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002; 17: 646–649. doi: 10.1046/j.1525-1497.2002.10750.x
27. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J Gen Intern Med. 2004; 19: 460–465. doi: 10.1111/j.1525-1497.2004.30091.x
28. Van Stralen KJ, Stel VS, Reitsma JB, Dekker FW, Zoccali C, Jager KJ. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney Int. 2009; 75: 1257–1263. doi: 10.1038/ki.2009.92
29. Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. Crit Care. 2004; 8: 508. doi: 10.1186/cc3000
30. Šeflová I, Chudoba J, Duncan M, Suchomel A, Bunc V. Motor Competence Prevalence in School-Aged Czech Children: A Cross-Sectional Study. J Mot Learn Dev. 2024; 1–13. doi: 10.1123/jmld.2024-0010
31. Unal I. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. Comput Math Methods Med. 2017; 2017: 3762651. doi: 10.1155/2017/3762651
32. Kosior-Lara A, Ortenburger D, Kuchta M, Korsak-Sabino Belo M, Wąsik J. Physical activity avoidance as a predictor of anxiety and sleep quality in women. Phys Act Rev. 2025;13(2):129-138. doi:10.16926/par.2025.13.26
33. Gosić E, Vlahović H, Lončarić Kelečić I. Prevalence, localisation, and contributing factors of injuries in team sports: a pilot analysis of Croatian football and handball. Phys Act Rev. 2025;13(2):115-128. doi:10.16926/par.2025.13.25
34. Jírovec J, Musálek M, Mess F. Test of Motor Proficiency Second Edition (BOT-2): Compatibility of the Complete and Short Form and Its Usefulness for Middle-Age School Children. Front Pediatr. 2019; 7: 153. doi: 10.3389/fped.2019.00153
35. Yoon D, Choi D, Kim M, Ji S, Joung Y-S, Kim EY. Validity of the BOT-2 Short Form for Korean School-Age Children: A Preliminary Study. Children. 2024; 11: 724. doi: 10.3390/children11060724
36. Deitz JC, Kartin D, Kopp K. Review of the Bruininks-Oseretsky Test of Motor Proficiency, Second Edition (BOT-2). Phys Occup Ther Pediatr. 2007; 27: 87–102.